Evaluation of Energy-Based Modelling For Medical Pathology Classification

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Statistik und Wirtschaftsmathematik

eingereicht von

Dominik Maruszczak, BSc
Matrikelnummer 01027869

an der Fakultät für Mathematik und Geoinformation

der Technischen Universität Wien

Betreuung: Professor Peter Filzmoser
Mitwirkung: Dr. Katja Bühler

Wien, 24. Oktober 2022

Dominik Maruszczak

Peter Filzmoser

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU WIEN Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Erklärung zur Verfassung der Arbeit

Dominik Maruszczak, BSc
Dammstraße 21, 2191 Gaweinstal

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 24. Oktober 2022

Dominik Maruszczak

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acknowledgements

The journey from inception to completion of this Master's Thesis was by no means a predictable linear function with a gradient and a y-intercept, but rather a random walk full of humbling steps back and euphoric steps forward. This work was one of the hardest, but also one of the most rewarding experiences of my life so far. The (loosely translated and paraphrased) words of Austrian Physicist and Nobel Laureate Dr. Anton Zeilinger: "like a child, he only ever pursued that which he found most exciting" aptly describe my own research journey. I had a specific vision of what I wanted to accomplish for my Master's Thesis: I wanted my years of study to culminate with a project that the excited me and would make me think back proudly to this time for years to come. In an almost child-like naivety, I wanted to use the tools I learned to delve into something meaningful because I might never have this opportunity again.

It is a long way from thought to reality and so my first thanks must go to the people without whom this journey would not have happened in the first place: from the TU Wien Dr. Peter Filzmoser and from VRVis Dr. Katja Bühler. They listened to me and allowed me the freedom of realising my visions and ideas, giving me the necessary tools and space for learning about the subject matter but also about myself. For this I am very grateful. Thank you to all University and VRVis colleagues for all the engaging and instructive conversations over the years; I hope all your future endeavours are successful.

The success or failure of any journey can be measured by one's companions on that journey. In Dimitrios Lenis I could not have asked for a better companion, guide and now friend. His patience, motivation and seemingly unbounded intellectual abilities were the catalyst that drove me towards my goal, even if the road came dangerously close to a precipice or doubled back on itself. While his own research and personal life, he welcomed two beautiful girls into this world, would have already kept him busy, he never wavered in his support of me and often the most productive discussions were had during a 2am diaper change. On those few days where he was not able to be there for me, he was replaced by David Major an equally calm, knowledgeable and motivating companion.

Last but not least, the most important people I must thank are with me not just in the course of a major research project, but always. My parents, siblings and closest friends.

Even if they didn't fully understand the details of the work or the nature of specific problems, they picked me up on the days where I was down and celebrated with me on the days I succeeded. They are, quite simply, my strength and stay and I owe them a great debt.

VRVis is funded by BMK, BMDW, Styria, SFG, Tyrol and Vienna Business Agency in the scope of COMET - Competence Centers for Excellent Technologies (879730) which is managed by FFG. Thanks go to our project partner AGFA HealthCare for providing data and valuable input.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Kurzfassung

Diese Arbeit untersucht und vergleicht die Eignung von zwei, auf neuronalen Netzen basierenden, Ansätzen für die medizinische Bilddiagnose von posteroanterioren Thorax-Röntgenbildern, wobei die Diagnose als Klassifikationsproblem interpretiert werden kann. Einerseits trainieren wir ein convolutional neural network basierend auf der standard Deep Learning Literatur. Des weiteren trainieren wir die selbe convolutional Architektur mit der energiebasierten Methodik, d. h. der Standardklassifikator von p(y|x) wird als energiebasiertes Modell für die gemeinsame Verteilung p(x,y) neu interpretiert. Dahinter steht die Überlegung, dass Deep Learning zwar sehr effizient ist und sehr hohe Genauigkeiten erreichen kann, aber im Zusammenhang mit der zuverlässigen Quantifizierung von Unsicherheiten erhebliche Nachteile aufweist, insbesondere in den Bereichen der Erkennung von Verteilungsabweichungen und der Modellkalibrierung. Sie liefern oft ßu sichere"Vorhersagen, was in sensiblen, risikoreichen Anwendungsbereichen wie der Medizin fatal sein kann. Daher ist die Quantifizierung der Unsicherheit für eine breite Anwendung in der Praxis unerlässlich. Das energiebasierte Modell ist wesentlich flexibler und kann zur Erstellung eines hybriden Modells verwendet werden, das sowohl generative als auch diskriminative Fähigkeiten kombiniert. Das führt zu Vorhersagen, welche die eigene Unsicherheit des Modells viel besser widerspiegeln. Darüber hinaus ist das Modell vielseitig genug, um für eine Vielzahl anderer zusätzlicher Aufgaben verwendet zu werden, die in dieser Arbeit kurz untersucht werden, wie z. B. die Erkennung von Ausreißern und die Generierung von Stichproben/synthetischen Daten. Die Theorie zu den energiebasierten Modellen in dieser Arbeit basieren zu einem großen Teil auf der von Will Grathwohl und Yann LeCun durchgeführten Forschung. Das Interesse an energiebasierter Modellforschung hat in den vergangenen Jahren aufgrund von Verbesserungen in der Technologie und ihrer Eleganz und Flexibilität stark zugenommen. Mit dieser Theorie können sowohl die Standardklassenwahrscheinlichkeiten als auch die nicht normalisierten Werte von p(x) und p(x|y) berechnet werden. Wir vergleichen den Deep Learning classifier und das energiebasierte Modell im Rahmen des von der Stanford Machine Learning Group veröffentlichten Lungenröntgendataset CheXpert. Die wichtigsten experimentellen Ergebnisse zeigen, dass das energiebasierte Modell zu starken Klassifikationsergebnissen führt und die Erkennung von Unregelmäßigkeiten in den Daten und Ausreißer im Vergleich zu einem standardmäßigen Deep Learning Modell verbessert. Gleichzeitig ist es auch in der Lage Stichproben von hoher Qualität zu erzeugen. In der bisherigen Literatur wurde festgestellt, dass energiebasierte Modelle auch die Modellkalibrierung verbessern, was jedoch nicht vollständig reproduziert werden konnte. Dieser Ansatz ist der erste, der an medizinischen Bilddaten getestet wurde und Ergebnisse erzielt, die mit dem generativen und diskriminativen Stand der Technik in einem Hybridmodell konkurrieren.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Abstract

This thesis investigates and compares the suitability of two neural network based approaches for medical image diagnosis of posteroanterior chest radiographs, where diagnosis can be interpreted as a classification problem. On the one hand, we train a convolutional neural network using the standard deep learning methodology. On the other hand we train the same convolutional architecture using energy based methodology, meaning the standard classifier of p(y|x) is reinterpreted as an energy based model for the joint distribution p(x,y). The rationale behind this is that deep learning, while being very efficient and able to achieve very high accuracies, has significant drawbacks in the context of reliable uncertainty quantification, specifically in areas of out of distribution detection and model calibration. They often give overly confident predictions, which can be fatal in sensitive, high-risk areas of application such as medicine. Thus, uncertainty quantification is essential for widespread real-world adoption. The energy based model is a lot more flexible and can be used to create a hybrid model that combines both generative and discriminative capabilities, resulting in predictions that reflect the model's own uncertainty much better. In addition, the model is versatile enough to be used for a host of other accessory tasks that are briefly investigated in this thesis such as outlier detection and sample/synthetic data generation. The energy-based work in this thesis is based in a large part on EBM research conducted by Will Grathwohl and Yann LeCun. Energy based modelling research has recently seen a strong increase in interest due to improvements in technology and its elegance and flexibility. In this setting, the standard class probabilities can be computed as well as unnormalized values of p(x) and p(x|y). We compare the deep learning classifier and the energy based model in the context of medical image classification using the chest x-ray dataset CheXpert, published by the Stanford Machine learning Group. The main experimental findings showed that energy based training results in strong discriminative results and improve out of distribution detection and outlier detection compared to a standard deep learning model, while also being able to generate samples of high quality. Previous literature has found that energy based models improve model calibration as well, this could not be fully reproduced. This approach is the first to be tested on medical image data and achieves results rivaling the generative and discriminative state-of-the-art within one hybrid model.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Contents

Kurzfassung **vii**

Abstract **ix**

Contents **xi**

1 Introduction **1**
1.1 Motivation 3
1.2 Statement Of Problem 6
1.3 Aim 7
1.4 Methodology 7
1.5 Structure 8

2 Related Work **11**
2.1 Deep Learning 11
  2.1.1 Deep Learning In General 11
  2.1.2 Deep Learning For Medical Image Classification 12
2.2 Out of Distribution Detection and Calibration In Deep Learning 12
  2.2.1 Out Of Distribution Detection 12
  2.2.2 Calibration 13
2.3 Energy Based Modelling 13

3 Introduction To Deep Learning **15**
3.1 Notation 15
3.2 Motivation 16
3.3 Historical Evolution 18
3.4 The Artificial Neural Network 22
  3.4.1 Definitions 22
  3.4.2 Universal Approximation 26
3.5 Training Deep Neural Networks 30
  Cost Functions 30
  Optimization 36
3.6 Network Architecture 40
  Convolutional Neural Networks 40

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Residual Network Implementation . . . 44

4 Introduction To Energy Based Modelling 47
4.1 Concept And Origin of Framework . . . 47
4.2 Energy Function . . . 50
4.3 Training and Inference . . . 52
4.4 Probabilistic Model . . . 54
4.5 Cost Functions . . . 57
4.6 Approximating The Partition Function . . . 59
4.6.1 Monte Carlo Markov Chain Methods . . . 59
4.6.2 Applied MCMC - Gradient Approximation . . . 60
4.6.3 Stochastic Gradient Langevin Dynamics . . . 62
Langevin Equation and Langevin Dynamics . . . 62
Stochastic Gradient Langevin Dynamics . . . 63
4.7 Benefits And Drawbacks Of The Energy Based Framework . . . 63

5 Evaluation And Comparison Of Energy Based Modelling And Deep Learning 65
5.1 Energy Based Modelling Compared to Deep Learning . . . 65
5.1.1 Deep Learning . . . 66
Calibration . . . 67
Out of Distribution Detection . . . 72
5.1.2 Energy Based Modelling . . . 76
5.1.3 Summary . . . 77

6 Methodology And Model Design 79
6.1 Contextualization Of Research Goals . . . 79
6.1.1 Network Definition . . . 81
6.2 Data Exploration . . . 82
6.2.1 CheXpert . . . 83
Technical details . . . 83
Exploration . . . 84
6.2.2 Outlier Detection . . . 87
6.2.3 ChestXray14 . . . 90
6.2.4 INbreast . . . 91
6.2.5 ImageNet . . . 91
6.3 Implementation . . . 91
6.3.1 Environment . . . 91
6.3.2 Data Pre-Processing . . . 92
Data Augmentation . . . 92
6.3.3 Model Design . . . 93
Model Architecture . . . 93
6.3.4 Model Training . . . 95
Statistical Considerations . . . 95

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Resolution . . . 96
Training A SOTA discriminative Deep Learning Classifier . . . 96
Training An Energy Based Model . . . 97
6.3.5 Evaluation . . . 100
Discriminative Power . . . 100
Out Of Distribution Detection . . . 100
Calibration Metrics . . . 102
Outlier Detection . . . 102

7 Results And Discussion 105
7.1 Discriminative Power . . . 105
7.2 Out Of Distribution Detection . . . 110
7.2.1 PCA . . . 110
7.2.2 Modelling Approaches . . . 112
7.3 Calibration . . . 115
7.4 Outlier Detection . . . 119
7.5 Sample Generation . . . 120
7.6 Discussion . . . 121

8 Conclusions 123

List of Figures 129

List of Tables 131

List of Algorithms 133

Index 135

Acronyms 139

Bibliography 141

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 1

Introduction

Medical imaging is a crucial tool in modern medicine for the diagnosis and treatment of various ailments Ostensen (2001). It plays a central role in confirming, assessing and documenting the course of many diseases in response to treatment. There is a variety of imaging technologies such as ultrasonography, x-rays, mammography, computed tomography (CT scans), and nuclear imaging. X-rays were discovered in the late 19th century by Wilhelm Röntgen Röntgen (1895) in the course of his investigation of special tubes that emitted unknown types of radiation. Following the publication of his research, they quickly became a popular diagnostic method. His research caused a sensation for giving actual view inside the body. Due to their relative non-invasiveness, they do not require any surgical procedures to expose afflicted areas, and relative simplicity, they quickly became vital to generate images of the body and they remain one of the most important tools today Spelic et al. (2010). Chest x-rays specifically, are the most popular tool for the evaluation of chest pathologies as they deliver a large amount of important information about a patient using only one image van Beek and Murchison (2019). The chest contains several of the body's most important organs such as heart and lungs, and x-rays allow for a fast assessment of their status van Beek and Murchison (2019). Eurostat reports diseases of the respiratory system accounted for 7.5% of all deaths in the EU in 2016 (not including lung cancer) Eurostat (2020) underlining the importance of fast and accurate chest pathology diagnostic tools in Europe. But also elsewhere, for instance in developing countries, 80-90% of imaging needs are covered with x-rays and ultrasounds, showing how fast and simple diagnostic tools save many lives Shah et al. (2010).

Early in the second half of the 20th Century, x-rays were began to be digitized, leading to increased development of tools for automatic diagnostics Litjens et al. (2017). Starting from the 1970s, the evolution of these systems has seen those which are completely designed by humans, to those that are trained by computers using machine learning models and example data Litjens et al. (2017), making it possible to automatically transform input x-rays to diagnoses Litjens et al. (2017). The class of models that has been shown to be highly adept at image classification are deep learning models, specifically convolutional neural networks (CNNs) LeCun et al. (1998). They provide a scalable approach to image classification and object recognition tasks. They are characteristic for having three main types of layers: the convolutional layer, the pooling layer and the fully connected layer Goodfellow et al. (2016). This combination useful to detect subtle, low-level features such as edges or shadows which, when combined on an aggregate level, makes it possible to identify high-level features and patterns.

Work on CNNs has been done since the late seventies Fukushima and Miyake (1982) and they were already applied to medical image analysis as early as 1995 Lo et al. (1995). CNN architectures have become the standard instrument for diagnosing medical images, achieving high accuracies, the proportion of correctly classified images, that challenge even the human experts in some tasks Litjens et al. (2017). However, high accuracy is only one part of a successful diagnosis; another important part is uncertainty Gal (2016). Models may perform well in experimental conditions, but when classifiers are used for real-world tasks, they tend to fail when the training and test distributions differ Hendrycks and Gimpel (2016). Simply put, this means that while the model might indicate high confidence in their output, they completely fail in detecting input firmly outside of their validate input range, yielding nonsensical results like tumor scores for cat images Goodfellow et al. (2014). Ideally, classifiers should be able to correctly quantify the level of uncertainty attached to the their prediction and thus indicate when they are likely to fail, since their errors have serious repercussions.

The energy based framework could provide an elegant solution to this problem. An energy based model is a generative model, meaning that it estimates a distribution over the input data. In doing so, they capture latent dependencies in the data. This is done by training a function, the energy function. This function returns a scalar output that expresses the un-normalized probability density of images, the energy LeCun et al. (2006). Energy can also be understood as a measure of compatibility between input and response variables, where lower energy means better compatibility. This means the computer would be able to learn the entire distribution of chest x-rays, and use this to make good quality predictions, estimate its own uncertainty, generate good-quality images and detect whether an input fits to the training distribution Grathwohl et al. (2019) These models are very flexible and can be used for downstream discriminative tasks.

The main focus of the thesis will be the comparison between the energy based framework and the classical deep learning approach. Specifically, focusing on how accurately the two approaches classify x-ray images, and how they behave in the wider context of understanding the underlying data such as identifying nonsensical input.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

1.1 Motivation

Chest radiographs are still the standard diagnostic tool for the evaluation of chest pathologies Litjens et al. (2017). While not without risks, they are accurate when investigating the overall status of the heart, lungs and skeletal system van Beek and Murchison (2019). A chest x-ray is essentially a two-dimensional projection of a three-dimensional volume where several different types of tissue are overlayed, making exact distinction of features with the human eye difficult Delrue et al. (2011). Radiologists traditionally evaluate the images directly, diagnosing conditions, categorizing diseases Drake et al. (2009). In addition to chest radiographs, technological advances have brought forth new imaging modalities such as multi-slice (volumetric) and multi- energy CT, multi-parametric and multi-frame (dynamic) MRI, multi-dimensional (3D+time) US, multi-planar interventional imaging, or multi-modal (hybrid) PET/CT and PET/MRI imaging technologies, giving radiologists many different diagnostic tools to choose from Suetens (2009).

This advance in the state of technology, the growing amount of images, the implicit urgency that comes with needing a x-ray and the relative shortage of personnel with the expertise to make a sound analysis, has put radiologists under pressure, increasing the probability of errors. In a study evaluating the miss rate of non-small cell lung cancer in patients who had presented with a lung lesion, 19% of cases were missed Quekel et al. (1999). These cases typically had small lesions, often with overlaying tissue. Furthermore, diagnostic delays caused 43% of patients to progress into a higher stage before being definitively diagnosed. Considering these issues and bottlenecks, improving and automating the analysis of chest x-rays would have an immediate improvement in patient care van Beek and Murchison (2019). Building on this notion, a separate study investigated whether a CAD system could detect cancer in the 89 x-rays of patients that were previously missed by Radiologists. The system was able to identify missed lesions in 46 of the 89 images (52%) White et al. (2009). A large study in rural Africa, involving 46,099 participants for screening of tuberculosis, was designed as a performance comparison between an automatic software to Radiologists both on site and at a remote, centralized location Melendez et al. (2017). It showed that the diagnostic accuracy of all three methods was very close, and the performance of the automated software was comparable to that of expert readers. Hence, the introduction of this system in a rural setting would be a major benefit for patient care in underserved areas in Africa. Images can be interpreted and relevant decisions can be made at the point of care, which is particularly useful in remote areas van Beek and Murchison (2019).

Thus it is evident that there is a great need for more efficient and reliable automated methods for medical image analysis. Driven by this need, the state of technology has developed to such an extent that these methods can realistically be integrated into daily clinical routine Maes et al. (2019). Machine learning, specifically deep learning, has developed into a crucial technology since the 2010s Goodfellow et al. (2016). Deep learning models are designed to automatically go through large amounts of data, focusing

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

on the underlying patterns and using this information to make predictions. Due to an increased number of datasets, a concurrent increase in the sizes of these datasets ter Haar Romeny (2019b), better computational resources and more efficient learning methods, deep learning has been able to achieve near human performance in certain areas Goodfellow et al. (2016). The advances in the digitization of hospitals and investments in their information systems, means that large databases of medical images and other medical information (including demographics, clinical findings, blood tests, pathology, genomics, proteomics) are being built up. Given an appropriate privacy set-up, these databases could be more accessible for research Maes et al. (2019), improving the quality of the systems used. As a result, there has been an organic shift from systems completely designed by humans to those that are trained by computers both with and without annotations (supervised/unsupervised learning) Litjens et al. (2017).

Deep learning, specifically learning based on convolutional neural networks (CNNs), has shown great promise in medical applications Greenspan et al. (2016) ter Haar Romeny (2019a). CNNs, are able to automatically identify low-level features such as edges and shadows without previous feature engineering. The aggregation of these low-level features results in high-level features that drive classification Goodfellow et al. (2016). This has already made them a valuable tool to radiologists for x-rays and CT scans, among other things, where the analysis is often time-consuming and subject to significant intra- and inter-observer variability, which could reduce the significance of the clinical findings Litjens et al. (2017). Using deep learning solutions for diagnostic purposes often involves the method of classification Litjens et al. (2017). In 2017, the AI platform Kaggle and the consulting firm Booz, Allen and Hamilton held the Data Science Bowl, inviting people to develop machine learning tools for the prediction of lung cancer given a chest CT investigation **?**. The competition was based on a training and validation dataset from the US National Cancer Institute. One of the winners, the medical imaging startup Aidence from Amsterdam, received regulatory approval in 2017 and subsequently several studies were undertaken. Their system demonstrated accuracies comparable to that of experts. The Aidence model is an example of deep learning being successfully applied to diagnostic tasks. The full system consists of two separate deep learning models to generate candidate locations (a classification network) and to filter out false- positive candidates.

Classification can be roughly split into multi-label classification, where an image could fall into multiple classes, and multi-class classification, where an image can only fall into one of at least two classes Goodfellow et al. (2016). When an image is classified as a specific class, it means that the classifier in question, in this case the neural network, calculates the conditional probability $p(y|x)$, where x represents the input image and y the output class, which is expressed as a vector of un-normalized scores given by the network. In the multi-class case, the softmax function, the multi-dimensional generalisation of the logistic function Goodfellow et al. (2016), is applied to the network output to create a

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

normalized distribution of scores, giving the conditional probability. This means that scores reflect model certainty, i.e. a prediction with high score implies a model is highly certain about its prediction. However, softmax scores are not appropriate uncertainty measures; a model can be uncertain even with a high softmax Gal (2016). In fact, with modern deep learning models there can be a disconnect between the output a trained model is delivering and whether this output is representative of the likelihood seen in a set of test data, i.e. models are miscalibrated Guo et al. (2017). It is not uncommon in real-world situations, that a classification model can receive an input image that is significantly different from the data it was originally trained on Hendrycks and Gimpel (2016). For instance, the test could be from a different domain or can shift and change over time away from the training data. Considering the dataset as a distribution, this notion can also be expressed as: the training and testing distributions being significantly different; an input image that is not part of the training distribution is considered out of distribution. This could be for instance: a mammogram which is given to a network that was originally trained on chest x-rays, or even a lateral chest x-ray is given to a network that was trained on posteroanterior chest x-rays. In both of these cases a classifier might still deliver a high confidence prediction while being wrong. It has been shown that even if both distributions (training and testing) deal with chest x-rays, but from different sources, this can have drastic effects on the model's performance Pooch et al. (2019). The fact that the models can silently fail in this way, and give no indication of their own uncertainty, can be a significant deterrent to the large scale adoption of these tools when diagnosing chest x-rays, or any other type of medical image, even though the benefits of widespread clinical use of AI technology in radiological diagnosis could be significant. Considering the fact that human lives are at stake (for example through misdiagnosed cancer or heart problems), this has serious implications for AI safety Amodei et al. (2016). Ideally, a model would behave in such a way that predictions are returned with the added information that the point lies outside of the data distribution and have low confidence Jiang et al. (2012). Models should know what they do not know Gal (2016). There are methods to increase the safety and certainty of classifiers by improving out of distribution detection, robustness and calibration Gal (2016). However, these methods are often "add-ons" that add to the complexity of the modeling set up, require more data and increase the need for hardware Goodfellow et al. (2016). Ideally, a framework could be used where one model is trained on images, in the case of this thesix x-rays, and it encompasses all necessary attributes to deliver calibrated predictions/diagnoses of chest pathologies and perform out of distribution detection, i.e. include information about uncertainty.

Generative models aim to estimate (joint) distributions and allow for the creation of new samples and confidence bounds that are useful for analysis and decision making Ghahramani (2015); whereas classifiers deliver point estimates of parameters and predictions Gal (2016). Generative models have been used in out of distribution detection for medical images Chen et al. (2018) and while they have delivered good experimental results, there remains room for improvement. Recently one type of generative model, the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

energy based model LeCun et al. (2006), has been used to achieve good experimental results on out of distribution detection Liu et al. (2020). This model is of particular interest as it is very flexible and can accommodate different mathematical settings, e.g. probabilistic or non-probabilistic LeCun et al. (2006), meaning they estimate a distribution and it is possible to calculate probabilities. The use of this framework is the foundation for the creation of a hybrid model that contains both generative and classification (or discriminative) abilities, delivering highly accurate predictions while being naturally calibrated and able to detect out of distribution inputs Grathwohl et al. (2019). Intuitively, this means the model does not just express a class for an image, but it understands what this means in the broader context of the entire dataset; the probability expressed by the model means something deeper. Conversely, deep learning classifiers are only focused on the classification aspect. The hybrid model in this thesis attempts to realize the idea that one model can organically encompass all necessary attributes to deliver trustworthy diagnoses and could thus potentially add a lot of value in the detection of chest pathology.

1.2 Statement Of Problem

There have been significant benefits in using automated systems when diagnosing chest pathology and supporting radiologists while they are facing rising pressures van Beek and Murchison (2019). The advent of deep learning suggests that as systems become more sophisticated and deliver better results, they can significantly contribute to overall patient care Maes et al. (2019). However, there are significant drawbacks to state of the art deep learning classifiers that, given the high risk environment that is medicine, hinder the roll-out of these models into a real clinical setting Gal (2016). While they are highly confident, i.e. produce predictions with very high scores, it has been shown that these predictions are at risk of not representing the true likelihood seen in the data, that is they are *uncalibrated* Guo et al. (2017). Furthermore, they show an inability to distinguish between different training and testing distributions, delivering high-confidence results while actually being incorrect, i.e. they do not perform well in detecting out of distribution inputs Hendrycks and Gimpel (2016). Overall, they are not well-suited to quantify (their own) uncertainty, which lowers trust in the technology and slows widespread adoption in medicine Gal (2016) Pooch et al. (2019). Generative models, those which can be used to estimate distributions of data, are a useful tool to quantify uncertainty and perform well on out of distribution detection Gal (2016) but do not show (close to) state of the art discriminative performance on their own Grathwohl et al. (2019). Thus, some approaches combine two or more deep learning models to achieve both good discriminative and generative performance Lakshminarayanan et al. (2017). One modelling framework that encompasses both discriminative and generative capabilities, and performs well in both domains, would be desirable.

This thesis investigates if the concept of a unifying framework, in the form of an energy based model LeCun et al. (2006), is suitable for medical imaging. Specifically, whether

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

it is able to achieve state of the art discriminative results, while performing well on out of distribution detection and delivering calibrated predictions. The energy based model is compared to a state of the art deep learning classifier. Both approaches are based on the convolutional neural network architecture, but interpreted and trained in different ways. This idea has recently been the central theme in an approach that uses energy based modelling to implement a hybrid model using CIFAR-10 Grathwohl et al. (2019) Krizhevsky (2009) that achieves good experimental results by delivering calibrated predictions and performing well on out of distribution detection.

1.3 Aim

The goal of this thesis is to study the suitability of a hybrid energy based model, as described in Grathwohl et al. (2019), for medical image pathology classification. The following questions will be addressed:

1. Does training a neural network classifier in a hybrid EBM scheme reach comparable results to a standard deep learning discriminative training setup?
2. Can the intrinsic generative model of the hybrid setup be utilised for out of distribution detection?
3. How does EBM training affect the model calibration?

In a nutshell, this work tries to answer if training the same neural network architecture in an energy based setup, yields comparable discriminative power while meeting significant requirements for applications in clinical setups. For this a battery of experiments and tests are performed, described in detail in the following section.

1.4 Methodology

Set Up A neural network based on the WideResNet architecture is chosen and trained using SOTA deep learning methodology and EBM methodolgy, respectively. The models are trained on the publicly available dataset chest x-ray dataset CheXpert Irvin et al. (2019) to discriminate between the four classes: cardiomegaly, lung lesion, pleural effusion and fracture. The training images are varied in size, namely the following resolutions: 64×64, 128×128 and 224×224. The discriminative power, the calibration and the out of distribution capabilities of the two models are compared on the different resolutions.

Evaluation

1. **Discriminative Power**

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

  - the DL classifier is evaluated using the area under the Receiver Operating Characteristic curve (AUROC) and compared to discriminative results published by Stanford in Irvin et al. (2019). To reduce overall variability, randomised resampling is used and an empirical confidence interval of the overall AUROC is calculated. In addition, visual evaluation using GradCam is generated. This is a type of visual explanation to make CNN results more transparent by showing what regions of the image drive classification.
  - the discriminative power of the energy based model is also calculated using the average AUROC and compared to the results generated by the DL classifier.

2. **Out of distribution detection**
    - In the course of an in-depth exploration of the data, an outlier detection is carried out on the x ray data using (robust) Principal Component Analysis Candès et al. (2011).
    - Out of distribution detection is investigated based using three different out of distribution datasets:
        a) CXR14 Wang et al. (2017)
        b) Inbreast Moreira et al. (2012)
        c) ImageNet Russakovsky et al. (2015)

      PCA is used to carry out a preliminary, non-model based out of distribution detection to get a first indication of how well the strongest features in the data, the principle components, can be used to separate the data. Post-modelling, the two models are exposed to the out of distribution datasets. The Maximum Prediction Probability is calculated for both models and based on this the respective AUROCs are calculated and compared Hendrycks and Gimpel (2016). This will give an insight into how well the models can tell apart data not part of the training distribution

3. **Calibration**
    - **Calibration for both models** is evaluated using the expected calibration error (ECE) Naeini et al. (2015) and with the use of reliability diagrams.

1.5 Structure

The diploma thesis is composed of six chapters and is organized in the following way:

Chapter 2 Related Works A survey of relevant publications are presented, describing the current state of the art. It focuses on the usage of deep learning in the context of medical image classification as a whole, as well as specific to the aims of this thesis. The current state of research on energy based modelling is also reviewed, giving an overview

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

of the concepts and tools used in this thesis.

Chapter 3 Introduction to Deep Learning This chapter introduces the deep learning framework. It motivates and develops the mathematical theory of the neural network model and details the training procedure for deep learning, highlighting optimisation methods, architectures, and loss functions.

Chapter 4 Introduction to Energy Based Modelling This chapter is dedicated to developing the the energy based framework. It motivates and develops the mathematical and statistical theory of the energy based model, highlighting the different possible interpretations and problems of the model. Furthermore, the training procedure is developed, specifically MCMC-based method inspired by Bayesian inference.

Chapter 5 Evaluation and Comparison of Energy Based Modelling and Deep Learning This chapter evaluates some of the characteristics observed in the two frameworks and highlights their similarities and differences.

Chapter 6 Methodology and Model Design Describes the way the models used in the experiments are designed and the overall training process of the two frameworks, including data generation, pre-processing and model-specific training specifications. Furthermore, it is dedicated to laying out how the experiments are conducted, what metrics are used for measuring the results and how this will give insight into answering the research questions.

Chapter 7 Results and Discussion Compares the results achieved by the deep learning model and the Energy Based Model with respect to the the research questions formulated in 1.3

Chapter 8 Conclusion Presents a thorough discussion of the results, contextualising them with the initial research questions, and drawing final conclusions.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

CHAPTER 2

Related Work

This chapter focuses on already published literature that has dealt with the topics addressed in this thesis. First, an overview of literature on convolutional neural networks and their applications in image classification is given in 2.1.1. This will be extended in 2.1.2 by presenting works that have successfully used deep learning methods for medical image classification, specifically chest x-rays. Section 2.2 will review literature on out of distribution detection and calibration. These are two vital concepts that can determine whether a model will be adopted into a real-life setting or will remain in experimental. Section 2.3 will review publications on energy based modelling and how it can be used on medical image classification with a focus on out of distribution detection and calibration.

2.1 Deep Learning

2.1.1 Deep Learning In General

Deep learning in computer vision has been a widely researched topic in the past 2-3 decades that has seen tremendous growth and yielded solutions that deliver incredibly high accuracies using deep convolutional networks Krizhevsky et al. (2017), Szegedy et al. (2015) and He et al. (2016). These early breakthroughs ushered in an ongoing period of high research activity to address increasingly complex problems, with increasingly complex modelling architectures and on increasingly varied data. For instance: text-to-image generation models Ramesh et al. (2021), high dimensional image synthesis models Niemeyer and Geiger (2021) and learning useful representations for images without incorporating knowledge of the 2D input structure Chen et al. (2020b) or human motion prediction Ma et al. (2022), to name a few. While deep neural networks have brought a lot of experimental benefits, this thesis also elaborates on some of their drawbacks. Yuille and Liu (2021) provides a survey of how much they have "really helped".

2.1.2 Deep Learning For Medical Image Classification

Deep learning for medical imaging has been a very active field of research. The most successful convolutional networks have arguably been the U-Net Ronneberger et al. (2015) and V-Net Milletari et al. (2016) and their variants. There has been considerable research with them on different image modalities, e.g. CTs and MRIs, and on different tasks, such as tumor segmentations on the brain Havaei et al. (2017) and liver Christ et al. (2017). Chest x-rays are one of the most important image modalities since one image already contains a large amount of important information about a patient's most important organs, and they are relatively non-invasive van Beek and Murchison (2019). In addition, several large, labelled chest x-ray datasets have been published Irvin et al. (2019) Bustos et al. (2020) Wang et al. (2017) that have encouraged the medical image community to publish a variety of specialized research, comparisons and surveys. Wang et al. (2017) published the first large scale results for the application of deep learning on chest pathologies which served as an inspiration for many papers. Baltruschat et al. (2019) compares the performance of various approaches to classify the 14 disease labels. Rajpurkar et al. (2018) compares the performance of an ensemble of deep learning models to professional radiologists. Novikov et al. (2018) uses a InvertedNet, a convolutional architecture, to investigate multi-class segmentation of anatomical organs, namely for lungs, clavicles and the heart. Junior et al. (2021) uses a DenseNet architecture for cardiomegaly detection. The COVID-19 pandemic has further fuelled research and development of chest x-ray diagnostic systems, especially early detection and severity evaluation. For instance, Samala et al. (2021) trains a GoogleNet to assess severity of COVID and Ibrahim et al. (2021) proposes a network to diagnose COVID-19, among other diseases. Diagnostic systems based on deep learning methods do not need to be trained on x-ray data from scratch. Many publications successfully use pre-training, or transfer learning. In a pre-trained network, the network is first trained on a large dataset for a different task (typically ImageNet Russakovsky et al. (2015)), and the resulting weights are used as an initialization for training the model on chest x-rays Yosinski et al. (2014). This not only increases efficiency but is also very beneficial for the performance of a model, as summarised in Çallı et al. (2021). Rajpurkar et al. (2017) uses transfer-learning with fine tuning, which raised the discriminative results on chest x-ray multi-label classification even higher. This thesis will leverage the benefits of pre-training and also generate highly accurate predictions.

2.2 Out of Distribution Detection and Calibration In Deep Learning

2.2.1 Out Of Distribution Detection

Besides creating highly accurate models, this thesis is focused on the context, if there is any, that network predictions deliver. Specifically, one important question to address is whether a CNN can discern between images that are from the distribution it was

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

trained on and those that are not. This is known as out of distribution detection (OOD). OOD has been studied in a series of publications which describe the phenomenon and offer possible solutions. Hendrycks and Gimpel (2016) is the defacto state of the art reference on this subject and provides ways to measure how well a classifier detects out of distribution samples. Bevandić et al. (2018) and Chen et al. (2020a) propose ways to perform robust out of distribution detection. Hendrycks et al. (2018), Hsu et al. (2020) and Lee et al. (2018) develop various methods to detect OOD samples. The importance of OOD has greatly increased in the last 3-5 years, especially in sensitive areas such as medical image classification. Cao et al. (2020) provides a benchmark for medical out of distribution detection. Pooch et al. (2019) and Chen et al. (2018) investigate the effects of OOD for the medical images, specifically chest x-rays, and finds that these can have drastic effects on the generalization of models and hinder real-world use.

2.2.2 Calibration

OOD is often viewed in conjunction with calibration. Calibration describes the situation whether a model delivers confident predictions, i.e. estimates representative of the true correctness likelihood in the data. Simply put, if a model predicts a disease with 90%, then this disease should also have a prevalence of 90% in the data. This can be a major challenge, especially with the strong class imbalances that often occur in medical data. Hence, calibration also give an intuition about how uncertain a model's prediction is. Rajaraman et al. (2022) publishes a systematic analysis of the effect of model calibration on its performance on chest x-rays, using deep learning classifiers. Guo et al. (2017) highlights the phenomenon that the hoghly accurate classifiers of today tend to be over-confident and miscalibrated, and evaluates the performance of various calibration methods on state of the art networks with proposals on possible metrics to measure calibration and possible solutions. Gal (2016) explores the uncertainty of neural networks and tools to estimate it. Nguyen et al. (2015) elaborates on the phenomenon of highly confident predictions by neural networks for nonsensical images. Lee et al. (2017) develops a simple and unified framework to detect miscalibration and out of distribution samples.

2.3 Energy Based Modelling

Energy based modelling in computer vision was prominently represented in the FRAME model Zhu et al. (1996) - a Markov Random Field model which is a type of energy model. This work was extended by Xie et al. (2016) where the energy function is parameterized by a CNN structure, as it is done in this thesis. Work with these types of models has been published over several decades in academic literature by Ackley et al. (1985), Hinton et al. (2006b) and Hinton and Salakhutdinov (2006). But implementations remained small or theoretical until hardware improved. Central to the work in this thesis are, Du and Mordatch (2019), Mnih and Hinton (2005), Hinton et al. (2006a) implement EBMs where inputs are directly mapped to outputs, which inspires the creation of models that could then be used for downstream classification tasks. The most comprehensive collection

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

of EBM theory comes from LeCun et al. (2006) which succinctly builds up the theory behind the learning and optimization problems in energy based modelling. Grathwohl et al. (2019), who advocates the usage of EBMs to leverage generative capabilities for downstream discriminative tasks, serves as the bedrock of this thesis. His work is extended and explored, for first time, to the medical domain. In fact, EBMs have currently not been tested on medical image data at all.

The different (probabilistic) EBM publications largely differ in their approach to estimate a partition function. This thesis uses MCMC sampling to approximate the partition function. Hinton et al. (2006b) and Salakhutdinov and Hinton (2009) apply Contrastive Divergence, that is MCMC chains initialized from training data, to estimate the partition function. Tieleman (2008) advocates the use of its extension: Persistent Contrastive Divergence, which propagates MCMC chains throughout training. This thesis, as in Du and Mordatch (2019), Welling and Teh (2011) and Grathwohl et al. (2019), initializes chains from random noise which and uses the idea of PCD, to keep past samples in a replay buffer to reduce mixing times. To further increase efficiency, Gradient based MCMC based on Stochastic Gradient Langevin Dynamics is used for sampling, which was also published by in Teh et al. (2003) and Xie et al. (2016).

In the context of energy based modelling, out of distribution detection and calibration has been studied in works such as Grathwohl et al. (2019), Liu et al. (2020) and Wang et al. (2021), where the generative capabilities of the model are used to determine out of distribution samples at the same time deliver well-calibrated predictions, out of the box.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU WIEN Bibliothek Your knowledge hub

CHAPTER 3

Introduction To Deep Learning

This chapter will develop the mathematical and statistical concepts that form the foundation of the deep learning (DL) framework. To begin, the most important notation used in this chapter is briefly outlined. DL will be motivated and contextualised, before formally defining the neural network model. Section 3.4.2 will investigate one of the most important applications of this model: the universal approximation of neural networks. This will be followed by an analysis of the training process and the optimization problem solved during training. The section will be closed by describing the development of network architectures and what kind of architectural set up will be used in this thesis.

3.1 Notation

For any learning model, the relevant spaces on which they act need to be defined. $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C} \subset \mathcal{A} \times \mathcal{B}$ are used for the general formulation of input, output and combined spaces, respectively. Unless otherwise stated, they are considered as subspaces of $\mathbb{R}^d$, $\mathbb{R}^k$ and $\mathbb{R}^{d \times k}$, where $k, d \in \mathbb{N}$ and $k \leq d$. Their lowercase counterparts a, b and $c :- (a, b)$ will be used to describe elements of the respective spaces. The hat notation will be used to indicate estimators, that is $\hat{b}$ is the estimate for an element $b \in \mathcal{B}$. Arbitrary functions will be denoted using f and networks will use the notation G. Weights, sometimes also called parameters, will be denoted using w and ω; the space of all possible weights will be written as Ω.

It will be necessary to define a relevant topology for the networks to operate on in order to prove their approximative capabilities: $C(K) := \{f \colon K \to \mathbb{R} \colon f \text{ continuous}\}$ is the space of continuous, real valued $d-$dimensional functions. $||f||_\infty := \underset{x \in K}{sup}|f(x)|$ denotes the the supremum norm. The norm gives an upper bound on f, and for two functions f, g it gives a bound on how much f and g differ from one another - an essential tool to determine whether approximation to an arbitrarily close degree is possible.

A measure imposes a notion of size or volume on a set, meaning they convey an idea of how much space that set takes up relative to the larger space in which it lies Kesavan (2019). Regular measures exhibit, as their name suggests, structural regularity Kesavan (2019). In other words, the measure of a compact set like $C(K)$, has *finite* measure when it is regular. Conversely, it would not make sense if $C(K)$ were compact but at the same time took up infinite space. In addition, regular measures enable the use of the Riesz Representation Theorem Rudin (1987) on a compact K. It tells us that the *dual space* is the space of finite, signed regular Borel measures on K defined above. Thus it is an important concept for proving universal approximation.
The space of finite, signed, regular Borel measures is denoted by:

$$\mathcal{M} \coloneqq \{\mu \colon \mu \text{a signed Borel measure}\}$$

Dual spaces are spaces of linear functionals, that is linear mappings from an arbitrary vector space V to its space of scalars, such as $\mathbb{R}$ or $\mathbb{C}$, with addition and scalar multiplication defined pointwise Rudin (1974). Any vector space V has a corresponding dual space V' Bourbaki (1966). Their importance for neural networks becomes clear by re-visiting the above idea that classification can be described by the search for a hyperplane that separates data into different classes. The equation of a hyperplane $w'x = \theta$, where w represents the weight vector, x the input vector and $\theta \in \mathbb{R}$ the threshold (see definition 3.2 the definition of the Rosenblatt perceptron), is a linear functional on $\mathbb{R}^n$: $F : \mathbb{R}^n \to \mathbb{R}, x = (x_1, \dots, x_n) \mapsto w'x = F(x) = \theta$. They are an essential tool for carrying out the linear algebra operations that are involved when training neural networks.

For probabilistic statements, an underlying probability space with probability measure $\mathbb{P}$ is assumed which exists on the probability space $(\Xi, \mathcal{B}, \mathbb{P})$. Where $\mathcal{B}$ is the Borel σ-algebra, Ξ is the set of all outcomes and $\mathbb{P} \colon \mathcal{B} \to [0, 1]$ is the probability measure. For a random variable X, $\mathbb{E}[X]$ as its expectation. The Lebesgue measure is assumed as a dominating measure.

3.2 Motivation

The over-arching task addressed in this thesis is the training, evaluating and comparing of two neural network-based models for the purposes of diagnosing, more generally classifying, chest radiographs. Thus, before going into detail about neural networks and the specifics of the models explored in this thesis, it is necessary to lay out the foundation of the learning problem considered in this work: classification in a supervised learning setting.

Supervised learning is a field of statistical learning that uses input-output pairs as examples to train a model for prescriptive or descriptive purposes Hastie et al. (2009). The key

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

difference to unsupervised learning is the fact that output information is available during training. On a high level, supervised statistical learning models can be distinguished by their type of output $b \in \mathcal{B}$ for a given input $a \in \mathcal{A}$; problems in which a model is used to produce an outcome measurement where $\hat{b} \in \mathcal{B} \subseteq \mathbb{R}$, or in other words a quantitative output, are typically referred to as regression problems. On the other hand, if a model is required to produce a qualitative or categorical output, it is referred to as a classification problem Hastie et al. (2009). The output of a classification model $\hat{b}$ is typically part of a discrete, finite set of labels of elements with no specific order, for instance: $\mathcal{B} := \{malignant, benign\}$.

Mathematically, learning can be considered a problem of function approximation Hastie et al. (2009). The output is generated from the input by ways of an unknown, underlying function f. The learning algorithm can be thought of as a procedure that uses the training data $\mathcal{C}$ to fit a model that approximates this function which performs the given task well on the training data and also generalizes to unseen data, as determined by a chosen evaluation metric Berner et al. (2021). For classification specifically, the data is grouped according to an unknown classification rule and training a classification model can be interpreted as the approximation of that underlying rule Hastie et al. (2009). The simplest approach in approximating a classification rule is the assumption of linear class boundaries. More generally, this means that classifying data can be thought of as a search for an appropriate space and hyperplane that is able to separate the data into different categories, with no observation assigned to the wrong category. However, in reality this is virtually impossible, especially in very large dimensions. On the one hand this leads to soft margins, i.e. boundaries that allow for some mis-classifications Hastie et al. (2009), and on the other hand this means that the introduction of non-linear decision boundaries must be considered.

The support vector machine algorithm Cortes and Vapnik (1995) is able to learn separating hyperplanes through Lagrangian optimization with strong convergence characteristics. They can use a non-linear projection to transfer the data into a higher-dimensional space to more easily find a classification boundary for more complex inputs Cortes and Vapnik (1995). Depending on the data and the task, computation can become very complex and require a great deal of domain knowledge Hastie et al. (2009). For example: in this thesis chest radiographs with a size of 224×224 of pixels are used. Taking into account that they are in gray scale (8 bit depth), this means there are 50176 features. With all second order terms, this would mean more than 25 billion terms that need to be fit. The features would be used to manually create new proxy features that would still be able to accurately represent a patient's condition, and be useful for fitting a model that could diagnose other x-rays. This is not only highly complex and, due to the domain, highly risky, but also highly subjective Nielsen (2015).

A more practical solution to estimate a classification rule that accounts for non-linearity

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

and is flexible enough to be used in high dimensions, could be to construct a more general and automatic approach where the model learns the most important features itself, directly from the data. This could help in learning better/more organic boundaries Goodfellow et al. (2016). Neural networks can be used to achieve just that kind of approach Nielsen (2015). On a high level, neural networks are a parametrized family of functions with differentiable parameters Berner et al. (2021). They provide powerful solutions for some of the issues faced when fitting traditional statistical models Goodfellow et al. (2016):

- There is no need for feature engineering Goodfellow et al. (2016)
- Compared to other modelling frameworks, deep learning can easily deal with complex types of data such as images, videos and text.
- Deep learning makes it possible to carry out efficient inference with a very high accuracy. For instance, Niu et al. (2019) implements a ResNet which needs 21ms to perform inference on images. This is especially useful for real time applications.

Neural networks form the basis of DL, the machine learning technique that has established itself as the dominant force in artificial intelligence Berner et al. (2021). Especially within the topic of image classification, deep learning has become the state of the art Berner et al. (2021). There exists a myriad of ways in which to define deep learning; this thesis refers to deep learning as techniques where *deep neural networks are constructed and subsequently trained with gradient-based methods Berner et al. (2021).* This type of definition allows for a clean distinction to energy based modelling later on. One of the most compelling characteristics is a network's capability to approximate *any* function Goodfellow et al. (2016), meaning a neural network will be able to learn a mapping that approximates the true class boundary to an arbitrarily small degree, no matter the shape or dimensionality Cybenko (1989). Deep learning enables a systematic, versatile and automatised approach to classification Nielsen (2015). The following section will put deep learning into historical context and briefly describe its development over the last decades.

3.3 Historical Evolution

Before formally defining the neural network model, the details of its architecture and the extension to deep learning, it is important to know its development over time. This helps to understand why it has its specific structure, how this helps in approximating any function and how this enables us to construct highly accurate automated classification systems.

Generally speaking, there were three major waves of increased popularity in deep learning Goodfellow et al. (2016):

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

1940-1960 This era was dominated by trying to build biologically inspired discriminative models Goodfellow et al. (2016). This was triggered by the works of McCulloch and Pitts McCulloch and Pitts (1943), who built a logical inference machine inspired by the inference capacity of the brain: the artificial neuron. These accepted binary inputs and were aggregated and compared to a pre-defined threshold parameter. If their aggregation exceeded the threshold, then the neuron would fire, i.e. return 1. Otherwise the neuron would remain inactive. By using the artifical neuron, it became possible to model simple boolean functions such as AND and OR. While models were influenced by the neuroscientific perspective and the function of the brain, they are **not** meant to be realistic representations of biological function Goodfellow et al. (2016). The mid-1950s saw the development of the first linear models, chief among them the revolutionary Rosenblatt Perceptron Rosenblatt (1958). The perceptron is a binary supervised learning algorithm that was inspired by the artificial neuron forms the basis of modern network models Goodfellow et al. (2016). It takes binary inputs $x \in \{0,1\}^d, d \in \mathbb{N}^+$, with d indicating the dimension, and assigns each one with a respective weight $w \in \mathbb{R}^d$, indicating that input's overall influence on the final result. A weighted sum is calculated and the binary output is determined by comparing the sum to a pre-defined threshold. The model can be visualised in figure 3.1.

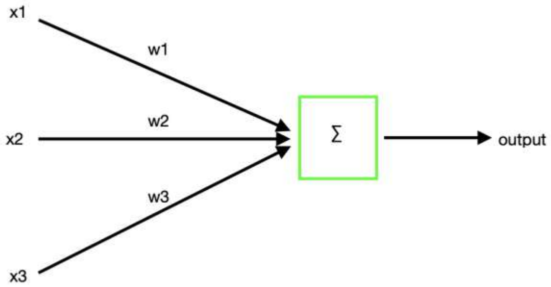


Figure 3.1: The perceptron model visualized

Training is done by adjusting the weights to minimize the difference between output and the true label. Perceptrons are not limited to one layer, they can be constructed with multiple layers by linking multiple neurons together, creating the well-known fully connected architecture. This gave rise to the extension of the perceptron, the Multilayer Perceptron (MLP) Goodfellow et al. (2016). However, perceptrons in their original form were binary and only worked on linearly separable data, meaning it could not account for more complex domains and problems. This becomes evident when considering a simple example: using the perceptron to learn the XOR function (exclusive or). This function is an operation on two binary values, x_1 and x_2. When exactly one of these binary values is equal to 1, the XOR function returns 1, otherwise 0. A linear model is not capable to learning a mapping to imitate this; its positive and negative instances cannot be separated by a line or hyperplane. For instance: When $x_1 = 0$, the model's output must increase as x_2 increases. When $x_1 = 1$, the model's output must decrease as

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

x_2 increases. However, a linear model must apply a *fixed* coefficient w_2 to x_2. In addition, due to the binary nature of the perceptron and other models of that time, gradient-based methods for the optimization of the weights could not be used, since the computation of derivatives necessary for these methods was not possible Goodfellow et al. (2016). This reason, among others, meant that the machines of that time were incapable of carrying out the calculations necessary for large networks to work well. Therefore, most work remained small-scale or theoretical and research activity soon waned.

1980-1995 Traditionally, the weights and thresholds would have been adjusted manually; however, due to the improvements in technology, automatic tuning algorithms were being created. The algorithms relied on differentiation techniques, which the Rosenblatt perceptron did not account for. Thus, research moved away from the binary nature of early models and started using differentiable and nonlinear activation functions instead, meaning the weighted sum in equation 3.2 became the input to a nonlinear function instead of the indicator function. The structure is visualised in figure 3.2

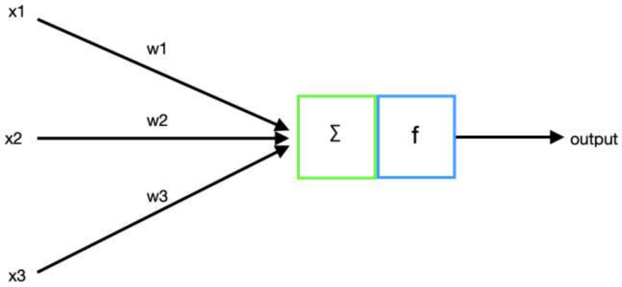


Figure 3.2: The artificial neuron including a continuous, nonlinear activation function f. Inputs x_i are weighted by weights w_i and added together before being fed into the activation function.

While many different activation functions are possible, the sigmoid function was, and still is, widely used for activation, because it can be seen as a smooth version of the indicator function used in the Rosenblatt Perceptron and thus acts as a natural extension for MLPs. The inclusion of different types of nonlinear activation functions allowed the models to learn more complex functions and process more complex input features Goodfellow et al. (2016). The term Artificial Neural Network (ANN) became popular to describe the class of models that is built up in a network fashion, often with fully connected neurons, with the MLP becoming the quintessential ANN. A major milestone in this era was the formulation and proof of the universal approximation theorem for ANNs that used sigmoid activation functions by Cybenko Cybenko (1989).

Interest for neural networks significantly re-emerged when Paul Werbos discovered an efficient gradient-based method to train networks through propagating errors back through

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

the layers of the network and modifying the weights at each node to reduce the overall error Werbos (1974). The algorithm makes use of the continuous and differentiable activation functions and computes the gradient of the loss function with respect to each weight using the **chain rule**. The chain rule helps to identify how much each weight contributes to the overall cost and the direction to update each weight to reduce it. It works one layer at a time from the last layer to avoid redundant calculations of intermediate terms, thus increasing efficiency Goodfellow et al. (2016). Yann LeCun published the first successful implementation of the back-propagation algorithm in neural network training LeCun (1987).

1990-1999 this decade saw significantly lower levels of broad Deep Learning research. Despite this, important breakthroughs were still being achieved, such as the development of Convolutional Neural Networks (CNNs) by Yann LeCun LeCun et al. (1998) and Long Short-Term Memory Networks (LSTMs) Hochreiter and Schmidhuber (1997a). CNNs excel at computer vision tasks and were particularly important for the advancement of deep learning. They use the convolution operation and specific aggregations called poolings to take into account that neighbouring pixels often share information. In other words, CNNs have locally receptive fields, meaning they collect information jointly from spatially close inputs Berner et al. (2021). In addition, they are invariant under translation allowing them to share parameters, which makes training a lot more efficient Goodfellow et al. (2016). A CNN is built by stacking multiple convolutional blocks, layers consisting of the convolution and pooling operations, on top of a fully connected architecture known from ANNs that produces the class estimates. Conceptually speaking, CNNs enable efficient learning based on different levels of composition Goodfellow et al. (2016), meaning that multiple units, or neurons, work together to identify low-level features in the data and, when combined, the network is able to make a decision based of these combined high-level features. These breakthroughs set the scene for the upcoming third wave of popularity.

2006-Present Improved hardware, mainly the development of GPUs, more structured and readily available data (ImageNet, MNIST) as well as more efficient algorithms Hinton et al. (2006b) led to increased research activity and a significant improvement to existing models. The general trend was to increase the number of layers and go deeper, while incorporating sophisticated mechanisms to be as efficient as possible Goodfellow et al. (2016). A trailblazing network implementation was the CNN AlexNet Krizhevsky et al. (2017). AlexNet won the ImageNet competition in 2012 by an 11% margin with 85% accuracy. It impressively demonstrated the benefits of CNNs and deeper architectures running on GPUs. This became the advent of deep learning as we know it today: very deep networks that are capable to operate in very high dimensional spaces in sensitive and complex domains, capable of processing the compositional structure of natural data.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

3.4 The Artificial Neural Network

The Artificial Neural Network (ANN) represents a large class of models that are built up as a network of inter-connected units, called neurons. The term is very broad and has become synonymous with many different concepts within neural network theory, such as individual networks and training methodologies Goodfellow et al. (2016). The preceding historical contextualisation 3.3 showed how the ANN concept grew from the perceptron Rosenblatt (1958), by combining multiple layers and different activation functions in order to account for complex domains and features. These networks are considered an important building block of many more sophisticated architectures, enabling the development of the deep architectures currently seen in today's research. The most recognizable architecture is the fully connected architecture , where all neurons in one layer are connected to all neurons in the subsequent layer.

A class of ANNs that encompasses many of today's essential neural network implementations is feedforward networks Goodfellow et al. (2016). They are characterized by a directed weighted graph where the information travels only in one direction: forward Goodfellow et al. (2016). The information travels from input node to output node with no cycles or loops and each neuron has directed connections to the neurons in the next layer Goodfellow et al. (2016). The standard and original example of an ANN is the MLP; ANNs and MLPs are often seen as synonyms. Another important example of a feedforward network for the context of this thesis is the Convolutional Neural Network (CNN) LeCun et al. (1998) Goodfellow et al. (2016). Feedforward networks stand in contrast to another important type of ANN: the Recurrent Neural Network (RNN) and its extension the Long Short-Term Memory Network (LSTM) Hochreiter and Schmidhuber (1997b). These are derived from feedforward networks and can be used to process inputs of varying lengths and are particularly useful in speech recognition Hochreiter and Schmidhuber (1997a) Abiodun et al. (2018). They are characterized by allowing cycles in the flow of information, meaning they keep information from prior inputs to influence the current input and output Goodfellow et al. (2016). RNNs will not be a subject of this thesis.

3.4.1 Definitions

The fundamental component of the Artificial Neural Network is the artificial neuron. Researchers McCulloch and Pitts developed a mathematical formulation that modelled the functionality of real biological neurons McCulloch and Pitts (1943). The core logic behind their work was that neurons communicate with each other by being activated; the activation of a neuron is binary and depends on whether the combined inputs are equal or larger than a threshold McCulloch and Pitts (1943) Goodfellow et al. (2016). This logic is summarised in definition 3.1 McCulloch and Pitts (1943) Petersen (2020):

Definition 3.1 (McCulloch and Pitts Neuron)**.** Let $x \in \{0,1\}$ be a binary input vector; $\mathbf{1}_{\mathbb{R}^+} : \mathbb{R} \to \mathbb{R}$ is the indicator function with $\mathbf{1}_{\mathbb{R}^+}(x) = 0$ for $x < 0$ and $\mathbf{1}_{\mathbb{R}^+}(x) = 1$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

everywhere else; furthermore, $w_i \in \{-1, 1\}$ are weights for $i = 1, \ldots, d$, where $d \in \mathbb{N}$ is the number of inputs; $\theta \in \mathbb{R}$ is a pre-defined threshold, then the artificial neuron is defined as

$$x \mapsto \mathbf{1}_{\mathbb{R}^+}\left(\sum_{i=1}^{d} w_i x_i - \theta\right)$$

The threshold determines whether the neuron in question is activated or not. If the weighted sum of the inputs exceeds the threshold the neuron is activated, otherwise not. Hence the name activation function Goodfellow et al. (2016). Frank Rosenblatt extended the definition of the neuron, making it more flexible and expressive Rosenblatt (1958). The key difference to the artificial neuron was that he introduced the usage real-valued weights. This gave the weights actual meaning because it expressed how important the corresponding input was for the output, this was not possible in the McCulloch Pitts framework. The Rosenblatt Perceptron is defined in definition 3.2:

Definition 3.2 (Perceptron)**.** Let $x \in \{0, 1\}$ be a binary input vector; $\mathbf{1}_{\mathbb{R}^+} : \mathbb{R} \to \mathbb{R}$ is the indicator function with $\mathbf{1}_{\mathbb{R}^+}(x) = 0$ for $x < 0$ and $\mathbf{1}_{\mathbb{R}^+}(x) = 1$ everywhere else; furthermore, $w_i \in \mathbb{R}$ for $i = 1, \ldots, d$ are the randomly initialized weights, where $d \in \mathbb{N}$ is the number of inputs; $\theta \in \mathbb{R}$ is the pre-defined threshold, then the Rosenblatt Perceptron is defined as

$$\text{classification output} = \begin{cases} 1 & \text{if } w' \cdot x + \theta \leq 0 \\ 0 & \text{if } w' \cdot x + \theta > 0 \end{cases} \tag{3.1}$$

The Perceptron is also known as a linear discriminator Hastie et al. (2009). On a high level, it can be thought of as a decision-making device that weighs up the evidence available to it Nielsen (2015). By connecting neurons such that the output of one becomes the input of another, the familiar fully connected network structure can be constructed, see figure 3.3. The model for this network is the Multilayer Perceptron (MLP) model, see 3.3 Petersen (2020), and is the simplest form of the Artificial Neural Network. By combining the decision making capabilities of all the perceptrons, this structure should be able to make subtle decisions, conceptually speaking Nielsen (2015).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

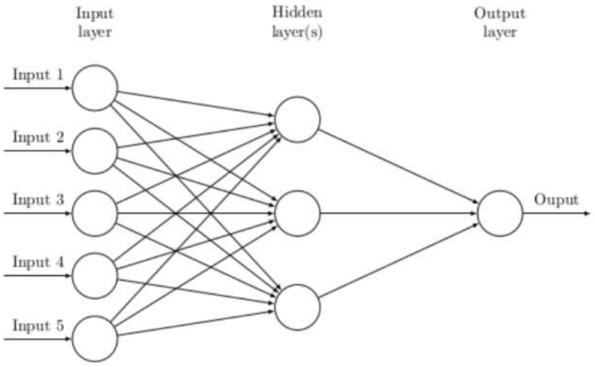


Figure 3.3: The multilayer perceptron model

Traditionally, the weights and thresholds would have been adjusted manually; however, modern day training algorithms tune are able to tune these automatically. The specifics of these algorithms will be covered in later sections, but a key component for training and improving the MLP is the knowledge of how the output changes following incremental changes in the weights, i.e. the derivative W.R.T. the weights. However, the binary nature of the perceptron does not allow for this, since the indicator function is not differentiable. Thus, the MLP model was extended to allow differentiable, non-linear activation functions, most prominently the sigmoid function which can be seen as a smoother version of the indicator function Nielsen (2015). Sigmoid functions make up a class of functions defined by Petersen (2020):

Definition 3.3 (Sigmoid function)**.** A continuous function $f: \mathbb{R} \to \mathbb{R}$ such that $f(x) \to 1$ for $x \to \infty$ and $f(x) \to 0$ for $x \to -\infty$ is called sigmoidal.

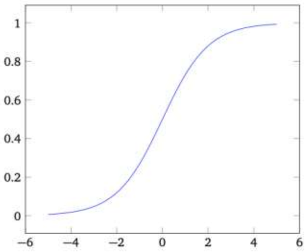


Figure 3.4: A sigmoidal function according to definition 3.3 Nielsen (2015) Petersen (2020)

A sigmoid function, shown in figure 3.4, is differentiable and has a non-negative derivative at each point with exactly one inflection point Han and Moraga (1995); it is often synonymous with the logistic function. It allows for a broader set of inputs, any real number, and, since it "squashes" its input to be between 0 and 1, it lends itself to

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

express likelihoods Goodfellow et al. (2016). In addition, being monotonically increasing, continuous, and differentiable at every point, allows for the use of gradient-based methods of optimization.

The preceding definitions can be used to further formally define the Multilayer Perceptron, the quintessential neural network and building block of more complex architectures seen later. Most importantly, it forms the centerpiece of one of the most important aspects of neural networks overall: universal approximation , see 3.4.2.

Definition 3.4 (Mulitlayer Perceptron)**.**

$$\begin{aligned} G\colon \mathcal{A} &\to \{0,\dots,k-1\},\\ x &\mapsto A_L\varphi_{L-1}(A_{L-1}\varphi_{L-2}(\dots\varphi_1(A_1(x))),\\ &k, L \in \mathbb{N},\\ &L \geq 2 \end{aligned}$$

be a *network* where

- $\mathcal{A}$ is the set of input data
- $$\begin{aligned} A_l : \mathbb{R}^{N_{l-1}} &\to \mathbb{R}^{N_l},\\ x &\mapsto wx + \theta,\\ 1 &\leq l \leq L, \quad (N_0 = n \in \mathbb{N}, N_L = k) \end{aligned}$$

  are affine mappings with weight matrix $w \in \mathbb{R}^{N_{l-1} * N_l}$ and bias $\theta \in N_l$.
- $\varphi_i : \mathbb{R} \to \mathbb{R}, \quad 1 \leq i < L$ continuous activation functions that act coordinatewise. As an extension to the Perceptron, these functions are arbitrary.
- L : the number of layers (also called depth) and N_L the dimension of the L-th layer (also called width)

The network according to definition 3.4 produces class labels $\hat{b} \in \{1,\dots,k\}$ represented as integers. Alternatively, the function G could give a probability distribution over the classes. In this case, the range would change to:

$$\{\hat{b} \in \mathbb{R}^k \colon 0 \leq \hat{b}[r] \leq 1,\ r \in \{1,\dots,k\} \wedge \Sigma_{i=1}^{k} y[i] = 1\}$$

That is, the output $\hat{b}$ of the function G would be a $k-$dimensional vector where the $r-$th element, corresponding to the $r - th$ semantic label, lies between 0 and 1, and whose elements sum to 1. The predicted label would then be the maximum element of $\hat{b}$

Network G is defined as a combination of affine, and non-linear functions, parametrized by a set of weights w. G characterizes a whole set of networks:

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Definition 3.5 (Set of networks)**.**

$$\zeta \coloneqq \{G \colon \text{G Network as defined in definition 3.4}\}$$

3.4.2 Universal Approximation

In this section the universal approximation theorem for neural networks is summarized. This theorem was originally formulated and proven by George Cybenko in 1989 Cybenko (1989). Universal approximation means that, under minor conditions, every continuous function on a compact set can be arbitrarily well approximated by a network $G \in \zeta$. This is of major importance because it means that a randomly initialized network can be trained to represent unknown structures that are modelled by some unknown mapping or distribution Goodfellow et al. (2016). For classification, this means that the *true* class boundary for the images, represented by some function f, can be approximated by a network G to an arbitrarily small degree. Universal approximation is what allows networks to be the powerful tools they have become, since any task that can be thought of as a function computation can be performed/computed by the neural networks – be it language translation, caption generation, speech to text, etc Goodfellow et al. (2016).

Cybenko's proof focuses only on the case of sigmoidal functions, see definition 3.3, and uses key results from functional analysis. He proved universal approximation by showing that the set ζ of all networks lies *densely* in $C(K)$, meaning that any continuous function $f \in C(K)$ has a network $G \in \zeta$ that is arbitrarily close. Effectively, this means that for an arbitrary precision $\epsilon > 0$ a neural network G exists that will approximate a continuous function f. More formally:

Definition 3.6 (Universality)**.** Let $\varphi \colon \mathbb{R} \to \mathbb{R}$ be a continuous activation function as in definition 3.4, $d, L \in \mathbb{N}$ and K and ζ defined according to the notation above. ζ is *universal*, if it is *dense* in $C(K)$.

However, in order to show definition 3.6, a relevant topology for the networks to operate on needs to be defined. The notation section 3.1 and definitions 3.4 and 3.5, lay out the most important components of this. However, additional conditions must be imposed on the activation function(s) φ:

Definition 3.7 (Discriminatory functions)**.** φ, continuous as in definition 3.4, is *discriminatory* if for $\mu \in \mathcal{M}$, $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ compact and

$$\int_K \varphi\left(wx + \theta\right) d\mu\left(x\right) \quad = \quad 0$$

for all $w \in \mathbb{R}^d, \quad \theta \in \mathbb{R}$ then $\mu = 0$

The discriminatory property of a function means that that for nonzero μ, there exist w and θ such that the integral in definition 3.7 is non-zero. Intuitively, this means that

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

no information from the input to the function is "lost" by being mapped to a subset of measure 0.

With these definitions the correct topological setting for the concrete formulation of the universal approximation theorem and its proof have been established.

Theorem 3.8 (Universal Approximation - Cybenko)**.** *Let* $d \in \mathbb{N}$*,* K *a compact subset,* $K \subset \mathbb{R}^d$ *and* $\varphi \colon \mathbb{R} \to \mathbb{R}$ *be discriminatory. Then the set of all networks,* ζ *defined in 3.5, is universal.*

Using the condition on activation functions, the universal approximation theorem can be proven Petersen (2020), Cybenko (1989).

Proof. Observe that $\Gamma \subseteq C\left(K\right)$ is a linear subspace of $C\left(K\right)$. Given 3.4, any network $G \in \Gamma$ is a composition of a series of continuous activation functions φ_i and affine functions A_l, thus any G is also continuous and therefore Γ a linear subspace of $C\left(K\right)$.

In order for Γ to be dense in $C\left(K\right)$, the following equality needs to hold for its closure $\overline{\Gamma}$:

$$\overline{\Gamma} = C\left(K\right)$$

By way of contradiction, suppose that $\overline{\Gamma} \neq C\left(K\right)$. Then $\overline{\Gamma}$ is a closed, proper subspace of $C\left(K\right)$.

By the Theorem of Hahn-Banach Rudin (1974), there exists a bounded linear functional $F \neq 0$ on $C\left(K\right)'$ (the dual space of $C(K)$), such that $F(\Gamma) = F(\overline{\Gamma}) = 0$ ($\overline{\Gamma} = \Gamma$ since Γ is closed).

By the Riesz Representation Theorem Rudin (1987), this functional is of the general form

$$F(h) = \int_K h(x) d\mu(x)$$

for some $\mu \in \mathcal{M}$, $\forall h \in C(K)$.

The Riesz Representation Theorem establishes that $\mathcal{M} = C(K)'$ is the relevant dual space, thus the functional is a signed, regular borel measure μ. This, in turn, implies that the following equality must hold, since φ was assumed discriminatory:

$$\int_K \varphi\left(wx + b\right) d\mu\left(x\right) \quad = \quad 0$$

which implies that $\mu = 0$ and thus $F = 0$, contradicting our assumption that $H \neq 0$. Hence, the subspace Γ must be dense in $C\left(K\right)$. □

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

This proves that all discriminatory activation functions lead to spaces of networks that can universally approximate every continuous function on a compact set K.

However, discriminatory functions are hard to identify directly, which is why it is useful to identify conditions that are more accessible but at the same time guarantee that the function is discriminatory Petersen (2020). At this point, the role of the the sigmoidal family of functions, defined in 3.3, becomes important. Cybenko proved that sigmoidality is a sufficient condition on the activation functions that still guarantees universal approximation and makes working with networks a lot easier Cybenko (1989). The following lemma and proof will formalise this Cybenko (1989):

Lemma 3.9. Any bounded, measurable *sigmoidal* function f is discriminatory.

Proof. Let $d \in \mathbb{N}$, $x, w \in \mathbb{R}^d$, $\lambda, \theta, \nu \in \mathbb{R}$, $K \subset \mathbb{R}^d$ compact and measure $\mu \in \mathcal{M}$.

In order to fulfil the definition for discriminatory functions 3.7, it must be shown that assuming for $f : \mathbb{R} \to \mathbb{R}$ sigmoidal, defined in 3.3, it holds that:

$$\int_K f(w'x + \theta))d\mu(x) = 0, \quad \forall w \in \mathbb{R}^d, \theta \in \mathbb{R}$$

then $\mu = 0$.

Consider the function

$$\gamma(x) = \lim_{n\to\infty} f_\lambda(x) = \lim_{n\to\infty} f(\lambda(w'x + \theta) + \nu) = \begin{cases} 1 & \text{for } w'x + \theta > 0 \\ 0 & \text{for } w'x + \theta < 0 \\ f(\nu) & \text{for } w'x + \theta = 0 \end{cases}$$

Let $\Pi_{w,\theta} \coloneqq \{x \colon w'x + \theta = 0\}$ be an affine hyperplane and let $H^+_{w,\theta} \coloneqq \{x \colon w'x + \theta > 0\}$ and $H^-_{w,\theta} \coloneqq \{x \colon w'x + \theta < 0\}$ be the half open spaces defined by $\{w'x + \theta > 0\}$ and $\{w'x + \theta < 0\}$, respectively. Note that f is bounded since $|f_\lambda(x)| \leq \max(1, f(\nu))$ for all x and K is compact.

Applying the Dominated Convergence Theorem Bartle (1995) gives

$$\begin{aligned}
\lim_{\lambda\to\infty} \int_K f_\lambda(x)\ d\mu(x) &= \int_K \lim_{\lambda\to\infty} f_\lambda(x)\ d\mu(x) \\
&= \int_K \gamma(x) d\mu(x) \\
&= \int_{H^-_{w,\theta}} 0 d\mu(x) + \int_{\Pi_{w,\theta}} f(\nu) d\mu(x) + \int_{H^+_{w,\theta}} d\mu(x) \\
&= 0 + f(\nu)\mu(\Pi_{w,\theta}) + \mu(H^+_{w,\theta}) \\
&= 0
\end{aligned}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

for all ν, θ, w. If $\nu \to \infty$ then, using the properties of sigmoidal functions, $f(\nu) \to 1$ and $\mu(\Pi_{w,\theta}) + \mu(H_{w,\theta}) = 0$. Similarly, if $\nu \to -\infty$ then $f(\nu) \to 0$ and $\mu(H_{w,\theta}) = 0$.

Consider w fixed and a step function h. Using the Riesz Fisher Representation theorem Rudin (1987) the corresponding linear functional $F \colon L^\infty(\mathbb{R}) \to \mathbb{R}$ has the following form:

$$F(h) = \int_K h(w'x)\ d\mu(x)$$

Note that F is a bounded linear functional on $L^\infty(\mathbb{R})$ since μ is a finite signed measure, and is well-defined for any $h \in L^\infty(\mathbb{R})$. This is because when we integrate with respect to a finite measure, there cannot be an infinite result.

Consider h to be the indicator function for the interval $[\theta, \infty)$ such that

$$F(h) = \int_K h(w'x) d\mu(x) = \mu(\Pi_{w,\theta}) + \mu(H^+_{w,\theta}) = 0$$

Where the indicator function $\mathbf{1} \colon X \to \{0, 1\}$ for a set $A \subseteq X$ is defined as:

$$\mathbf{1}(x) := \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

Similarly, $F(h) = 0$ if h is the indicator function for the open interval (θ, ∞). By linearity, F is 0 for the indicator function on any interval, hence for any step function. Step functions are dense in $L^\infty(\mathbb{R})$, so $F = 0$ for every bounded continuous function h.

In particular, the bounded, measurable functions $h(x) = \sin(wx)$ and $h(x) = \cos(wx)$ give

$$F(a + ib) = \int_K = \cos(m'x) + i \sin(m'x)\ d\mu(x) = \int_K e^{im'x}\ d\mu(x) = 0$$

for all m. Thus, the Fourier transform of μ is 0 and so μ must be zero as well Rudin (1987).

Therefore f is discriminatory by 3.7. □

Remark. A key element of this proof is the choice of a specific function h that drives the measure μ to zero. This was shown to be the Fourier transform.

Remark. The proof shown above only accounts for sigmoidal activation functions. Shortly after Cybenko's proof of the the universal approximation theorem, Hornik was able to extend the theorem to a broader class of activation functions by using arbitrary non-linearities Hornik et al. (1989).

Remark. The proof is not constructive, i.e. it gives no way to explicitly construct or find such a neural network or choose its architecture

Remark. A remark on notation: the work thus far has shown that a network G is parametrized by weights w. Therefore, networks will be written as G_w in the following sections.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

3.5 Training Deep Neural Networks

The preceding sections have motivated and defined the underlying mathematical concepts of neural networks classifiers. Deep learning was characterised in 3.2 as techniques where deep neural networks are trained with gradient based methods, where depth is achieved by stacking many layers of neurons on top of each other Berner et al. (2021). Increasing depth can greatly increase a neural network classifier's performance, however it also increases their mathematical complexity and computational intensity, making training much harder Goodfellow et al. (2016). Thus, efficient training methods are necessary and will be explored below.

The definition a neural network classifier however, does not give any details about the concrete parts that actually make up an implementation and how it is trained. Goodfellow et al. (2016) describes all machine learning algorithms as a combination of:

1. **a specific dataset:** the dataset
2. **a model:** section 3.4.1 lays out the mathematical foundation of the neural network model, on which the remaining specific components will now be built.
3. **a cost function:** training involves continuously evaluating the model's predictions. They are measured by the *error*, *loss* or *cost* function, which expresses the deviation between the model's predictions and the corresponding ground truth.
4. **an optimization procedure:** training a supervised model means that the model's variable components that parametrize the model, the weights in this case, are continuously adjusted based on the behaviour of the cost function. Since the goal is to produce predictions that are highly accurate, training means that the cost function is minimized. How this happens specifically is expressed by the optimization procedure.

For the following sections $\mathcal{C} \coloneqq \{(a_1, b_1), \dots, (a_n, b_n)\}$, will be referred to as the training data. The element b_i represents the $i-th$ input's a_i label.

Cost Functions

In supervised learning, the cost function measures the discrepancy between the prediction a machine learning algorithm produces and the corresponding ground truth. It always contains a term, usually a type of parameter ω, that makes the learning process, a process of statistical estimation Goodfellow et al. (2016). Define the cost function as:

$$J : \Omega \to [0, +\infty]$$

where $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$ represents the set of all possible parameter vectors.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The choice of a cost function largely depends on the type of problem addressed by the model Goodfellow et al. (2016). For classification, the neural network G is trained to produce a class label that *most likely* corresponds to a given input image. This can be reformulated as maximising the conditional probability $p(b|a)$ of a label b given an input a, where the conditional probability is expressed using likelihoods parametrized by the network weights w. This means that training a classifier can be interpreted as a Maximum Likelihood estimation problem, where the weights are optimised to yield the highest probability Fisher (1925) Goodfellow et al. (2016). It is intuitively clear that finding the most likely label, i.e. maximising the conditional probability, is equivalent to minimizing the cost between the network's prediction and the image's true label, meaning the optimal parameter estimates ω of J are the same as the optimal weights w of G. Since maximising a function is equivalent to minimizing the negative of that function, cost can be expressed by the negative likelihood. Thus, minimizing cost amounts to minimizing the negative likelihood or, more simply, the Negative Log Likelihood (NLL). In short, training a neural network classifier is a problem of Maximum Likelihood and can be accomplished by minimizing the NLL loss W.R.T. the network's weights. The NLL loss is one of the most popular cost functions for classification, regardless of the modelling algorithm used.

The considerations above intuitively describe how the choice of the NLL as a cost function is an organic fit to this classification task. The underlying statistical theory showing why its use is justified and what its benefits will be shown. Fisher formalised the concept of likelihoods and likelihood functions as a tool to carry out inference on unknown population parameters, based on a known sample Fisher (1992). He explicitly distinguishes the concept of likelihood from the concept of probability, the difference can be seen in the following definition Shao (2006):

Definition 3.10 (Likelihoods and Likelihood Functions)**.** Given a parametrized family of probability density functions with parameter ω, that is a collection of functions f characterized by the parameter ω that indicate the relative likelihood of a random variable being within a range of values.

$$I \mapsto f(a \mid \omega)$$

The likelihood function Λ is defined as a function of the unknown parameter, given the data which is known:

$$\omega \mapsto f(a \mid \omega) =: \Lambda(\omega \mid a)$$

In other words, when $f(a \mid \omega)$ is viewed as a function of a with ω fixed, it is a probability density function, and when viewed as a function of ω with a fixed, it is a likelihood function.

A key theorem used here is the Radon–Nikodym Theorem Billingsley (1986) Bartle (1995). It states that the probability (measure) can be expressed as the integral of a density

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

function over a subset of possible outcomes, W.R.T. to a dominating measure, typically the Lebesgue measure. The density is also called the Radon–Nikodym derivative.

Remark. The above definition is based on results from measure theory. The Radon–Nikodym Theorem is a key tool to extend probability densities to probability measures Billingsley (1986).

Remark. The likelihood can be seen as the joint probability between the data and the parameters of the underlying statistical model, or in other words the model distribution.

To justify using probability density functions to specify the likelihood function the following steps need to be considered Billingsley (1986):

Proof. For an observation a, the likelihood for the interval $[a, a+h]$, where $h > 0$ is a constant, is given by $\Lambda(\omega \mid a \in [a, a+h])$. This means that:

$$\underset{\omega}{arg\,max} \quad \Lambda(\omega \mid a \in [a, a+h]) = \underset{\omega}{arg\,max} \quad \frac{1}{h}\Lambda(\omega \mid a \in [a, a+h])$$

since h is positive and constant. Because

$$\begin{aligned} \underset{\omega}{arg\,max} \quad \frac{1}{h}\Lambda(\omega \mid a \in [a, a+h]) &= \underset{\omega}{arg\,max} \quad \frac{1}{h}\mathbb{P}(a \leq a_j \leq a+h \mid \omega) \\ &= \underset{\omega}{arg\,max} \quad \frac{1}{h}\int_a^{a+h} f(a_j \mid \theta) da_j \end{aligned}$$

where $f(a_j \mid \omega)$ is the probability density function, it follows that

$$\underset{\omega}{arg\,max} \quad \Lambda(\omega \mid a_j \in [a, a+h]) = \underset{\omega}{arg\,max} \quad \frac{1}{h}\int_a^{a+h} f(a_j \mid \omega) da$$

Using the fundamental theorem of calculus Spivak (1967) and the l'Hospital's rule de L'Hospital (1696) it follows that:

$$\begin{aligned} \lim_{h\to 0^+} \frac{1}{h}\int_a^{a+h} f(a_j \mid \omega) da &= \lim_{h\to 0^+} \frac{\frac{d}{dh}\int_a^{a+h} f(a_j \mid \omega) da}{\frac{dh}{dh}} \\ &= \lim_{h\to 0^+} \frac{f(a+h \mid \omega)}{1} \\ &= f(a \mid \omega). \end{aligned}$$

Then

$$\begin{aligned} \underset{\omega}{arg\,max} \quad \Lambda(\omega \mid a) &= \underset{\omega}{arg\,max} \quad \left[\lim_{h\to 0^+} \Lambda(\omega \mid a_j \in [a, a+h])\right] \\ =\underset{\omega}{arg\,max} \quad \left[\lim_{h\to 0^+} \frac{1}{h}\int_a^{a+h} f(a_j \mid \omega) da\right] &= \underset{\omega}{arg\,max} \quad f(a \mid \omega). \end{aligned}$$

Therefore

$$\underset{\omega}{arg\,max} \quad \Lambda(\omega \mid a) = \underset{\omega}{arg\,max} \quad f(a \mid \omega)$$

□

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

The reasoning above can be used to formalise the basic idea behind Maximum Likelihood: $f(a \mid \omega)$ is maximised W.R.T. to the parameter ω and for a fixed input a. The resulting maximiser, also known as the Maximum Likelihood estimator, is the parameter value that results in the distribution that delivers the highest probability for that input. The parameters are estimated based on known quantities, hence the term estimator. More specifically, an estimator is a rule for calculating an estimate of a given quantity based on known data Tukey and Frederick (1965 - 1986). While this method makes intuitive sense, it also provides estimators with important statistical properties Fisher (1992), namely:

- consistency: for $n \to \infty$, where n is the number of samples, the estimator converges in probability to its true value Fisher (1992).
- efficiency: the variance of the estimator converges for $n \to \infty$ to the Cramer-Rao bound and is thus the lowest possible variance Cramér (1999).

In other words, the maximum likelihood estimator is, asymptotically, the best estimator, where "best" is to be understood as the estimator with lowest variance relative to the Cramer-Rao bound Cramér (1999). The Cramer-Rao bound characterizes the performance of an estimator; it describes a lower bound for the variance of estimators of the parameter ω.

In the context of classification, this means that we maximize over all *conditional probabilities*. Assuming the data is identically, independently distributed (IID), meaning every sample comes form the same distribution and all are mutually independent, the likelihood is expressed as:

$$\Lambda(\omega \mid a) = \prod \mathbb{P}_\omega \left(y_i \mid a_i\right) \tag{Likelihood}$$

or the log-likelihood:

$$log(\Lambda(\omega \mid a)) = \ell(\omega \mid a) = \Sigma_{i=1}^{n} log \left(\mathbb{P}_\omega \left(y_i \mid a_i\right)\right) \tag{Log-Likelihood}$$

The Maximum Likelihood estimators ω_{ML} are obtained my maximising Likelihood or, equivalently, Log-Likelihood.

$$\begin{aligned} \omega_{ML} &= \underset{\omega}{arg\,max} \quad \Lambda(\omega \mid a) \\ &= \underset{\omega}{arg\,max} \quad \prod_{i=1}^{n} \mathbb{P}_\omega \left(y_i \mid a_i\right) \\ &= \underset{\omega}{arg\,max} \quad \Sigma_{i=1}^{n} log \left(\mathbb{P}_\omega \left(y_i \mid a_i\right)\right) \end{aligned}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

The likelihood function can also be understood as an expression of the model distribution:

$$\ell(\omega \mid a) = p_{model}$$

Intuitively, the search for the Maximum Likelihood estimators is meant to deliver a model distribution p_{model} that best resembles the true underlying empirical data distribution, $\hat{p}_{data}$ Goodfellow et al. (2016). This implies that, in addition to finding the most suitable model parameters, it is necessary to quantify the difference between two distributions. This can be done using the statistical distance Kullback-Leibler Divergence Kullback and Leibler (1951). Thus, the term "best" can be understood as the model distribution p_{model} with *minimal divergence* to the true underlying empirical data distribution $\hat{p}_{data}$ Kullback and Leibler (1951).

Definition 3.11 (Kullback-Leibler Divergence)**.** The Kullback-Leibler Divergence is a statistical distance measure that describes how (dis-)similar two distributions are Kullback and Leibler (1951). The divergence is quantified as:

$$D_{KL}(\hat{p}_{data}||p_{model}) = \mathbb{E}_{a_i, y_i \sim \hat{p}_{data}} \left[log\left(\hat{p}_{data}\left(y_i \mid a_i\right)\right) - log\left(p_{model}\left(y_i \mid a_i\right)\right)\right]$$

While the intuition seems straightforward, the equivalence between maximising the likelihood and minimizing the Kullback-Leibler Divergence is not. The following proof will make the equivalence clear:

Proof. Consider the IID training data $\mathcal{S} = ((a_1, y_1), \ldots, (a_n, y_n))$, where the samples are generated according to a probability distribution, that is $\mathcal{S} \sim \hat{p}_{data}$. This distribution is estimated by finding the ω_{ML} that will maximize the likelihood p_{model}:

$$\begin{aligned}\omega_{ML} \Leftrightarrow & \underset{\omega}{argmax} \Lambda(\omega \mid \mathcal{S}) \\ \Leftrightarrow & \underset{\omega}{argmax} \prod_{i=1}^{n} p_{model} \\ \Leftrightarrow & \underset{\omega}{argmax} \sum_{i=1}^{n} log\left(p_{model}\right)\end{aligned}$$

Since maximising the log-likelihood is equivalent to minimising the negative log-likelihood:

$$\begin{aligned}\Leftrightarrow & \underset{\omega}{argmin} - \sum_{i=1}^{n} log\left(p_{model}\right) \\ \Leftrightarrow & \underset{\omega}{argmin} \frac{-1}{n} \sum_{i=1}^{n} log\left(p_{model}\right)\end{aligned}$$

this describes the expectation w.r.t $\hat{p}_{data}$ that assignes each sample the probability $\frac{1}{n}$. Therefore, using the law of large numbers Dekking et al. (2005) and letting $n \to \infty$:

$$\Leftrightarrow \underset{\omega}{argmin} \mathbb{E}_{a_i, y_i \sim \hat{p}_{data}} \left[-log\left(p_{model}\right)\right]$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

using the properties of the logarithm we have:

$$\Leftrightarrow \underset{\omega}{argmin} \mathbb{E}_{\hat{p}_{data}} \left[log \left(\frac{\hat{p}_{data}}{\hat{p}_{data}} \right) \right] - \mathbb{E}_{a_i, y_i \sim \hat{p}_{data}} \left[log \left(p_{model} \right) \right]$$
$$\Leftrightarrow \underset{\omega}{argmin} - \mathbb{E}_{\hat{p}_{data}} \left[log \left(\hat{p}_{data} \right) \right] - \mathbb{E}_{\hat{p}_{data}} \left[log \left(\hat{p}_{data} \right) - log \left(p_{model} \right) \right]$$

The expression above is known as cross entropy. The first term is the average amount of information contained in $\hat{p}_{data}$, otherwise kown as entropy. The entropy above is constant. Since monotonic transformations such as adding/multiplying by a constant do not change the maximiser, it follows that:

$$\Leftrightarrow \underset{\omega}{argmin} \mathbb{E}_{\hat{p}_{data}} \left[log \left(\hat{p}_{data} \right) - log \left(p_{model} \right) \right]$$
$$\Leftrightarrow \underset{\omega}{argmin} D_{KL} (\hat{p}_{data} || p_{model})$$

☐

Thus, minimizing the negative log-likelihood, or maximizing the log-likelihood, is asymptotically equivalent to minimizing the KL-Divergence, the difference between model and data distribution, and minimizing the crossentropy Goodfellow et al. (2016). For a classification problem, the likelihood is expressed using the conditional probabilities calculated form the model output. The formal expression for the negative log-likelihood cost function becomes:

$$J : \Omega \rightarrow [0, +\infty] \tag{3.2}$$
$$(\omega) \mapsto J(\omega) = -\mathbb{E}_{(a_i, y_i) \sim \hat{p}_{data}} \left[log \left(p_{model} \left(y_i \mid a_i ; \omega \right) \right) \right] \tag{3.3}$$
$$\tag{3.4}$$

where a_i and y_i are the $i-$th elements of the training data is the class output by the classification network as defined in 6.3

The workings in this section show that the learning process involves maximum likelihood estimation on the parameter ω which is induced by the minimization of the negative log-likelihood cost function. Looking at the expression 3.4, the parameter ω controls the conditional probabilities calculated from the network output; the network output is itself controlled by the weights w of the network, see 3.4. Consequently, the parameters in the cost functions represent the weights in the neural network, meaning that learning occurs by minimizing cost W.R.T. to the *weights* that parametrize the neural network. The final form of the cost function over the entire dataset can be expressed as:

$$J : \Omega \rightarrow [0, +\infty]$$
$$(w) \mapsto J(w) = -\mathbb{E}_{(a_i, y_i) \sim \hat{p}_{data}} \left[log \left(p_{model} \left(y_i \mid a_i ; w \right) \right) \right]$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Remark. Given that the parameter ω represents the weights w of the neural network G_w, the following sections will omit ω and only reference w to avoid confusion in the notation.

Remark. This thesis chooses to work with the Negative Log Likelihood cost function due to its statistical properties. Nevertheless, the choice of the cost function is, in principle, arbitrary and must be chosen according to the specific *task* and *input*.

Optimization

The preceding section constructed and formalised the principle of neural network training through minimization of the cost function. However, the formalisation alone does not give details as to how the function should be minimized. In general, the optimum of a function can be approximated with iterative methods or found in closed form algebraically Bertsimas and Tsitsiklis (1997), which depends on the structure and complexity of the function. Section 3.5 has established that the cost function depends on the weights w of the network, see equation 3.4, with deep networks typically having tens of millions of weights to optimize Niu et al. (2019). This means that the cost function has a very complex structure, with potentially many local optima in addition to the global optimum Goodfellow et al. (2016). Optimisation of this kind of functional landscape is not straightforward and solvers typically cannot efficiently deliver a closed form solution. Iterative methods can be successfully used to find approximate solutions; however, these methods come with the caveat of possibly getting stuck at a point of inflection or delivering only a local optimum Bertsimas and Tsitsiklis (1997).

Gradient based methods is a popular set of optimization techniques, especially when training deep learning models Ruder (2016). Central to these optimization methods is, as the name suggests, the gradient of a function

Definition 3.12. Let $f \coloneqq \mathbb{R}^n \to \mathbb{R}, n \in \mathbb{N}$ be a differentiable function. Given a point $x \in \mathbb{R}^n$, the gradient of the function at that point x is expressed by:

$$\nabla f(x) \coloneqq (\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x))$$

The gradient $\nabla f(x)$ at a point x indicates the direction of steepest ascent, while $-\nabla f(x)$ indicates the direction of the steepest descent, respectively. The main reason for the popularity of these methods is that the accurate and efficient computation of pointwise derivatives is possible using a method for efficient differentiation, often referred to as the backpropagation algorithm. Intuitively, these methods iteratively calculate a sequence of points, starting from a randomly initialized point, until a minimum is reached.

Backpropagation The challenge with (deep) neural networks is the calculation of all the partial derivatives of the cost function with respect to each individual weight. Directly calculating the derivatives for each weight is highly inefficient; backpropagation Rumelhart et al. (1986) enables a methodical calculation of these partial derivatives, making training

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

more efficient. On a high level, backpropagation calculates the derivative for the last layer, and uses it to inductively go "backwards" through the network, calculating the partial derivatives of each layer until it reaches the first layer of the network. Mathematically, backpropagation exploits the use of the multivariate chain rule Leibniz (2012).

Theorem 3.13. *Let* f *and* g *be differentiable, real-valued functions. If* $y = g(x)$ *and* $z = f(g(x)) = f(y)$*, then the chain rule states:*

$$\frac{\partial z}{dx} = \frac{dz}{dy}\frac{dy}{dx}$$

Generalizing this to the multidimensional case with $x \in \mathbb{R}^m$*,* $y \in \mathbb{R}^n$*,* g *maps from* $\mathbb{R}^m$ *to* $\mathbb{R}^n$*, and* f *maps from* $\mathbb{R}^n$ *to* $\mathbb{R}$*. If* $y = g(x)$ *and* $z = f(y)$*, then*

$$\nabla_{x^z} = \left(\frac{\partial y}{\partial x}\right)^T \nabla_{y^z}$$

Meaning that the gradient of z *with respect to* x *can be computed by multiplying the* $n \times m$ *Jacobian* $\frac{\partial y}{\partial x}$ *by the gradient of* z *with respect to* y*.*

The chain rule allows for the calculation of the derivative of a function that is a composition of other differentiable functions. This is important for training the network because: any of the network's layers is a composition of functions, see 3.4, thus the chain rule is directly applicable to calculate the gradient of a layer W.R.T. to its weights w. This also implies that there are several nested differentiations that need to be carried out, with many of the same calculations being carried out multiple times - this is highly inefficient, especially for deep neural networks. The backpropagation algorithm is designed to avoid the continued repetition the number of subexpressions and reuse the expressions that have already been calculated. It tells us how to incrementally adjust the weights in response to the difference between the generated and desired output vectors for each training example. The algorithm is described in the following pseudocode Goodfellow et al. (2016):

Backpropagation is often wrongly understood to mean the whole optimization process; however, it is merely a method of efficient differentiation for pointwise partial derivatives using the chain rule used *within* an optimization method Berner et al. (2021). There are countless ways that use backpropagation to optimize the cost function.

Gradient Descent Given a neural network G and training data $\mathcal{S}$, the cost function J describes, through its graph, a high-dimensional surface called a loss landscape Berner et al. (2021). The surface may have regions associated with lower cost values which resemble valleys of a landscape, if they are surrounded by regions of higher values of cost. Training the neural network includes the minimization of the cost function, which informally amounts to starting from an arbitrary point on the surface and finding a path to a minima Berner et al. (2021). Gradient descent is a broad class of optimisation techniques and the core optimization methodology in machine learning Du et al. (2017).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Algorithm 3.1: Backpropagation algorithm for a Multilayer Perceptron. The computation yields the gradients with respect to the parameters of each layer, starting from the output layer and going backwards to the first layer. These gradients can be interpreted as an indication of how each layer's output should change to reduce error.

1. For a sample (a_i, b_i), propagate the input a_i through the network to compute the outputs $(\hat{b}_{i_1}, \dots, \hat{b}_{i_n})$ (in topological order).
2. Compute the cost J_i and its gradient
$$\frac{\partial J_i}{\partial b_{i_n}}. \tag{3.5}$$
3. For each $j = n, \dots, 1$ compute
$$\frac{\partial J_n}{\partial w_j} = \frac{\partial J_n}{\partial b_{i_n}} \prod_{k=j+1}^{n} \frac{\partial b_{i_k}}{\partial b_{i_{k-1}}} \frac{\partial b_{i_j}}{\partial w_j}. \tag{3.6}$$
where w_j refers to the weights in node i_j.

The gradient of the objective function is used to identify the direction of the strongest incline/decline and accordingly moving the parameters in small steps in order to reach the optimum Cauchy et al. (1847). One iteration of the gradient descent algorithm can be expressed as:

$$x^{(t+1)} \longleftarrow x^{(t)} - \eta \nabla f(x^{(t)})$$

where η is the step size, otherwise known as the learning rate. The learning rate controls how large of a step to take in the direction of negative gradient so that we can reach a (local) minimum Lu (2022). For convex problems, gradient descent converges to an optimum and the rate of convergence can be precisely characterised Du et al. (2017). However, the optimisation involved when training a neural network is not a convex problem Du et al. (2017) and the algorithm may oscillate around a local minimum or even diverge, this is due to gradient descent not exploiting a function's *curvature*- information contained in the function's Hessian Matrix Goodfellow et al. (2016). Thus, many variants of gradient descent have been developed to improve this.

Stochastic Gradient Descent Gradient descent in its basic form typically uses the whole dataset for calculating the gradient each iteration - this becomes infeasible for large networks. Stochastic Gradient Descent (SGD) Ruder (2016) selects a random sub-sample from the dataset, calculates its gradient and determines the update of all the parameters according to the direction of that gradient. A general formulation of the problem is Berner et al. (2021):

In algorithm 3.2, if $D^{(k)}$ is deterministic, i.e. $D^{(k)} = \nabla r(\Theta^{(k-1)})$, it is the original gradient descent algorithm introduced above. In stochastic gradient descent $D^{(k)}$ is a random

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Algorithm 3.2: Stochastic Gradient Descent

```
Input: Differentiable function $r\colon \mathbb{R}^d \to \mathbb{R}, d \in \mathbb{N}$, sequence of step-sizes
  $\eta_k \in (0, \infty)$, $k \in [K]$, $\mathbb{R}^d$-valued random variable $\Theta^{(0)}$
output: Sequence of $\mathbb{R}^d$-valued random variables $(\Theta^{(k)})_{k=1}^{K}$
for $k = 1, \dots, K$ do
    Let $D^{(k)}$ be a random variable such that $\mathbb{E}[D^{(k)}|\Theta^{(k)}] = \nabla r(\Theta^{(k-1)})$
    Set $\Theta^{(k)} := \Theta^{(k-1)} - \eta_k D^{(k)}$
end
```

variable, reflecting the inherent stochastic nature involved when selecting the sub-sample for calculating the gradient. More concretely, a parameter update can be expressed as Berner et al. (2021):

$$\Theta^{(k)} := \Theta^{(k-1)} - \frac{\eta_k}{m} \sum_{s \in S} \nabla_w J(w)$$

where J is the cost function, w the network weights, m is the sub-sample size and the sub-sample S is selected uniformly at random. SGD is generally noisier than typical Gradient Descent, because of the randomness in its descent. It requires a higher number of iterations to reach the minima than gradient descent in its basic form, but it is still computationally much less expensive than typical Gradient Descent Goodfellow et al. (2016).

An important issue Stochastic Gradient Descent faces is the fact that one of the most important hyperparameters, the learning rate, must be set a-priori and remains fixed Ruder (2016). The learning rate is the step size at every iteration of the optimization; intuitively, the gradient determines the direction of the strongest decline but the learning rate determines how large the step in that direction will be, thus representing how fast the machine *learns* Goodfellow et al. (2016) - the setting of the learning rate is, inherently, a trade-off Goodfellow et al. (2016). Extensions to SGD implement adaptive methods to adjust the learning rate while searching for the minima. This thesis makes use of the Adaptive Moment Estimation (ADAM) Kingma and Ba (2014) optimizer. In this optimization algorithm, running averages of both the gradients and the second moments of the gradients for every parameter are used. Next to its adaptive capabilities, it has attractive convergence behaviour.

For convex cost functions, basic gradient descent and SGD guarantee convergence to a global minimum Berner et al. (2021). SGD has a convergence rate of $\frac{1}{\sqrt{K}}$. However, typically the cost function is not convex and convergence to a global minimum can in general not be guaranteed and SGD, for instance may converge to a local minimum. While there is no convergence guarantee, SGD has nonetheless shown that it is the highly successful at minimizing complex, non-convex cost functions Zou et al. (2020). It has

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

been proven that the adaptive extension of SGD, ADAM, converges when applied to smooth, non-convex cost functions. The gradient of the cost function, averaged over the trajectory, has an upper-bound which can be made arbitrarily small, with a rate of convergence of $\mathcal{O}(\frac{ln(K)}{\sqrt{K}})$.

The specification of the optimization technique, the method for efficient differentiation, and the cost function complete all of the components necessary to implement a machine learning algorithm that can be used for the specific task in this thesis. The following section specifies how the model architecture will look like.

3.6 Network Architecture

The previous sections have formalised the general mathematical and statistical concepts necessary for the definition and implementation of a deep learning model. However, they do not give an indication of the specific architecture that model should have beyond the basic MLP; but they rather represent a large class of different types of algorithms and architectures that can be used to address any number of questions Goodfellow et al. (2016). What model is most useful for a given problem is at the discretion of the modeller.

This thesis is focused on the field of computer vision, where the goal is to create machines that can derive meaningful information from visual inputs such as images or videos. In general, images are stored as matrices, with each element of the matrix representing a fraction of the image, known as a pixel. The numeric matrix element indicates the intensity of a color: either red, green and blue for color images or black and white for non-color images. The network processes the matrices by applying different calculations and transformations to the pixels in order to derive characteristics about the image that are useful for solving the task at hand. Given the large size of the matrices and the, possibly very, small size of the relevant details within them, especially with x-rays, the networks need to be *large*, meaning that they require a large number neurons to collectively behave in an intelligent way Goodfellow et al. (2016).

Convolutional Neural Networks

A fully connected network like in figure 3.3 can, in principle, be successfully employed for image classification, see 3.4.2. These networks can have a very large number of parameters and they do not take spatial structure of images into account. This means the network treats input pixels which are far apart and close together exactly the same way, when in image classification it is often the case that neighboring pixels share information Berner et al. (2021). For fully connected networks, these concepts of spatial awareness must be inferred from the training data. However, an architecture that takes spatial proximity into account would be a more natural and efficient approach Berner et al. (2021). The Convolutional Neural Network (CNN) LeCun et al. (1989) is able to realise this notion. CNNs are characterised by three basic ideas: local receptive fields, shared

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

weights/parameters, and pooling. These ideas will be briefly elaborated before providing a mathematical formulation

Local Receptive Fields The CNN is particularly good at collecting information jointly from spatially close inputs Berner et al. (2021). The network does this by making connections only in a small localized region of the input image, called the local receptive field. This is a small weight matrix, for instance 5×5, that is incrementally slid across the input image. At each step, the inputs (pixels) within the sliding window are aggregated and passed on to one neuron in the second layer, thus each neuron of the second layer learns about a small area in the input image and how the pixels within it are related Nielsen (2015). The step size, or stride, is a tunable parameter; however, in practice it is typical to use a stride of 1 for the input image Goodfellow et al. (2016).

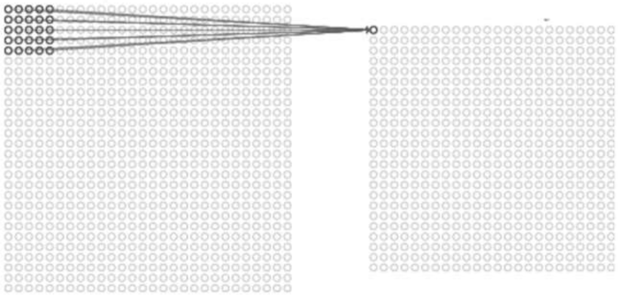
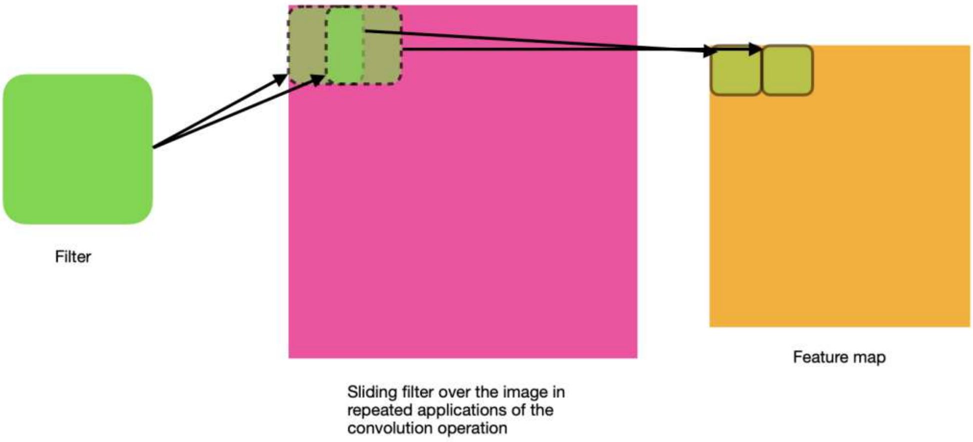


Figure 3.5: visualisation of a local receptive field Nielsen (2015). The small window of input pixels is weighted and aggregated and passed to a single neuron in the second layer.

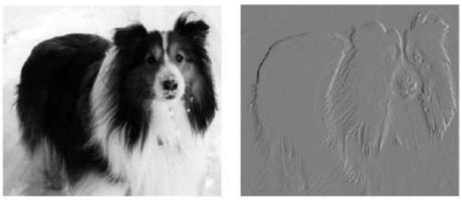
The weight matrix used to process a local receptive field is called a kernel or filter.

Parameter Sharing The weights used within a kernel are designed to identify a specific feature, such as an edge. Thus, the layer of neurons receiving the information from the local receptive fields is called a feature map. The weight matrix applied to a local receptive field is the same for the whole input matrix, meaning all neurons in the subsequent layer share their weights. In addition, this implies that the weight matrix characterises the feature map. Since the weights are designed to detect characteristics in the image, using the same weights implies that all the neurons in the second layer learn to detect the same feature, but in different locations of the image. Once this feature is detected, the location of the feature becomes irrelevant, implying that the network is invariant to translation in the image. Depending on what features are being looked for, many different kernels can be used to create many different feature maps that make up a convolutional layer. Sharing the weight parameters in this way greatly reduces the complexity of the network, speeds up the learning process and reduces the chance of obverfitting Nielsen (2015). The weights are processed using the convolution

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.


Filter
Sliding filter over the image in repeated applications of the convolution operation
Feature map

(a)



(b)

Figure 3.6: (a) Graphical representation of creating a feature map. (b) Application of a filter Goodfellow et al. (2016)

operation. Convolution expresses the amount of overlap of one function g as it is shifted over another function f; in other words "blends" one function with another. When used in image processing, this means interesting features are highlighted and others removed. In practice, the function f is the input image expressed as a matrix and g is the kernel or filter, where each filter is specific to a certain characteristic, like an edge or a shadow. The process of sliding the kernel over the image is visualised in figure 3.6a, and the result of a filtered image is shown in figure 3.6b.

Pooling Pooling is an operation that is used to simplify the information contained in the feature maps of a convolutional layer. Summary statistics such as the *max* or *average* are used to downsample the feature maps and reduce dimensions. This is again beneficial for the efficiency of training by using summary statistics.

Mathematical Formulation A CNN is constructed by multiple convolutional blocks, which are made up of a series of convolutions and poolings across channels. More specifically, consider a group P, the cyclic group of order $d \in \mathbb{N}$ defined by the equivalence

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

classes $[d] = \mathbb{Z}\ d\mathbb{Z}$ for one-dimensional convolution or $[d] = (\mathbb{Z}\ d\mathbb{Z})^2$ for two-dimensional convolution, the convolution of two vectors is defined as Berner et al. (2021)

Definition 3.14. Let $x, y \in \mathbb{R}^P$ be two vectors, where $\mathbb{R}^P$ is the space of mappings from P to $\mathbb{R}$. Then (group) convolution is defined as

$$(x * y)_i = \sum_{j \in P} x_j y_{j^{-1}i}, \quad i \in P$$

Based on this, a convolutional block is defined as Berner et al. (2021)

Definition 3.15. Let $\tilde{P}$ be a subgroup of P and let $p\colon P \to \tilde{P}$ be an operator known as the pooling operator with $t \in \mathbb{N}$ representing the number of channels. For a series of kernels $\kappa_i \in \mathbb{R}^P, \quad i \in [t]$, a convolutional block can be formulated as

$$\begin{aligned} \Psi\colon \mathbb{R}^P &\to (\mathbb{R}^{\tilde{P}})^t \\ v &\mapsto (p(v * \kappa_i))_{i=1}^P \end{aligned}$$

Simply put, the output of a convolutional block Ψ is the composition of the convolution with kernels κ_i and a pooling operation p, along the channel dimension t. Intuitively this means that t convolutions are carried out on the same image with different kernels and then downsampled by a pooling operation such as *max* or *average*. This concept is also called grouped convolutions which is expressed in the group theoretical approach to the definition Berner et al. (2021). A CNN is built by stacking multiple convolutional blocks, possibly with non-linear activation functions between them. At a certain point, the output is mapped to a single vector and is fed into a fully connected structure which can either be one layer or another network of arbitrary size. The concept is shown in figure 3.7.

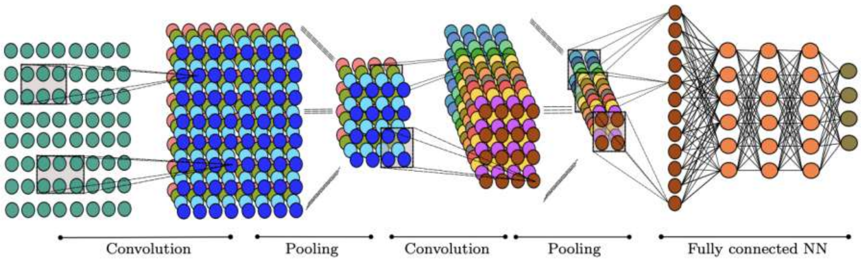


Figure 3.7: Illustration of a convolutional neural network Berner et al. (2021)

Following these considerations, it is important to address whether CNNs can also be considered as universal approximators. Yarotsky (2022), Oono and Suzuki (2019) and Zhou (2020) were able to prove the universality of Convolutional Neural Networks, meaning that it can be used to approximate any continuous function to an arbitrary accuracy when the depth of the neural network is large enough, thereby mimicking the universality of fully connected networks.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Residual Network Implementation

Using sophisticated optimization and differentiation techniques, auch as SGD and backpropagation, it has become a lot easier to train very deep networks, with many publications noting the superiority of deeper networks Bianchini and Scarselli (2014). Thus, the last decade has seen a strong increase in network depth Krizhevsky et al. (2017) Szegedy et al. (2015) He et al. (2016) and a concurrent strong increase in performance in different areas, such as image recognition Goodfellow et al. (2016).

Though it became feasible and efficient to train deep networks, it was observed that, at a certain point, with increased depth, model performance started to degrade Zagoruyko and Komodakis (2016). One of the reasons this occurred was the vanishing gradient, where the gradients of the cost function shrink to zero after several applications of the chain rule, meaning the weights never update and no learning occurs Goodfellow et al. (2016). Another reason is a problem called degradation, which refers to the phenomenon where accuracy gets saturated and then rapidly sinks Niu et al. (2019)

In 2015, the ResNet He et al. (2016) introduced a new framework based on *residual learning* to address these problems. This type of learning refers to the process of learning residual mappings instead of directly learning the overall desired function, implying that the overall desired function is implicitly learned. Directly learning the overall mapping by passing an image through many layers is very computationally intensive and quickly becomes inefficient He et al. (2016). This is because, when an image is passing through the deeper layers of the network, the marginal changes done to the matrix from one layer to the next, or equivalently the added information learned, are very small. In other words, if one views the pass through any one layer as learning a small sub-mapping that assigns the layer input to its output, the network continuously tries to learn a function whose output only deviates slightly from its input. The overall mapping is learned by the composition of all of the sub-mappings, in line with 3.4, but this is highly inefficient when considering the full scale of a very deep network. Instead, the ResNet architecture is constructed by stacking residual blocks, which are collections of (convolutional) layers surrounded by a "short cut" known as skip connection. The intuition to use these connections is that for each block the network learns the small changes between the block's input x and output y, represented by a residual function $\mathcal{F}(x)$ defined as $\mathcal{F}(x) \coloneqq \mathcal{H}(x) - x$, where $\mathcal{H}(x)$ is the the actual underlying mapping the block is trying to learn, thus the input is simply added to the residuals to give the acutaly underlying mapping. Formally, these can be expressed as:

$$y = x + \mathcal{F}(x, w_x)$$

where y represents the output of the block, x the input into the block, $\mathcal{F}$ the function representing the residuals of the block and w_x are the weights within the block. This technique counteracts the accuracy degradation and the vanishing gradient because: if the changes within a block approach zero, the identity mapping is learned which results in at least no higher training error than a shallower counterpart. This allows for

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

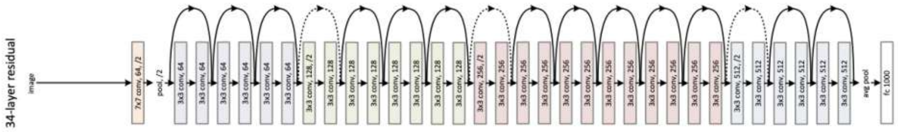


Figure 3.8: The architecture of the ResNet with the residual blocks bordered by the short cuts (loops). In practice there are different types of residual blocks depending on the dimensions of the convolutional layers within. Niu et al. (2019)

deeper structures that can still be trained more efficiently. A full ResNet architecture implementation, including skip connections, is visualized in figure 3.8

Using skip connections, residual networks were able to achieve great depths, with less parameters and better accuracies than their predecessors He et al. (2016). An important architectural detail is that the ResNet is focused more on its depth and less so on its width Zagoruyko and Komodakis (2016). This could be because the ResNet reduces parameters while increasing layers and wider networks would result in more parameters. However, wider networks are found to have very useful mathematical properties: Berner et al. (2021) Lu et al. (2017) report on the benefits of wider networks in terms of optimization, with Lee et al. (2019) even finding that wide networks evolve as linear models under gradient descent. Zagoruyko and Komodakis (2016) notes how residual blocks with a skip connection can be seen as a weakness in terms of optimisation. For instance, it is possible that only a few blocks learn useful representations, or many blocks learn very little information, also known as diminishing feature reuse Srivastava et al. (2015). Considering these drawbacks, Zagoruyko and Komodakis (2016) created the Wide Residual Network (WRN), a wider, shallower variant of the ResNet. He showed that widening ResNet blocks can provide a more effective way of improving performance of residual networks compared to increasing their depth. For instance, they report a wide 16-layer deep network as having the same accuracy as a 1000-layer thin deep network and a comparable number of parameters, while training several times faster. This thesis aims to leverage these mathematical and computational benefits of (wide) residual networks and will implement a WRN as a backbone for the subsequent experiments.

Sophisticated architectures and optimisation algorithms have enabled the efficient training of deep neural networks, helping them achieve state of the art successes and giving rise to the dominance of deep learning in areas such as image recognition. However, research has also shown that there are significant disadvantages attached to deep learning that decidedly affect its adoption in the real world. The following section will highlight the specific aspects this thesis focuses on and how a possible solution could look like.

TU Bibliothek, WIEN Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 4

Introduction To Energy Based Modelling

The previous chapters have described the motivation and the fundamental theory of a state of the art neural network classifier, as well as a description of the most widely used architectures and how they can be optimized. Neural networks are so-called universal approximators, meaning that under minor conditions, every continuous function on a compact set can be arbitrarily well approximated Petersen (2020). For efficient training, they are optimized using gradient based methods such as stochastic gradient descent, or one of its variants. This chapter will introduce the Energy Based Model (EBM) and elaborate on its role in the context of the task presented in this thesis. The concept of energy-based learning will be thoroughly introduced and motivated by expanding on its historical origin. The model is defined and explored, starting from the definition of the core component of this framework: the energy function. It will subsequently be applied to the context of structured probabilistic modelling, with a special focus on sampling and its challenges. The probabilistic form of the energy based model will be derived and the effects on the learning process will be elaborated, especially considering loss functions and their optimization.

4.1 Concept And Origin of Framework

Energy based modelling is in itself not a new field of research. In fact, theoretical foundations go back to the late 19th century to the time of Ludwig Boltzmann. The characterising element of energy based modelling is the scalar-valued energy function E, it assigns individual states/observations an energy value. The terms energy and energy function are very abstract and do not immediately give a clear picture of what an EBM is or how it can be used. Their exact specification always depends on the underlying task, which implies that the energy based framework is very broad and versatile, encompassing many models that can be applied to many different types of problems LeCun et al. (2006). Boltzmann saw energy as the actual, *physical* energy of a system, with the energy function describing the way states in that system behave Boltzmann (1868). However, it can also be viewed in a less physical and more statistical sense: the energy function's purpose is to encode latent dependencies within the data LeCun et al. (2006). The distinguishing feature of the energy function is that it has a-priori no restrictions and its form is always specific to a given task, allowing for a lot of flexibility when designing the model LeCun et al. (2006). These considerations imply that an EBM can be categorised as a type of generative model, where the energy function is trained to capture the underlying (joint) distribution of data. Thus, energy can be understood as a type of likelihood, or in other words an expression of how well a state or observation fits into the data overall LeCun et al. (2006). However, given the lack of a-priori restrictions, energy does not necessarily need to be normalised, meaning it is interpreted as an unnormalised probability LeCun et al. (2006). In this case the EBM is known as a non-probabilistic model, since its output is not a real probability. On the other hand, if a specific task requires the calculation of actual probabilities, a normalisation constraint can be defined and the the energy output simply normalised. Conversely, this case is known as a probabilistic model. This demonstrates how the energy-based framework can be elegantly used to express both probabilistic and non-probabilistic output, depending on a specific purpose. Equation 4.5 is an example of a probabilistic model, meaning the energy function is normalised to produce probabilities.

A purely discriminative model, as in definition 3.4, will try to fit a hyperplane that can be used to separate and classify data Cortes and Vapnik (1995). But the generative model learns how to do more: it gives a complete understanding of how the data was generated and imitates this by fitting a *model distribution* Foster (2019). The trained model knows how the real data is placed within the space and can be used to *generate* very similar synthetic/fake data, by pulling random samples from the model distribution using sampling algorithms such as Monte Carlo Markov Chain methods (MCMC)Foster (2019). In addition to generating data, a generative model's knowledge of the data can be used to express likelihoods Foster (2019). This means that the distribution is modelled by the energy function and energy can be understood as an indicator of how well an observation fits into the data overall LeCun et al. (2006). A-priori the energies are not normalised, meaning the general form of an EBM is a non-probabilistic model; in the probabilistic case, meaning if a specific task requires the calculation of actual probabilities, the EBM's flexibility allows for the energy function to be normalised. The model distribution in this case is given by the Boltzmann distribution, see equation 4.5 below. These probabilities are "real" in the sense that they are rooted in the real underlying data, thus delivering valuable context LeCun et al. (2006). The EBM's deeper understanding of the data and its ability to express real probabilities can be exploited for downstream discriminative tasks, leading to better quality predictions than a purely discriminative model could deliver Grathwohl et al. (2019).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Overall, energy based modelling is a framework that encompasses a broad class of models LeCun et al. (2006). One prominent example of a probabilistic EBM is the Boltzmann machine, developed by Geoffrey Hinton in the 1980s Hinton et al. (1984). Its energy function to capture interesting underlying features, and models them into a distribution using the Boltzmann distribution. The input is binary and the energy function takes the shape of a neural network with two layers. The network structure for a Boltzmann machine is special because all neurons are connected to each other, both within a layer and between layers. In a fully connected network, neurons are usually connected to each neuron in the subsequent layer, i.e. only connected between layers, and not within a layer.

While this was an elegant concept, the interconnectedness of the Boltzmann machine meant that training was very difficult and highly inefficient Goodfellow et al. (2016). Thus, Boltzmann machines were extended into different variants to reduce this inefficiency, chief among them the Restricted Boltzmann Machine (RBM), another important example of an EBM Goodfellow et al. (2016). The key difference between Restricted Boltzmann Machines and Boltzmann machines is the restriction on the energy function, i.e. the underlying neural network. In a RBM no intra-layer connections are allowed, meaning neurons are no longer connected within a layer but rather only in a fully connected fashion between layers. This resulted in significant efficiency gains and increase the model's versatility Goodfellow et al. (2016). For instance, RBMs lend themselves to be easily stacked, thus allowing for the construction of deep generative networks to learn distributions over the inputs very well (the restriction of no intra-layer connections is key here).

RBMs were initially invented by Smolensky in the mid 1980s Smolensky (1986) as a model for information theory, originally known as the Harmonium. However, with the publication of Geoffrey Hinton's fast learning algorithms in 2006 Hinton and Salakhutdinov (2006), they became commonly known as RBMs. Being an extension of the Boltzmann machine, it is also an unsupervised learning model trained on binary inputs that uses the Boltzmann distribution to assign probabilities to various states of, what Smolensky called, harmony. He referred to harmony as a synonym for energy to underline the strong connection between cognition/information theory and physics Smolensky (1986). Boltzmann machines and restricted Boltzmann machines are compared in figure 4.1.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

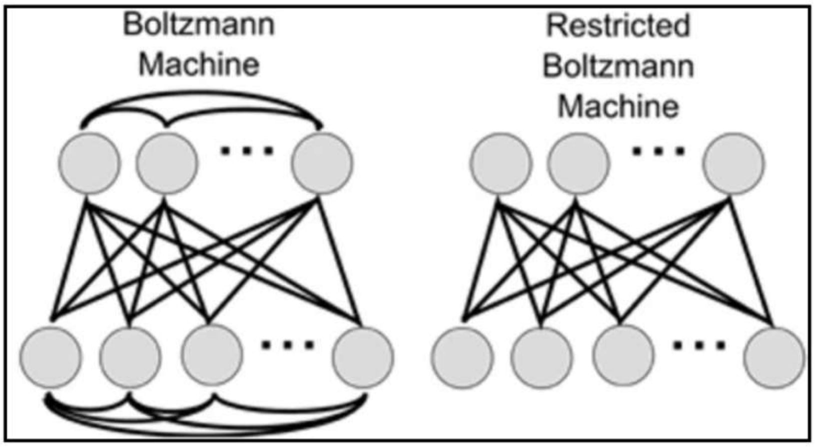


Figure 4.1: Comparison of the energy functions, i.e. the underlying neural networks, for the Boltzmann and restricted Boltzmann machines. The original Boltzmann machine (left) has every neuron within a layer and between layers connected to every other neuron. The restricted Boltzmann machine has no intra-layer connections. Every neuron in a layer is only connected to every other neuron in the subsequent layer. O'Connor et al. (2013)

RBMs have a series of attractive characteristics that make them a particularly important inspiration for the work done in this thesis. One of these is the possibility of stacking multiple layers, thus building energy functions that are deep neural networks. In addition, these networks can be further combined with a classifier that exploits the distributional properties of the generative model to make highly accurate classifications that include important context about the underlying data Smolensky (1986) Hinton and Salakhutdinov (2006). We want to exploit these aspects and use the RBM as a foundation for our EBM to build a model on chest radiographs that can compete with a state of the art discriminative classifier, and use the "built-in" generative capabilities for higher quality predictions, more on this below.

4.2 Energy Function

The energy based framework is a-priori very broad and abstract, meaning the model has no restrictions or conditions concerning the form of the energy function or the inputs LeCun et al. (2006). This offers a lot of flexibility when designing a model; however, in order to make use of it for real-world problems, further specification is needed LeCun et al. (2006). The EBM is characterized by the shape of its energy function LeCun et al. (2006). Its purpose is to capture the underlying dependencies in a dataset and based on these assign a scalar energy to different states LeCun et al. (2006). The interpretation of the information contained in energy is always specific to an underlying task. One interpretation of energy is as a measure of how well an observation fits into the dataset

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

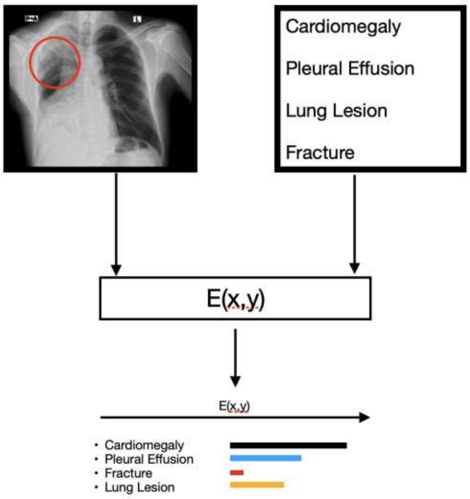


Figure 4.2: Visualisation of how the energy function can be applied to a x-ray image for classification. The function's output can be understood as a measure of compatibility between the x-ray image and each of the possible diseases. A scalar energy for each of the combinations is calculated, with the lowest resulting energy representing the best combination. The disease that determined that the lowest energy, becomes the predicted class.

overall; another is how well a given input/output combination fit together, relative to the dataset overall.

Definition 4.1. Let $\mathcal{A}$ once again be the set of inputs, see section 3.1. Each input $a \in \mathcal{A}$ has a corresponding label b describing the class the input belongs to, where $b \in \mathcal{B}$. The energy function E is defined as:

$$E : \mathcal{A} \times \mathcal{B} \to \mathbb{R} \tag{4.1}$$

The energy function in definition 4.1 is constructed W.R.T. the underlying classification task. How images are classified is outlined by the following intuition: given a trained energy function, an input image is combined in a pairwise manner with each of the possible labels that the EBM was trained on. Each pair represents a different state of the system and is assigned an energy. A low energy implies higher stability, and thus better compatibility LeCun et al. (2006). The combination that results in the lowest energy corresponds to the predicted class. For the specific case of chest radiograph classification, this procedure is visualised in figure 4.2.

The definition of the energy function alone does not give a clear idea of how this model can be trained and used for predictions and other tasks. This is elaborated in the following section.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

4.3 Training and Inference

Training Procedure Training the energy function means shaping the function's landscape, i.e. determining the location of peaks in valleys. To train an EBM is to search for a function E within a family of energy functions $\mathcal{E}$ that gives the best b for any a LeCun et al. (2006), with b and a defined as in definition 4.1. The family of energy functions can be defined as:

Definition 4.2. $\mathcal{E} \coloneqq E_w(I, y) \colon w \in W$

where w describes an index parameter.

There is no restriction on the structure of the elements in $\mathcal{E}$, b, a or w; for example, $\mathcal{E}$ could contain functions that are a linear combination of basis functions LeCun et al. (2006). Section 4.1 introduced the RBM model and its ability to have an energy function approximated by a many-layered network. Building on this idea, the energy function in this thesis will have a CNN architecture, the same as the state of the art neural network classifier, meaning the backbone of the two models is the same, but they are trained in different ways. The key differences are: the form of the loss function and the optimization technique.

Remark. The parameter in definition 4.2 formally represents an index parameter. In this thesis, the energy function is defined as a CNN, meaning the index parameter w will be interpreted as the weights of the network. In short: the energy function is parametrized by the weights of the network.

Finding the best energy function, no matter the underlying architecture, includes the use of a tool that assesses the *quality of the fit* LeCun et al. (2006). In other words, the task of training an EBM includes choosing an appropriate *cost function*, similar to training the state of the art classifier. Considering the fact that the training data $\mathcal{C} \coloneqq \{(a_1, b_1), \ldots, (a_n, b_n)\} \subset \mathcal{A} \times \mathcal{B}$ defined as in 6.3 is labelled, i.e. contains information about which diseases are in each radiograph, the model can be trained in a supervised fashion. The relevant cost that takes label information into account is defined by the following functional LeCun et al. (2006):

Definition 4.3.

$$\mathcal{L}(E, \mathcal{C}) \coloneqq \frac{1}{P} \sum_{i=1}^{P} L(Y_i, E_w(X_i, \mathcal{Y})) + R(w) \tag{4.2}$$

where $\mathcal{L}$ is the loss functional, i.e. a function of a function, L is the cost of prediction and $R(w)$ is the regularizer that represents any prior knowledge about energy functions, e.g. prior restrictions like non-negativity.

Overall, the loss functional is *the average over the training set of a per sample loss.* The per sample loss is described by a function that quantifies the deviation between the model

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

prediction for an input a_i and its corresponding true label b_i:

$$L(b_i, E_w(a_i, \mathcal{B}))$$

The model prediction $E_w(a_i, \mathcal{B})$ is determined by the input/output combination that gives the smallest energy, see figure 4.2. The image a_i is combined in a pairwise fashion to all classes and their respective energies are calculated. The class responsible for the lowest energy becomes the model prediction. This means that for each sample of the training data (a_i, b_i), a slice of the energy surface is evaluated LeCun et al. (2006). However, calculating the loss using each sample can be infeasible for large datasets. Thus, the loss can also be aggregated to a per batch level. Training a model using mini batches not only increases efficiency, but larger batches result in more stable gradient estimates than smaller batches Goodfellow et al. (2016).

Considering that the energy function has a CNN architecture parametrized by weights w, choosing an appropriate loss *functional* can be simplified to defining a loss *function* that is optimized W.R.T. the weights of the network. Thus, the learning problem can be formally described as LeCun et al. (2006):

Definition 4.4.

$$w^* = \underset{w \in W}{min} \mathcal{L}(w, \mathcal{C}) \tag{4.3}$$

In other words, training the EBM is an optimization problem to find the set of network weights w that minimize the loss function in definition 4.3. The specific form of the loss function and how it will be optimized, is detailed further below.

Inference Intuitively, inference for a classification EBM involves finding the class label that results in the lowest energy for a given input image. Mathematically, this can be stated as:

$$\hat{b} = \underset{b \in \mathcal{B}}{argmin} \quad E(a_k, b) \tag{4.4}$$

Thus the inference problem is also interpreted as a *minimization problem* LeCun et al. (2006). In general, the appropriate technique for solving equation 4.4 depends on the form of $\mathcal{C}$ and, by extension, that of the energy function LeCun et al. (2006). In the discrete case where $\mathcal{C}$ has a low cardinality, as is the case in this thesis, finding the optimum is straightforward since it consists of finding the label that results in the lowest energy output, see figure 4.2. In general though, exact, exhaustive searches within $\mathcal{B}$ are infeasible for very large cardinalities. Thus, the inference procedure also depends on the choice of an optimization technique LeCun et al. (2006). Typically gradient based methods such as stochastic gradient descent are used.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

4.4 Probabilistic Model

The preceding sections have laid out the general principles of the energy based framework. It has been shown that the framework is very flexible and its exact form is always use-case specific. Flexibility implies the lack of prior constraints on the energy function, including normalisation constraints. This means that the energy function is a non-probabilistic model. If a task requires the calculation of real probabilities, the energy function can re-interpreted as a probabilistic model using the Boltzmann distribution equation, see equation 4.5 LeCun et al. (2006).

$$p_i \propto \frac{1}{Z} exp(\frac{-E_i}{kT}), \quad Z \text{ normalization denominator } \quad k, \quad T \text{ constant} \tag{4.5}$$

Remark. p_i is the probability of the system being in state i, E_i is the scalar energy of that state, and a constant kT, the product of Boltzmann's constant k and temperature T.

The distribution expresses the probability that a system (a collection of atoms, for instance) will be in a certain state, as a function of that state's energy and the temperature of the system. The intuition is that the probability of a state is specified by the scalar energy of that state E_i: the lower the energy, i.e. the more stable the state, the higher the probability of that state. This serves as an inspiration for the work done by Will Grathwohl in his energy-based paper **Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One** Grathwohl et al. (2019). This thesis applies his findings to the medical image field. The fact that probabilistic models are a special case of the general EBM, is another facet of the flexibility of this framework. This section will elaborate on this specific aspect of energy based modelling. Expression 4.5 can be derived in the following way Boltzmann (1868)

Proof. There are N states where a state is understood to be an image/label combination or just an image. Each of these states has a scalar energy ϵ_j attributed to it. Some states could have the same energy, i.e. I_1 and I_2 could both have energy ϵ_1 while I_3 takes up the rest. Or I_1, I_2 and I_3 could all have different energies. In general, the number of states that have the $j-th$ energy level ϵ_j is given by n_j; therefore, the total energy $E_{tot} = \sum_j n_j \epsilon_j$ and the total number of states is $N = \sum_j n_j$. There are clearly many different ways to arrange the energy among the different states, such that the total energy remains constant. Using combinatorics, the total number of ways to distribute the energy among these different states is:

$$\Upsilon = \frac{N!}{n_1! \dots n_j!}$$

Taking the logarithm of Υ and using Stirling's approximation of $ln(n!) = Nln(N) - N$ Dutka (1991):

$$ln(\Upsilon) = ln(N!) - (\sum_j ln(n_j!)) = Nln(N) - N - \sum_j ln(n_j!)$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Under the constraints of total energy and total number of states, we seek to maximise this expression W.R.T. the n_j, because out of all possibilities to group the states according to their energy n_j, we want the one that will give the highest probability. We can construct the following Lagrangian:

$$\begin{aligned} L &= ln(\Upsilon) - \alpha(\sum_j n_j - N) - \beta(\sum_j n_j\epsilon_j - E_{tot}) \\ &= ln(N!) - (\sum_j ln(n_j!)) = Nln(N) - N - \sum_j ln(n_j!) - \alpha(\sum_j n_j\epsilon_j - E_{tot}) - \beta(\sum_j n_j - N) \end{aligned}$$

Taking the derivative and setting equal to zero gives for n_j:

$$\begin{aligned} \frac{\partial L}{\partial n_j} &= -ln(n_j!) - \alpha n_j - \beta\epsilon_j n_j = 0 \\ &\Rightarrow n_j = exp(-\alpha)exp(-\beta\epsilon_j) \\ &n_j \propto exp(-\beta\epsilon_j)\text{since } exp(-\alpha) \text{ is a constant} \end{aligned}$$

This is normalised using the constant $Z = \frac{\sum_j exp(-\beta\epsilon_j)}{N}$ which yields:

$$n_j \propto \frac{1}{Z}exp(-\epsilon_j\beta)$$

This is the exact expression in 4.5 with $n_j = p_j$; $\epsilon_j = E_j$ and $\beta = \frac{1}{kT}$ □

In very general terms, any statistical model is meant to approximate a dynamic and complex system in order to help describe or infer different types of situations Hastie et al. (2009). The model must learn to encode interesting latent dependencies in the data and any a-priori knowledge/assumptions about model parameters. Considering that the model is an abstraction, all inferences derived from the model are inherently uncertain Bishop and Nasrabadi (2006). A probabilistic model takes the inherent randomness into account. This approach makes intuitive sense for real world problems, especially for the one addressed in this thesis, simply because the model cannot account for every single detail and all of the circumstances surrounding the diagnosis of every single x-ray and its subtle details. The probabilistic approach incorporates distributions into the model.

A state of the art discriminative classifier, as defined in section 3.4.1, produces a distribution across the classes via the final softmax layer, and can also considered to be a probabilistic model. The distribution across the classes is an example of the categorical distribution, a special case of the multinomial distribution Bishop and Nasrabadi (2006). These classifiers are known to deliver highly accurate results Baltruschat et al. (2019); however, it is a common misconception, because of this high accuracy, to consider the softmax output a probability representative of the underlying data and a real expression of uncertainty with deeper meaning attached to it. In fact, the softmax output gives no deeper insight into the data and its generation process, see 5.1.1 Hendrycks and Gimpel

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

(2016).

The probabilistic EBM differs from the state of the art classifier because it is a *generative model* and learns to emulate the data generation process delivers as output a distribution estimate for the original raw data, making it possible to express likelihoods and confidence intervals. The EBM's output is *directly* representative of the underlying data and contains *actual* insight about the uncertainty attached to a prediction, making it useful for downstream discriminative tasks. Another aspect of the energy based framework's flexibility and (mathematical) elegance. The energy function, in this case represented by a CNN, is re-expressed using the **Boltzmann distribution** Boltzmann (1868) LeCun et al. (2006), see also 4.5, by simply "inserting" it into the distribution and appropriately normalizing. The most important question in this probabilistic point of view, aside from the definition of the energy function, will be *how* the function is normalized - this is one of the key differences to the state of the art classifier.

The mulit-class classification setting for the EBM is defined by LeCun et al. (2006) in the following way:

Definition 4.5. Let E_w be an arbitrary energy function indexed by the parameter w and a and b represent input and output, respectively. Then

$$P_w(a,b) = \frac{exp(-\beta E_w(a,b))}{\int\limits_{b \in \mathcal{B}} exp(-\beta E_w(a,b))da} \tag{4.6}$$

is the joint probability distribution over input and output. E is represented by a CNN architecture parametrised by weights w, see 3.4. Furthermore, a and b represent an input and its corresponding label, respectively, $-\beta$ represents a tempering parameter, usually set to $\beta = 1$, that controls the distribution's shape. The integral in the denominator normalizes the expression.

The probabilistic interpretation of an EBM is a form of a structured probabilistic model Goodfellow et al. (2016). These models represent distributions by using graphs to describe the dependencies of random variables in a probability distribution, thus they are also called graphical models. Structured probabilistic models reduce the computational costs of training, inference and sampling compared to unstructured models Goodfellow et al. (2016). Only direct interactions between random variables are modelled, instead of every possible interaction, which results in a lot less parameters to be estimated, while retaining reliability in the distribution estimates. This also has effects on statistical efficiency: since there are less parameters, the model will tend to overfit less and require less training data Goodfellow et al. (2016). Given the fact that there is typically no clear direction of influence, or causality, in the dependencies between the random variables encoded in the energy function, EBMs are undirected graphical models. Undirected graphical models are often also called Markov Random Fields or Markov Networks Kindermann and Snell (1980).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The output of the energy function in 4.5 represents an un-normalized distribution over an un-directed graph and must be normalized, so that it integrates to 1. Thus, to successfully transform an energy function to a probabilistic model, i.e. re-formulate it as the Boltzmann distribution, proper normalisation is essential. In theory, this is achieved by the partition function:

Definition 4.6.

$$Z_w = \int\limits_{b \in \mathcal{B}} exp(-\beta E_w(a, b)) da \tag{4.7}$$

The partition function is an integral over all possible combinations between input images a and elements of $\mathcal{B}$. Depending on the structure of $\mathcal{B}$, this is typically *intractable* to calculate Goodfellow et al. (2016) which, in turn, creates difficulties in training and ultimately also hinders using the probabilistic EBM for inferential purposes, such as sampling and calculating likelihoods. So, while expressing the EBM as a distribution is mathematically straightforward, it is intractable to carry out probabilistic inference. For this reason, approximative methods need to be used to solve the intractable integration and deliver an approximate solution. How the model is trained and how the problem of the partition function is approached in this thesis is detailed in the following sections.

4.5 Cost Functions

Training an EBM, much like traning a classical DL model, see section 3.5, involves optimizing a cost function. Considering the overall flexibility of the energy based framework, there are many types of cost functions for the different ways an EBM can be interpreted and trained LeCun et al. (2006). When training the EBM, the goal is to shape the energy function in such way that it assigns "good" inputs low energy and bad ones high energy, the general formulation of this optimization problem is shown in 4.3. Thus, the loss function needs to be optimized in such a way that it pulls the energy function up for wrong combinations, and pulls it down for right ones LeCun et al. (2006).

This thesis uses the Boltzmann distribution model to formulate the EBM as a probabilistic model, meaning it becomes possible to calculate likelihoods. The natural way to train this type of model Goodfellow et al. (2016) is using maximum likelihood to minimize the **negative log likelihood** loss, analogously to section 3.5. This gives the set of parameters that maximises the likelihood under the distribution for given inputs. Furthermore, minimizing NLL equivalently minimizes the distance between the distribution estimated by the model, p_{model} and the true data distribution p_{data}. However, the following derivation by LeCun et al. (2006) shows how the EBM also involves some additional considerations that differentiate its NLL cost function:

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The likelihood to be maximized over training data $S := \{(a_i, b_i)|i = 1 \dots n\}$ is:

$$P_w(a|b) = \prod_{i=1}^{n} P_w(b_i|a_i) \tag{4.8}$$

This is equivalent to minimizing:

$$\sum_{i=1}^{n} -log(P_w(b_i|a_i)) \tag{4.9}$$

Using the Boltzmann distribution 4.5 this becomes:

$$\sum_{i=1}^{n} -log(P_w(b_i|a_i)) = \sum_{i=1}^{n} \beta E_w(a_i, b_i) + log\left(\int_{b\in\mathcal{B}} exp(-\beta E_w(a_i, b))da\right) \tag{4.10}$$

Averaging over the data set and dividing by β yields the expression for negative log likelihood loss

$$\mathcal{L}_{nll}(w, \mathcal{S}) = \frac{1}{n}\sum_{i=1}^{n}\left(E_w(a_i, b_i) + \frac{1}{\beta}log\left(\int_{b\in\mathcal{B}} exp(-\beta E_w(a_i, b))da\right)\right) \tag{4.11}$$

The derivation shows how the loss is decomposed into two key components, the *positive phase* and the *negative phase* Goodfellow et al. (2016):

- $E_w(a_i, b_i)$ the energy between the $i - th$ input and $i - th$ response (positive phase).
- $log\left(\int_{b\in\mathcal{B}} exp(-\beta E_w(a_i, b))da\right)$ the log of the partition function definition 4.6 also referred to as free energy (negative phase). This term contains the energies between the $i - th$ input and *all* possible responses $b \in \mathcal{B}$ (even the wrong ones).

Minimizing the negative log likelihood loss requires the calculation of the gradient of definition 4.11. The gradient of the likelihood for one sample can be expressed as LeCun et al. (2006):

$$\nabla_w \mathcal{L}_{nll}(w, (a_i, b_i)) = \nabla_w E_w(a_i, b_i) - \int_{b\in\mathcal{B}} \nabla_w E_w(a_i, b) P_w(b|a_i) \tag{4.12}$$

It is not straightforward to evaluate this gradient due to the intractability of the partition function, which directly impacts the calculation of the gradient for the whole likelihood function LeCun et al. (2006); thus, the gradient must be approximated. Considering that $P_w(b|a_i)$ is given by the probabilistic definition of the EBM 4.5, the gradient of the negative log likelihood objective function can be re-expressed using an expectation over the model distribution:

$$\nabla_w \mathcal{L}_{nll}(w, (a_i, b_i)) = \nabla_w E_w(a_i, b_i) - \mathbb{E}_{P_w}\left[\nabla_w E_w(a_i, b)\right] \tag{4.13}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

The gradient 4.13 contains an inherently random term, implying that minimising the loss function using maximum likelihood becomes difficult and requires a different set of methods to find a solution Goodfellow et al. (2016). The problem of approximating the partition function and its gradient is a central topic when training a EBM. How this problem was approached in this thesis will be explained in the next section.

4.6 Approximating The Partition Function

The previous section has shown the importance of evaluating the (gradient of the) partition function for optimising the loss function and training the model. The partition function is given by

$$\int_{b \in \mathcal{B}} exp(-\beta E_w(a_i, b)) dI \tag{4.14}$$

From 4.13, one can see that the gradient of the partition function can expressed by the expectation:

$$\mathbb{E}_{P_w} \left[\nabla_w E_w(a_i, b) \right] \tag{4.15}$$

This expression forms the the basis for Monte Carlo Markov Chain methods (MCMC) to approximately maximize the likelihood with intractable partition functions Goodfellow et al. (2016). MCMC methods are a class of techniques that rely on random sampling to calculate numerical results, for instance for intractable integrals. A brief outline is given below.

4.6.1 Monte Carlo Markov Chain Methods

Classical Monte Carlo methods can be used to solve any problem of a probabilistic nature. They are mainly used in three problem classes Kroese et al. (2014): optimization, numerical integration, and generating draws from a probability distribution. In all cases, random numbers are repeatedly sampled and evaluated for statistical analysis. In machine learning, they are especially popular for approximating expected values of random variables whose computation is intractable Ahn (2015). Hence, the rationale of using Monte Carlo methods to evaluate the gradient in equation 4.15. In a nutshell: Many independent, random samples are drawn and evaluated using the random variable. The expected value/integral is approximated by calculating the arithmetic mean of the resulting functional values which, by the law of large numbers, converges to the expected value. Thus, the more random trials that are performed, the more accurate the approximated quantity will become Goodfellow et al. (2016). This implies, that the number of samples provides control over the precision of the quantity that is being approximated, often limited by the computational complexity of drawing a sample Murphy (2012).

However, sampling from target distributions can be very difficult and independence of samples is not always a correct assumption, especially in high dimensions with probabilistic

TU WIEN Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

models Goodfellow et al. (2016). This makes calculating equation 4.15 with classical Monte Carlo methods difficult. In these cases, Markov chains can be used to generate the samples. Markov chains are stochastic processes where the current state only depends on the state that came immediately before. The mathematical formalisation of this is known as the Markov Property and is the defining element of the Markov chain Norris and Norris (1998):

$$P(X_t = i | X_{t-1} = j)$$

The probability is also knows as transition probability. MCMC methods unify Markov chains and the random sampling of Monte Carlo methods: a Markov chain is used to sample from a distribution of interest, and the Monte Carlo method uses these samples to approximate an expectation using the arithmetic mean. The distribution the Markov chain samples from, i.e. the model distribution, should be invariant or stationary, meaning transition probabilities do not change the distribution Ahn (2015). In practice, the chain is modelled by an algorithm called a sampler. The most popular algorithms to generate samples for MCMC are the Gibbs sampler and the Metropolis-Hastings algorithm Norris and Norris (1998).

The Markov chains are arbitrarily initialized and they move around randomly, looking for places with a high contribution to the integral, also known as modes, to move into next, assigning them higher probabilities Norris and Norris (1998). MCMC algorithms are sensitive to their starting point, and often require a warm-up phase or burn-in phase to move in towards an area of high probability, after which prior samples can be discarded and useful samples can be collected. The burning-phase is very costly because it takes time until the chain finds useful samples. In addition, it can be challenging to know whether a chain has converged and collected a sufficient number of steps. Often a very large number of samples and multiple chains need to be run for a large predefined, fixed number of steps to produce a representative sample Murphy (2012).

The main problem with traditional MCMC is that they do not scale well to large scale problems Ahn (2015), meaning they they are too computationally intensive for a lot of data, which is a significant disadvantage for machine learning. The variance of MCMC estimates converges to 0; however, the amount of real-world data necessary to achieve this is too much for traditional algorithms considering their failure to scale Ahn (2015), meaning their estimates will have high errors.

4.6.2 Applied MCMC - Gradient Approximation

In general, problems where the function that needs to be solved is the expected value of another function, as in equation 4.15, can be solved using stochastic approximation techniques, that is without using MCMC, created by Robbins and Monro Robbins and Monro (1951). This is a class of iterative techniques that use noisy observations to find the root of a function. Stochastic approximation can be applied to optimisation, if the function in question is the gradient of a function; it works by finding local optima using

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

noisy subgradient observations ?. In practice, these methods work by processing small batches of data at each iteration, updating model parameters by taking small gradient steps in a cost function Welling and Teh (2011).

$$w_{t+1} = w_t + \epsilon_t \nabla f$$

Stochastic approximation techniques are particularly useful because they are guaranteed to converge in L_2, given constraints on the step size ϵ_t of the iterations Robbins and Monro (1951).

$$\sum_{t=1}^{\infty} \epsilon_t = \infty \quad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty$$

However, these do not capture parameter uncertainty and can potentially overfit data Welling and Teh (2011).

The benefit of using MCMC methods is that in addition to approximating the intractable likelihood gradient, they allow for the expression of uncertainty by generating samples from the model distribution and they do not overfit the data Welling and Teh (2011). This makes them especially interesting for training a probabilistic EBM. One of the most popular gradient approximation algorithms in deep learning is the contrastive divergence algorithm Hinton et al. (2006b). It is a special case of standard MCMC methods because it does not initialize the chains with random points in space but instead with random points from the data distribution, thereby reducing burning-in time and increasing efficiency. Naively, every time a gradient approximation is needed, a mini-batch of data is randomly selected and a chain is run for a pre-defined number of steps. During each step, a random sample is generated, using a Gibbs sampler for instance, that depends on the previous sample Hinton et al. (2006b). While this method does increase efficiency slightly, it still can become computationally infeasible because in every gradient step new chains are initialized Goodfellow et al. (2016). Contrastive divergence converges to the set of parameters that minimize the Kullback-Leibler divergence between model distribution and data distribution, which is equivalent to minimizing crossentropy and the negative log likelihood.

A strategy that further increases effciency is the Persistent Contrastive Divergence algorithm Younes (1999) Tieleman (2008). The intuition here is to create a so-called replay buffer to store states from earlier running Markov Chains and use these to initialize the new Markov chains, meaning the chains are not reset every time. In essence, the gradient estimates happen continuously, or *online* Tieleman (2008). This method reduces the time for the current chain to reach equilibrium, since the distributions in the individual steps are similar. The algorithm works better with smaller learning rates, which improves the gradient estimations. Conversely, the method becomes inaccurate and training diverges.

TU WIEN Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

4.6.3 Stochastic Gradient Langevin Dynamics

The preceding sections have highlighted some of the main benefits, drawbacks and applications of traditional MCMC methods. The main drawback is that these methods do not scale well to large amounts of data, making them uncompetitive compared to optimization techniques such as SGD. An algorithm that can approximate the gradient in equation 4.15 while combining the distributional benefits of traditional MCMC and the efficiency/convergence benefits of optimization, would be particularly useful to train the EBM. One such method is a class of MCMC methods called Stochastic Gradient Langevin Dynamics (SGLD) Welling and Teh (2011). This algorithm was used to train EBMs successfully by Grathwohl et al. (2019), enabling efficient training and sampling of probabilities from the target distribution which happens to be the Boltzmann distribution Langevin (1908). This thesis uses the same methodology to train an EBM.

Langevin Equation and Langevin Dynamics

The Langevin equation is a stochastic differential equation describing the motion of a particle when subjected to a combination of deterministic and random forces Langevin (1908):

$$m\frac{\partial^2 x}{\partial t^2} = -m\gamma\frac{\partial x}{\partial t} + F_t \tag{4.16}$$

x is the particle's position, m its mass, $\frac{\partial^2 x}{\partial t^2}$ its acceleration, $\frac{\partial x}{\partial t}$ its velocity and F_t a random fluctuating force; $m\gamma\frac{\partial x}{\partial t}$ can be understood as the viscous friction force on the particle that is proportional to its velocity (Stokes' Law). The fluctuating force F_t is assumed to be a Gaussian process such a Brownian motion MacKay et al. (2003). We can rearrange the terms, rename the variables and write a discrete approximation of the stochastic differential equation, since continuous time cannot be simulated by computers Ahn (2015):

$$x_{t+1} = x_t - \frac{\alpha}{2}\nabla E(x_t) + \epsilon, \quad x_0 \sim \mathcal{U}(-1,1) \quad \epsilon \sim \mathcal{N}(0,\alpha) \tag{4.17}$$

where x_0 is sampled uniformly. Since via Stokes' law, velocity is proportional to the viscous friction force, $\frac{\partial^2 x}{\partial t^2}$ becomes the gradient of the force E, which can also be considered as the energy; ϵ is normally-distrbuted noise; α is the step size and the standard deviation. In practice the step-size α and the standard deviation of ϵ is often chosen separately because it allows for faster training Grathwohl et al. (2019). Expression 6.1 is the discretization of the stochastic differential equation.

The Markov chain defined by SGLD is non-stationary, such that the $t-th$ step transition will have as its equilibrium distribution the posterior Welling and Teh (2011). Since the chain is not stationary, it is not immediately guaranteed that it converges to the posterior, or target, distribution. However Borkar and Mitter (1999) proves that this holds, and the stationary distribution is equal to the target distribution Welling and Teh (2011). The solution of the Langevin equation is the Boltzmann distribution Langevin (1908),

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

meaning the stationary distribution of the Markov chain is the Boltzmann distribution.

Remark. Stochastic gradient langevin dynamics is an inherently Bayesian method Welling and Teh (2011). Expression 6.1 can be adapted using likelihoods, priors and posteriors to have the following form Welling and Teh (2011):

$$\omega_{t+1} = \omega_t \frac{\epsilon}{2} \left(\nabla log(p(\omega_t) + \sum_{i=1}^{N} \nabla log(p(x_i|\omega_t) \right) + \eta_t \tag{4.18}$$

where ω are the distribution parameters; ϵ the step size; η Gaussian noise

Stochastic Gradient Langevin Dynamics

When we approximate 4.15 with stochastic gradient langevin dynamics, we run Markov chains across the Boltzmann distribution and average the resulting samples in order to approximate the expectation/gradient. This information then flows into the original maximum likelihood problem for the cost function. Overall, training the EBM involves implicitly generating samples from the target distribution and improving by minimizing the resulting loss Du and Mordatch (2019). To leverage the efficiency of stochastic optimization techniques, generating the samples can also be done batch-wise Welling and Teh (2011). The proof that the properties of the Markov Chain and the Langevin dynamics still hold is given by Borkar and Mitter (1999). As with all MCMC methods, the burn-in phase is inefficient. It can be extended analogously to persistent contrastive divergence to keep a replay buffer of useful samples from previous chains and initialize new chains using these.

Stochastic gradient langevin dynamics is an extension of both stochastic optimization and traditional MCMC methods. Not only is it computationally efficient Welling and Teh (2011) and enables efficient training of EBMs Du and Mordatch (2019), but it allows us to directly sample from our target distribution, the Boltzmann distribution.

4.7 Benefits And Drawbacks Of The Energy Based Framework

This chapter has introduced energy based modelling and how it can be used in an image classification setting. The energy based framework is an elegant, unifying framework that gives the modeller a lot of freedom and flexibility when designing a model; however, it is also a very abstract concept that needs exact specification, meaning it is not an "out of the box" algorithm but needs a lot of architectural work.

The central element of an EBM is the energy function. It encodes the latent variables within the dataset and assigns states, for instance images or image/label combinations, a scalar energy that indicates how well this state fits into the data. In other words the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

energy function is an unnormalized density. This thesis considers the special case of the energy function beeing re-formulated as a probabilistic model using the Boltzmann distribution.

Following the work of Grathwohl et al. (2019), MCMC techniques from Bayesian inference to train it to fit a distribution over the data Du and Mordatch (2019). This allows for reliable uncertainty quantification, sample generation and the expression of likelihoods Welling and Teh (2011). Overall, the goal is to create a hybrid model which delivers a fits a distribution over the data that can then be used for downstream discriminative tasks Du and Mordatch (2019)Grathwohl et al. (2019).

While the energy based framework is mathematically elegant and flexible, its realisation has some key drawbacks that make the widespread/commercial use difficult.

- The biggest drawback of EBMs is the fact that training is very difficult and time consuming, especially in very high dimensions. The reason for this is the computations necessary for the stochastic gradient langevin dynamics training algorithm. The number of steps a Markov Chain must take to reach stable solutions is not known. Samples from the estimated distribution are highly correlated, meaning that a representative sample can only be achieved with many parallel chains, which adds to the computational effort.
- The partition function that normalizes the energy function is typically intractable, which necessitates the use of approximative methods to calculate the gradient of the cost function, increasing the complexity of training.
- Training the EBM as a hybrid model creates a trade off between fitting a high-quality distribution and producing class predictions the meet the state of the art in accuracy. Typically, some of the accuracy will be sacrificed in order to produce higher quality out (class predictions or probabilities).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 5

Evaluation And Comparison Of Energy Based Modelling And Deep Learning

This chapter will highlight the similarities and differences of the deep learning and energy based frameworks. Special attention is paid to how the interpretation of their their respective outputs differ.

5.1 Energy Based Modelling Compared to Deep Learning

Chapters 3 and 4 have introduced the deep learning and the energy based framework, respectively, showing the differences in their motivation and training. The two modelling frameworks share a significant similarity: a CNN architecture as their backbone. The CNN is the state of the art model for computer vision tasks, especially classification. They are able to far outperform their well-known discriminative counterparts such as the support vector machine or logistic regression models Goodfellow et al. (2016). This is because their usage of local receptive fields, the convolution operation and efficient downsampling allows them to be particularly good and efficient at finding specific patterns within an image and memorizing them Goodfellow et al. (2016).

Training a CNN using deep learning methodology has delivered unprecedented results to many different types of highly complex problems and domains, including the medical domain. It has re-defined the state of the art in classification accuracy in experiments, even on medical image data Baltruschat et al. (2019); however, they have also been shown to have specific characteristics that potentially slow down, or even hinder, widespread adoption.

5.1.1 Deep Learning

Black Boxes While networks are able to approximate any function, see section 3.4.2, the approximation does not give any details about the underlying structure, thus they are often referred to as black boxes. In other words, when a network generates a class prediction, the interpretation of these results is non-trivial, meaning there is no intuitive indication what specifically drives the prediction, which reduces trust in the network Adebayo et al. (2018) Rudin (2019). This lack of transparency invariably creates issues of accountability and can have severe consequences Rudin (2019). One possible way forward is to produce separate solutions that aim to explain the behaviour of the black box, known as explainable machine learning Rudin (2019). Proper explanations could potentially help users improve the quality of follow up tasks and, perhaps, reveal bias or other unintended effects learned by a model Lakkaraju et al. (2017). In computer vision, there has been a lot of research and development into the creation of such, for instance GradCam Selvaraju et al. (2019). This uses the gradient of the class output flowing into the final convolutional layer to produce a coarse localization map highlighting the regions in the image where the model is looking when predicting the class. However, these methods typically cannot explain why a prediction is the way it is, i.e. it is an inaccurate representation of the original model and not completely faithful to what the original model computes Rudin (2019). A combination of inaccurate explanations and an opaque model lowers trust in the approach overall.

Data and Hardware Requirements Networks need a lot of data for training in order to learn the underlying dependencies well and produce accurate predictions Goodfellow et al. (2016). In addition, supervised learning problems require the data to be labelled data which requires extra effort to create and maintain. Furthermore, training networks requires sophisticated hardware Goodfellow et al. (2016). For computer vision tasks, GPUs have become indispensable to carry out efficient, high-dimensional calculations. State of the art deep learning algorithms can take several weeks to train completely from scratch Goodfellow et al. (2016). Inferior hardware hinders proper training because it cannot carry out the necessary calculations and it does not have the capacity for the amount of data needed to train the model, even if a significant amount of data exists.

Uncertainty A high accuracy is only one part of what a successful deep learning implementation should include; a model should have the ability to capture model uncertainty, or epistemic uncertainty Gal (2016). This type of uncertainty is linked to the knowledge that is available, or lack thereof. More specifically, many different types of models can be used to solve a specific type of problem: the architecture chosen and the way the parameters are estimated, all induce uncertainty into the predictions Gal (2016). Failure to capture this can have severe effects, especially when human lives are involved Gal (2016). Overall, quantifying uncertainty is important information for both modellers and end-users Gal (2016). This includes understanding if a model is over- or underconfident, i.e. uncertainty estimates are too small or too large, which can help get better performance Gal (2016). In addition, it is very important in real world applications

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

for a model to understand when a testing observation is too different from the overall training data Hendrycks and Gimpel (2016). Intuitively, a model should not make a highly confident prediction on an image that is completely different than the data it was trained on, otherwise wrong information could be derived from that prediction Gal (2016). Unfortunately, deep learning classifiers often do deliver highly confident predictions for nonsensical inputs and do not account for this uncertainty information Hendrycks and Gimpel (2016). This scenario is not purely theoretical. Aside from human error, training and testing distributions can naturally shift over time, implying that models would need to be continuously re-trained, which can be very costly. Alternatively, a model could be trained to include an expression of high uncertainty or low confidence to flag these observations for human intervention Gal (2016). The softmax output of a classification network is often erroneously interpreted as such an expression of model confidence Gal (2016), or even a real likelihood relative to the entire dataset, but this is false Guo et al. (2017) Gal (2016). The uncertainty attached to predictions implies a certain level of context and knowledge within these predictions, which can deliver valuable insights; this means: the numerical output for a class *should* be in line with the ground truth occurrence of this class in the dataset and can in fact be interpreted as a likelihood Gal (2016).

The topic of uncertainty in neural networks is a central focus in this thesis. Specifically, the experiments focus on the following two aspects:

1. Out of Distribution Detection (OOD) capabilities of the model Hendrycks and Gimpel (2016)
2. The calibration capabilities of the model Guo et al. (2017)

These will be elaborated separately below.

Calibration

If neural networks are to be entrusted with sensitive decisions such as medical diagnosis, they should be able to indicate when they are likely to be incorrect Guo et al. (2017). This means the score produced by a network for a specific class, should reflect its ground truth likelihood Nixon et al. (2019), this is known as calibration. More intuitively: given 100 chest x-rays, each with a score of 0.8 for having a fracture. If the model in question is well-calibrated, it can be expected that the ground truth shows that 80% of the chest x-rays also have a fracture present and are correctly classified. This is essential in machine learning applications, especially in high stakes tasks like medical imaging where confident, but incorrect, predictions could have disastrous consequences Minderer et al. (2021). Even though neural networks and their output are considered as black boxes and cannot be fully explained, confidence calibration provides a way for avoiding major mistakes by associating each prediction with an uncertainty/confidence score that reflects the ground truth data. Calibrated probability scores associated with each prediction allow

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

low-quality predictions to be identified and discarded Gal (2016).

Convolutional Neural Networks have been able to achieve excellent accuracies on image tasks but, paradoxically, this has resulted in a rising level of miscalibration as well Guo et al. (2017). The softmax scores produced by classification networks are typically interpreted as confidence scores, i.e. the softmax is interpreted as a probability backed by the ground truth data. In other words: a classification model produces a softmax score for a class of 0.996, does this mean the chance that the input belongs to that class is truly 99.6%? Networks often make poor/incorrect predictions with softmax scores of nearly 100%, meaning they are over-confident and there is a risk of actual misinterpretation of these values and how likely the correctness truly is Guo et al. (2017). The reason for this cannot be distilled into one causal driver, but rather a combination of different factors, including: the large increase of model depth, models commonly have 100s of layers with 100s of convolutional filters per layer, and the regularization techniques that enable efficient training of these very deep networks, such as weight decay (adding a penalty term to the cost function) and batch normalization (normalizing across batches) Guo et al. (2017). In addition, the use of softmaxes also contribute to the high confidence predictions Hendrycks and Gimpel (2016). Softmax scores are computed with the exponential function; minor additions to the inputs can already lead to substantial changes in the output distribution. Rising miscalibration in the face of near-human accuracy seems counterintuitive and poses a severe risk to open adoption of deep learning models, as it greatly reduces trust in them Chen et al. (2020a).

Formally, perfect calibration is defined as Guo et al. (2017):

Definition 5.1 (Calibration)**.** Let $\mathcal{A}$ and $\mathcal{B}$ be the sets of inputs and their corresponding labels as laid out in section 3.1, respectively. $a \in \mathcal{A}$ and $b \in \mathcal{B}$ can be considered as random variables that follow a joint distribution π representing the ground truth in the following way:

$$\pi\left(a, b\right) = \pi\left(b|a\right)\pi\left(a\right)$$

The joint distribution is based on the probability measure $\mathbb{P}$.

Let G_w be the network defined in definition 3.4 and consider the output

$$G_w(a) = \left(\hat{b}, \hat{q}\right)$$

where $\hat{b}$ is a class prediction and $\hat{q}$ is its associated confidence, or in other words the score, like the softmax, indicating to what degree this prediction is correct, e.g. $(3(= \text{fracture}), 0.8)$. Ideally, $\hat{q}$ is calibrated, meaning it is equal to the ground truth likelihood of the predicted label. A model is perfectly calibrated if:

$$\mathbb{P}(\hat{B} = b|\hat{Q} = q) = q \quad \forall p \in [0, 1] \tag{5.1}$$

This means that the probability of predicting a label, given that the corresponding confidence score is the likelihood of that label in the ground truth, is exactly the observed

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

likelihood of that label in the ground truth Guo et al. (2017). Achieving perfect calibration is virtually impossible, due to time constraints and the fact that the probability in 5.1 cannot be computed using finitely many samples since $\hat{Q}$ is a continuous random variable Guo et al. (2017). This motivates empirical measures to evaluate calibration.

Reliability Diagrams A simple and easily accessible visual indicator of a model's calibration are reliability diagrams Degroot and Fienberg (1983). These diagrams plot expected sample accuracy as a function of confidence. Predictions are grouped into interval partitions, called bins, based on the prediction confidence value (softmax), and the accuracy for each bin is calculated. Any deviation from the diagonal represents miscalibration. Expected prediction accuracy can be estimated by Guo et al. (2017):

For a given set of predictions from a classification model, group the predictions into M bins each of size $\frac{1}{M}$. Let B_m be the set of indices of the predictions whose score falls into the $m-th$ bin defined as $r_m = \left(\frac{m-1}{M}, \frac{m}{M}\right]$.

Definition 5.2. The accuracy of B_m is

$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(y_i = \hat{y}_i) \tag{5.2}$$

where y_i and $\hat{y}_i$ are the predicted and true class labels for prediction i and $\mathbf{1}$ is the indicator function. $acc(B_m)$ is an unbiased and consistent estimator of $\mathbb{P}(\hat{y} = y|\hat{P} \in r_m)$, meaning $\mathbb{E}[acc(B_m)] = \mathbb{P}(\hat{y} = y|\hat{P} \in r_m)$ and $acc(B_m)$ converges in probability to the true value of the accuracy Guo et al. (2017).

The average confidence within a bucket B_m is defined as:

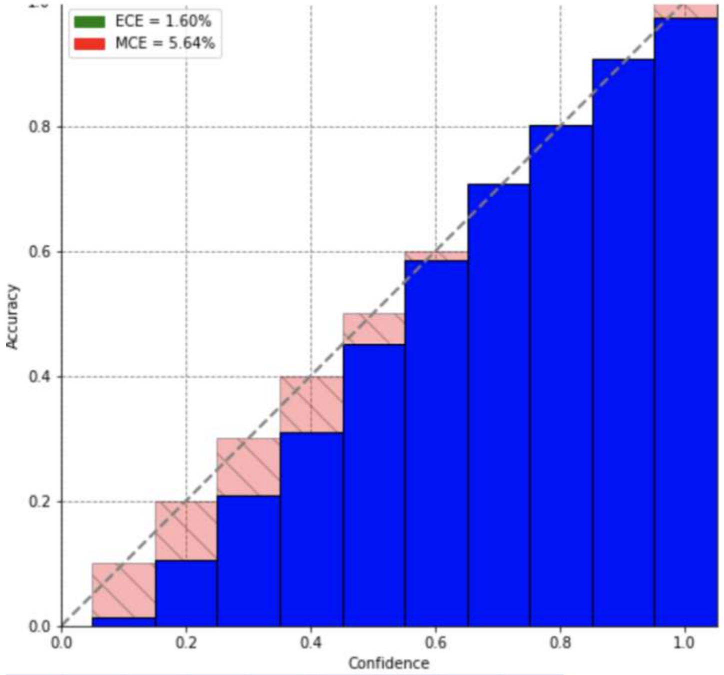
Definition 5.3. Let B_m be the set of indices of predictions that fall into the $m-th$ bin. The average confidence of B_m is

$$conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \tag{5.3}$$

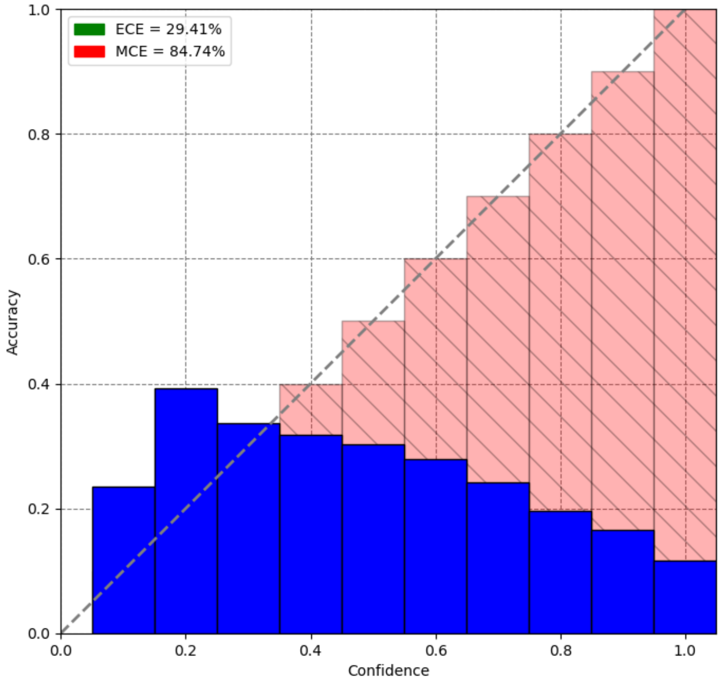
where $\hat{p}_i$ is the confidence of the $i-th$ prediction Guo et al. (2017).

A perfectly calibrated model's output will reflect ground truth likelihood, thus for a given bin, the confidence expressed by the prediction score will be equal to the accuracy, i.e. $conf(B_m) = acc(B_m), \forall m \in M$. If confidence and expected prediction accuracy are plotted against each other, the resulting histogram can be used to visually assess calibration. Perfect calibration corresponds to the alignment of the bars with the 45 degree line. 5.1 show an example of a very well calibrated reliability curve for a state of the art classifier 5.1a and an over-confident classifier 5.1b, respectively.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub



(a)



(b)

Figure 5.1: An example visualization of reliability curves. An alignment with the 45 degree line indicates perfect calibration. Under the the 45 degree line indicates over-confidence of the classifier. Over the 45 degree line indicates under-confidence of the classifier. (a) shows a very well-calibrated classifier and (b) a very poorly, over-confident classifier

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Reliability curves are an important tool in quickly and simply assessing a classifier's calibration. However, it does not numerically quantify the level of miscalibration. A more convenient method may have a scalar statistic that expresses how the classifier behaves. (Mis-)calibration can be numerically measured by calculating the **Expected Calibration Error** Naeini et al. (2015) and the **Maximum Calibration Error** Naeini et al. (2015).

Definition 5.4. The Expected Calibration Error expresses the expected value of a classifier's confidence and accuracy:

$$\mathbb{E}_{\hat{P}}\left[|\mathbb{P}(\hat{Y}=y|\hat{P}=p)-p|\right] \tag{5.4}$$

We can approximate miscalibration using 5.4 by binning the model's predictions into M equally spaced bins and taking a weighted average of the bin's accuracy/confidence difference Guo et al. (2017):

$$\Sigma_{m=1}^{M}\frac{|B_m|}{n}|acc(B_m)-conf(B_m)| \tag{5.5}$$

n being the number of predictions.

Proof. The exact definition of miscalibration is

$$\mathbb{E}_{\hat{P}}\left[|\mathbb{P}(\hat{Y}=y|\hat{P}=p)-p|\right]$$

Let $F_{\hat{P}}$ be the cumulative distribution function of $\hat{P}$ such that $F_{\hat{P}}(b)-F_{\hat{P}}(a)=\mathbb{P}(\hat{P}\in[a,b])$. Using the Riemann-Stieltjes integral on to re-express the expectation:

$$\int_0^1 |\mathbb{P}(\hat{Y}=y|\hat{P}=p)-p|dF_{\hat{P}}(p)$$
$$\approx \Sigma_{m=1}^{M}|\mathbb{P}(\hat{Y}=y|\hat{P}=p_m)-p_m|\mathbb{P}(\hat{P}\in r_m)$$

where r_m represents the interval of bin B_m. Looking at the summand and comparing it to the defintions 5.2 and 5.3 it can be seen that for large n $|\mathbb{P}(\hat{Y}=y|\hat{P}=p_m)-p_m|$ is approximated by $|acc(B_m)-\hat{p}(B_m)|$ Hence expected calibration error using M bins converges to the M-term Riemann-Stieltjes sum of $\mathbb{E}_{\hat{P}}\left[|\mathbb{P}(\hat{Y}=y|\hat{P}=p)-p|\right]$

□

It is also useful to analyze, and minimize, the maximum difference between confidence and accuracy versus the mean in 5.4 if reliable confidence measures are absolutely necessary Guo et al. (2017).

Definition 5.5. The Maximum Calibration Error expresses the greatest deviation between confidence and accuracy:

$$\underset{p\in[0,1]}{max}|\mathbb{P}(\hat{Y}=y|\hat{P}=p)-p| \tag{5.6}$$

The approximation of this summary statistic also involves binning:

$$\underset{m\in\{1,\dots,M\}}{max}|acc(B_m)-conf(B_m)| \tag{5.7}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Out of Distribution Detection

When deep learning classifiers are deployed in real-world situations, they often fail because they cannot distinguish when the distribution of data used differs too greatly from the distribution of the training data, i.e. it is *out of distribution* Cao et al. (2020). With medicine specifically, failure could result in wrong diagnoses or introduce biases that affect a healthcare professional's judgement Gal (2016). Despite the failure of classifiers, they still provide high confidence predictions while being effectively wrong Hendrycks and Gimpel (2016). These weaknesses significantly amplify the associated danger of using deep learning in a real clinical setting Goodfellow et al. (2014). Trying to classify an input as out of distribution, meaning it was sampled from a different testing distribution compared to the training distribution, is what is known as Out of Distribution Detection (OOD).

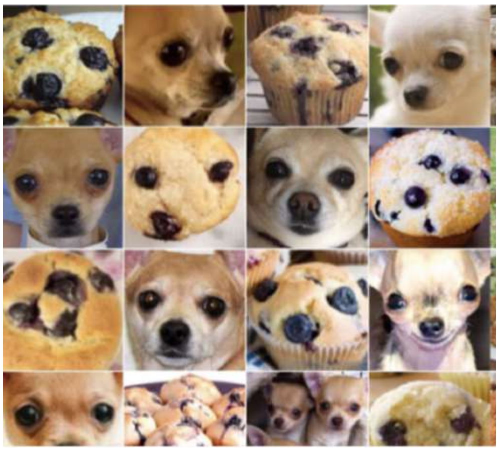


Figure 5.2: Visualisation of OOD: the goal is to classify an input as out of distribution, meaning it was sampled from a different testing distribution compared to the training distribution. If a model is trained on dog breeds, can it successfully distinguish between the blueberry muffing and the chihuahua?

Image 5.2 visually describes the OOD problem. The network in this example has been trained on images of dog breeds and can distinguish between them. The important question to consider: if it is given an image of a blueberry muffin, will it be able to distinguish between whether it is a muffin or a chihuahua? While this is a light example, it shows the principle of OOD very well. It is obvious to a human observer that blueberry muffins and chihuahuas are different; however, they do share some striking similarities that would lead a network to possibly classify a muffin as a dog. This can be easily extrapolated to more serious domains, such as chest radiographs with pathologies a diagnostics system has never observed before, leading to wrong diagnoses and follow up treatments that could be potentially harmful.

Before formally defining what Out of Distribution Detection is, it is necessary to understand how data is viewed in a statistical sense. The general assumption is that the population is generated by an underlying, unknown process Goodfellow et al. (2016). In order to derive information about the population, this process can be abstractly modelled by a probability distribution based on a known sample, *the* data. Thus, it is often useful to consider the dataset a distribution when dealing with probabilities/uncertainties. An observation, in this case an image, is considered in-distribution if it is part of the training data and out (of) distribution if it is not.

Out of Distribution Detection is in essence a separate binary classification problem that is applied after to a classifier after model training. In an abstract sense, OOD can be thought of as separating a test dataset into those examples that are part of the training distribution and those that are not. While there are different approaches to do this, Lee et al. (2018), this thesis focuses on the work of Hendrycks and Gimpel (2016) which uses a classifier's predictive score, such as the maximum softmax probability, more on this below. The score from the original model is used to assign in and out of distribution labels to the observations. The formal definition of the problem is as follows Chen et al. (2020a) Bendale and Boult (2016) Sehwag et al. (2019):

Definition 5.6. Chen et al. (2020a) Let $P(I)$ and $Q(I)$ be the **in-** and **out-**distribution on the space $\mathcal{I}$ of images, respectively. P and Q are sufficiently different and Q has a label set that is disjoint from that of P.

The out of distribution problem is characterised by a function that separates in and out of distribution samples. It is defined as

$$D := \mathcal{X} \rightarrow \{0, 1\}$$

this is a binary classifier that assigns in-distribution samples a 1 and out of distribution samples a 0, based on their classification scores. **in-**distribution samples are originally drawn from $P(I)$ and denoted as $\mathcal{D}_{in}$; **out-**distribution are samples drawn from $Q(I)$ and are denoted as $\mathcal{D}_{out}$

Threshold Free Metrics An important issue with this concept is the fact that if one class is more prevalent than the other, the model could have a high accuracy based on always guessing the class with the higher frequency - this can be misleading Hendrycks and Gimpel (2016). Inadvertently, this will lead to the issue of choosing a threshold that reflects a certain trade-off between false positives and false negatives. Therefore, one tool to assess OOD is the use of the Area Under Receiver Operating Characteristic (AUROC), which is a threshold-free metric Davis and Goadrich (2006). A ROC curve plots true positive rate against false positive rate.

Maximum Softmax probability baseline The baseline metric used in this thesis, the defacto standard metric developed by Hendrycks and Gimpel (2016), takes the maximum

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

probability from of a softmax distribution to determine whether an input can be classified as out of distribution or not. The metric is called maximum softmax probability metric. The rationale here is that incorrect and out-of-distribution examples have lower softmax scores than the prediction probability for correct examples Hendrycks and Gimpel (2016). Specifically, the test set consists of in and out of distribution examples and for each example, the maximum softmax probability is recorded. From this, AUROC values can be calculated and the score distributions for the two different groups can be visualized to assess the OOD capabilities.

Another approach to OOD is the use of models that estimate a data density to understand the data generation process. The idea is that a sample is out of distribution if it lies within low density regions Kingma and Welling (2013). The benefit of this method is the assumption that the density captures elements of the data generation process and has more information about the underlying structures and dependencies.

Ideally, there should be one modelling framework that can fit an accurate model and at the same time deliver the valuable uncertainty information that deep learning classifiers do not, possibly by ways of a combined density estimation. For instance, in addition to a class prediction, the model could produce a measure that conveys a high level of uncertainty, or low level of confidence if the image lies outside of the training distribution Gal (2016). Having more structural information about the data embedded into the modelling could increase trust in these methods an contribute to AI safety Amodei et al. (2016). Specifically autonomous detection of OOD examples has become an important component for trustworthy AI Amodei et al. (2016). Such frameworks exist and this thesis will focus on the development of one such framework called Energy Based Model.

In the case of the EBM, the intuition is that the energy function E is approximated by a CNN, meaning the (negative) energies in $E_w(a,b)$ are given by the logits. A natural then question is: how is the joint distribution and the (marginal) data distribution derived from the logits? The following derivation based on Grathwohl et al. (2019) explains this central connection.

$$G_w : \mathbb{R}^D \to \mathbb{R}^K$$

Let $G_w(a)[b]$ be the projection onto the y-th component of $G_w(a)$, known as the logits. Given that the energy function E in equation 4.5 is being interpreted as a neural network, $-E_w(a,b)$ can be understood to represent the logits $G_w(a)[b]$, with the parameter $\beta = 1$. The logits $G_w(a)[b]$ are not normalized, which fits into the energy based framework since energy is a-priori without any restrictions, especially normalization restrictions. Thus, equation 4.5 becomes:

$$P_w(a,b) = \frac{exp(G_w(a)[b])}{\int\limits_{b \in \mathcal{B}} exp(G_w(a)[b])da} \tag{5.8}$$

giving the joint distribution of inputs and labels. From this joint distribution we can construct the marginal distribution of just the inputs $P_w(a)$ by summing 5.8 across the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

labels b, i.e. marginalizing b out:

$$P_w(a) = \sum_b P_w(a,b) = \frac{\sum_b exp(G_w(a)[b])}{\int\limits_{b\in\mathcal{B}} exp(G_w(a)[b])da} \tag{5.9}$$

Which defines the distribution over the inputs. The joint and input distributions can be used to express the conditional distribution $P_w(b|a)$:

$$P_w(b|a) = \frac{P_w(a,b)}{P_w(a)} = \frac{exp\left(G_w(b)[a]\right)}{\sum_{b'} exp\left(G_w(a)[b']\right)} \tag{5.10}$$

The expression in equation 5.10 is the familiar softmax function Gibbs (1902), the multidimensional extension of the logistic function,

$$\sigma : \mathbb{R}^K \to [0,1]^K$$
$$z \mapsto \sigma(\mathbf{z})_b = \frac{exp(z_b)}{\sum_{j=1}^{K}}$$

where $z = G_w(a)[b]$. The softmax function normalises the network output to a probability distribution over the K different possible responses. It is the typical output layer for classification networks and assigns decimal probabilities to each class, where the probabilities must add up to 1. The class corresponding to the highest probability is selected as *the* class for the image in question. However, the number itself has no deeper meaning for the image, or the prediction in relation to the dataset overall.

The derivation above, formulated by Grathwohl et al. (2019), shows how a generative model lies hidden within a standard classification model, in this case a "normal" CNN's. The negative logits from the CNN are the output of the energy function, that is they are interpreted as energy. The energy is normalised based on the *whole dataset* to give the joint probability distribution $P(a,b)$, this is the key difference to the state of the art classifier that does not do this. The joint distribution can be marginalised to give the data distribution $P(a)$; the joint and data distributions can be used to express the conditional probability $P(b|a)$. The EBM in this context is considered a *hybrid model* and has come to be known as the Joint Energy Based Model (JEM) constructed by Will Grathwohl Grathwohl et al. (2019).

Remark. This treatment of logits to define a generative model within a classification model is not limited to neural networks. Any classifier can be used.

A deep learning classifier is trained in such that the pattern recognition skill of the CNN is leveraged to approximate a function/hyperplane that will group the data into the relevant classes and minimize the cost of misclassification. Classification occurs by producing a set of output scores and selecting the maximum one as the class for the specific image. The output score itself however, does not have any deeper meaning that could be attached to the image or related to the dataset overall. For instance, one cannot conclude that a

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

patient is more ill than another patient because they have a higher classification score. One can also not determine whether a given test image fits into the data distribution overall or not. Even so, the classification score is often erroneously interpreted as a reliable uncertainty measure for the specific image Guo et al. (2017). On a high level, the score itself is the result of a series of weighted sums passing through the different transformations, such as convolutions and activation functions, before being normalised by a softmax function across the number of classes. More formally, DL classifiers give point estimates for the weight parameters that minimize a specific loss. This works well in a tightly controlled, very specific domain but often does not reflect reality and the overall data distribution well which can result in significant drawbacks that directly influence real-world adoption. This includes the phenomenon where the model produces highly accurate predictions in experimental settings, but the predictions do not reflect likelihoods found in the data, i.e. the model is not well calibrated an over-confident. In addition, the model does not have the ability to identify when it is given an image that is significantly different to the training distribution, either through distribution shift or false inputs, also known as Out of Distribution Detection.

5.1.2 Energy Based Modelling

The key difference between the models is the way they quantify uncertainty. The EBM is trained in such a way that it learns the distribution of the data, i.e. it is a generative model, and this knowledge about the data is then used in downstream discriminative tasks, hence the term hybrid model.

The role of the CNN is to approximate the energy function, the representation of the latent dependencies within the data. The energy function by is an unnormalized distribution that can be transformed and normalized to a probabilistic model, i.e. it gives reliable probability estimates, using the Boltzmann distribution. The distribution is fit by using MCMC methods typically used in Bayesian model training and inference, where random samples are generated by many Markov chains and updated stochastically using the information about the function's gradient. This allows for a comprehensive quantification of uncertainty. The samples are compared to real images and the divergence/difference is minimized; in addition to learning the patterns present in the data, the model learns deeper structural information about the data.

The function's flexibility can be leveraged to subsequently produce classification scores that benefit from this deeper knowledge; as the distribution estimate iteratively improves, so do the class predictions. Because of this procedure, the classification scores contain deeper meaning and can be used to make reliable conclusions and assumptions about the inputs. Furthermore, the model can be used to make more reliable out of distribution estimates and deliver more calibrated predictions.

The distributional properties can be used to conclude how well an image fits into the data overall and generate new and realistic samples of data. While it does have many benefits, the EBM has key deficiencies that could make widespread use impractical.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

EBMs are very difficult to train, especially in very high dimensions LeCun et al. (2006). It involves a lot of manual effort to find the correct hyperparameters for all of the involved components (the network, the markov chains etc). In addition, EBMs need a lot more time and computational power to deliver useful results, compared to "traditional" deep learning. Since it is a growing field of research, at present there are only few established best-practices for training.

5.1.3 Summary

The similarities and differences of the deep learning and energy based frameworks is summarised in table 5.1

	Deep Learning	**Energy Based Modelling**
Pros	A lot of resources and best practices are available	Very flexible and applicable to many different types of problems and types of data
	Highly accurate on image recognition tasks, pattern recognition, outlier detection fast inference/prediction times	Allows reliable uncertainty quantification
	A lot of resources and best practices are available	One unifying framework that allows a generative model to also be used for discriminative purpose
	Scale well	No prior restrictions
	Fast and efficient training algorithms available	Increases model interpretability
Cons	Black Boxes	Not a lot of best practices available
	High data and hardware requirements	Highly computationally intensive
	Lack of reliable uncertainty quantification which influences OOD and model calibration	Requires a lot of manual tuning and high mathematical complexity
	Tend to only work well in very specific, narrow areas of application	High hardware requirements and do not scale well
		Can be unstable to train, i.e. loss often diverges

Table 5.1: Comparison of the deep learning and energy based frameworks

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU Bibliothek WIEN Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 6

Methodology And Model Design

This chapter is divided into three sections, each will describe a different aspect of the experiments carried out in the course of this thesis. To begin, section 6.1 will recall the overarching goals of this thesis and view them in the context of the theory developed in the previous chapter. The following section, section 6.2, is dedicated to all data used in the experiments. It will detail the chest x-ray data used for training, as well as the additional datasets, both medical and non-medical, used the evaluation phase of the model. In the final section of this chapter, section 6.3, the implementations of both the state of the art classifier, 6.3.4, and the EBM, 6.3.4, are addressed. This includes a detailed description and formal justification of the choices made when constructing the two pipelines, such as the reasoning behind the choice of pre-processing, what network architecture was used, how the different models were trained and, finally, how the results were evaluated such that they can be compared to the aim of the thesis.

6.1 Contextualization Of Research Goals

The rationale for conducting the research in this thesis is to study the suitability of a hybrid energy based model, as constructed in **Your Classifier is Secretly an Energy Based Model and You Should Treat It Like One** by Will Grathwohl Grathwohl et al. (2019). The idea is to leverage the information captured by the *generative* model to create a medical image *classifier* that delivers highly accurate predictions that also contain uncertainty information, hence the hybrid nature of the EBM. An additional innovation this thesis brings, is the fact that EBMs have never been applied on the medical domain before.

In relation to the aims formalised out in 1.3, a chosen state of the art deep learning classifier is trained on posteroanterior chest x-rays. The same neural network architecture will be reinterpreted within the energy based framework and examined for its discriminative and out of distribution capabilities. Specifically:

TU Bibliothek
WIEN
Your knowledge hub

- the discriminative power of the models will be compared using the respective AUROCs, where the AUROC is the area under the ROC curve and an aggregate measure of performance across all possible classification thresholds.

- a significant benefit of the EBM, especially compared to a classical state of the art deep learning classifier, is the ability to detect of wrong inputs very well, which is also known as Out of Distribution Detection Hendrycks and Gimpel (2016), see also 5.1.1. OOD will be evaluated with maximum prediction probability score, according to Hendrycks and Gimpel (2016) because it allows comparability between the state of the art classifier and the EBM.

- the distributional properties should allow for the "out of the box" calculation of calibrated likelihoods, that is the predicted confidence such as a logit is representative of the true likelihood seen in the data. Calibration will be compared using the Expected Calibration Error, see 5.1.1.

In all of the above aspects, the EBM will be compared to a a state of the art discriminative classifier. The hypothesis is that the energy based setup yields comparable discriminative power while meeting significant requirements for applications in clinical setups.

The foundational theory on EBMs, presented in 2.3, details how the models in question work and how they can be used to address the above points. EBMs were presented as a useful and flexible tool for generative modelling, which is what this thesis hopes to leverage for downstream discriminative tasks. They estimate a, not necessarily normalised, density over the data, which allows for the calculation of likelihoods, confidence intervals and other quantities that rooted in the data and quantify uncertainty. As a special case, they can be viewed in a probabilistic way, meaning they give a probability distribution, which allows for the calculation of likelihoods and the construction of confidence intervals to assess uncertainty - all of which gives the user valuable understanding about the latent connections in the data. This thesis focuses on exactly this special case of the energy based framework. These models have become increasingly popular to explore, due to their mathematical simplicity Du and Mordatch (2019), especially how well the estimated density can be used for downstream discriminative tasks, such as multiclass classification Grathwohl et al. (2019). The motivation is to use the added information the generative nature of the model provides for better quality predictions. In short, EBMs allow to correctly quantify uncertainty, which has potentially large benefits for AI safety and real-world adoption of the concept Amodei et al. (2016). This idea forms the central part of the research carried out in this thesis, with the capabilities of the EBM directly compared to a state of the art classifier.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

6.1.1 Network Definition

After presenting the foundational theory and providing the context for the research in this thesis, it is important to understand what learning means conceptually and provide the specifics of our set up within that concept. Tom Mitchell Mitchell and Mitchell (1997) defines (machine) learning in the following way:

Definition 6.1. A computer program is said to learn from experience with respect to some class of tasks and performance measure, if its performance at the tasks, as measured by a relevant measure, improves with experience.

This definition can be applied in the following way:

- **The Task:** the task in this thesis is to classify chest radiographs. Thus, the task of finding a classifier means approximating a function G that will correctly map x-rays to their disease or category. Mathematically, this function can be defined as:

  Definition 6.2. Let $\mathcal{I} := \{I_1, ..., I_n\}$ be the set of $n \in \mathbb{N}$ chest radiograph with pixels $\mathbf{x} = (x_1, x_2)$ on a discrete grid $m \times m$, $m \in \mathbb{N}$, and intensities $I_i(\mathbf{x}) \in \mathcal{A} \subset \mathbb{R}$ Novikov et al. (2018). Each image $I \in \mathcal{I}$ has a corresponding label l describing the disease present, where $l \in \mathcal{L} := \{l_1, \ldots, l_k\}$, $k \in \mathbb{N}$. Let $\mathcal{Y} := \{y \colon y \in \{0, \ldots, k-1\}$, where $k \in \mathbb{N}\}$, be the set of integers such that $y_i \in \mathcal{Y}$ represents the $i-th$ radiograph's label l_i. The values in $\mathcal{Y}$ correspond to the labels $\mathcal{L}$ of the images $\mathcal{I}$. Let $\hat{y}$ be the network G's predicted label, based on the image, then

  $$\begin{aligned} G \colon \mathcal{I} &\to \{0, \ldots, k-1\}, \\ &\mapsto G(I) = \hat{y} \end{aligned}$$

  Alternatively, the function G could give a probability distribution over the classes. In this case, the range would change to:

  $$\{\hat{y} \in \mathbb{R}^k \colon 0 \leq \hat{y}[r] \leq 1,\ r \in \{1, \ldots, k\} \wedge \Sigma_{i=1}^{k} y[i] = 1\}$$

  That is, the output $\hat{y}$ of the function $G(I)$ would be a $k-$dimensional vector where the $r-$th element, corresponding to the $r-th$ semantic label, lies between 0 and 1, and whose elements sum to 1. The predicted label would then be the maximum element of $\hat{y}$

- **The Performance Measure:** In order to determine the performance of an algorithm, an appropriate quantitative measure must be selected. For classification this is often accuracy, the proportion of correctly classified inputs Goodfellow et al. (2016). However, accuracy can be a misleading measure depending on the dataset and its balance. Measures that are particularly important in medical classification

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

are: sensitivity/recall, or the true positive rate, i.e. the probability of a positive test, or $\frac{\text{no. of true positives}}{\text{no. of true positives + number of false negatives}}$, conditioned on being truly positive, and specificity, or the true negative rate, i.e. the probability of a negative test , or $\frac{\text{no. of true negatives}}{\text{no. of true negatives + number of false positives}}$ conditioned on being truly negative Hastie et al. (2009). These rates are used to construct threshold-free metrics such as the Area Under Receiver Operating Characteristic (AUROC) which are more reliable when datasets are unbalanced Hastie et al. (2009). The area under the curve that shows a classifier's true positive rate against its false positive rate.

- **The Experience:** models are differentiated by what information they are exposed to during training. Generally speaking, the main methods for training models are: unsupervised and supervised learning, though there are also other important methods such as semi-supervised or reinforcement learning. This thesis only deals with supervised learning, meaning that the model receives information about the true class label, or disease type, of each chest x-ray. This can be formalised as:

  Definition 6.3. Let $\mathcal{S} \coloneqq \{(I_1, y_1), \ldots, (I_n, y_n)\}$ be a finite sequence of pairs made up of images and labels known as the *training dataset*, where $S \subset \mathcal{I} \times \mathcal{Y}$ with $\mathcal{I}$ and $\mathcal{Y}$ defined as above in 6.2. The element y_i represents the $i - th$ x-ray's label as an integer that corresponds to the labels $\mathcal{L}$.

6.2 Data Exploration

All data used within this research is publicly available data. Training the models is carried out using the chest radiograph dataset CheXpert Irvin et al. (2019). During evaluation, both medical and non-medical images were used, depending on the specific research question under consideration:

- To evaluate the discriminative power of both models, a test set from CheXpert is separated out before training. When carrying out the random partitioning of the data, it was made sure that the test set has no overlapping patients.
- To evaluate the calibration of both models, the test set from CheXpert is used.
- To evaluate and compare the OOD abilities of the models, three different datasets that increasingly differ in similarity to the original CheXpert data. The intuition behind using three increasingly dissimilar datasets for the evaluation of OOD, is that it allows us to gauge how, if at all, the models progressively improve separating the original data, the in distribution data, and the out of distribution data, the data that is completely different than the training data. It is expected that the EBM, with its deeper understanding of the latent connections within the data, will be able to separate all the datasets very well. The out of distribution datasets include:

  1. ChestXray14 Wang et al. (2017)

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

2. INbreast Moreira et al. (2012)
3. ImageNet Russakovsky et al. (2015)

Figure 6.1 shows a small extract from each of the out of distribution datasets.

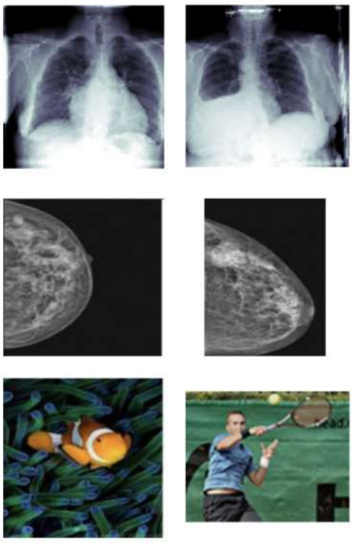


Figure 6.1: samples of the out of distribution datasets; top is the ChestXray14 data; middle is INbreast data; bottom is ImageNet data

6.2.1 CheXpert

Both the EBM and the state of the art classifier are trained on CheXpert data Irvin et al. (2019). CheXpert is a dataset consisting of 224,316 chest radiographs of 65,240 patients who underwent a radiographic examination from Stanford University Medical Center between October 2002 and July 2017, in both inpatient and outpatient centers. The radiographs are labelled for the presence of 14 common chest radiographic observations. The experiments in this thesis are concentrated on only four different pathologies: cardiomegaly, pleural effusion, fracture and lung lesion. These were chosen because their respective manifestations in the radiographs are typically all sufficiently different from each other, so that a non-medical professional could differentiate between them.

Technical details

The dataset consists of a mixture of frontal and lateral images. The resolution is, in general, not standardised across the dataset. The original raw images typically have a size of at least 1024 × 1024; however, to efficiently work with them a separate, downsized dataset is also available for download. In this version of the dataset, also known as "CheXpert small", the images are reduced to a size of at at least 320 × 320, this is also not

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

standardised across all images. For the purposes of our experiments we used "CheXpert small". The images were originally saved in the DICOM format, the standard for storing medical images, but were converted to 8 bit .png images before being made publicly available, and the corresponding paths are stored in accompanying csvs that also include and information about the 14 pathologies. Each row of the file represents an image with each column describing the patients age, gender and what (if any) pathologies are present. A pathology-column can contain one of four labels:

- -1 = uncertain
- 0 = negative
- 1 = positive
- NA/blank = unmentioned

The labels were assigned by a separate labelling model developed by Stanford, whose output was evaluated by two board-certified radiologists. A key decision in the set up for the classification task is how to deal with the uncertainty and unmentioned labels. Unmentioned cases are interpreted as the patient in question not having a certain illness. The authors of the original CheXpert paper explored different ways of dealing with uncertainty labels, before ultimately carrying out their discriminative experiments by mapping the "uncertain" (=0) labels to the "present" (=1) labels. This is similar to zero imputation and aims to mimic approaches where missing examples are used as negative labels Kolesov et al. (2014). This approach is more risk-averse and makes intuitive sense considering the sensitivity of the medical domain. In this thesis the choice was made to emulate the CheXpert authors and consider the uncertain cases as cases that have the specific pathology. This also makes sense in an experimental setting; the state of the art discriminative classifier is intended to reproduce the CheXpert results and act as a benchmark for the EBM. Taking similar steps in data manipulation ensures comparability amongst the models.

Exploration

The following section delves into the data to explore its attributes. To understand the overall structure of the data, first patient-specific variables such as age and gender will be analysed, followed by more general structural variables such as and x-ray type will be explored.

Gender	**Amount**
Male	132764
Female	90883
Unknown	1

Table 6.1: Table showing the distribution of gender in the CheXpert data

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The distribution of gender is shown in table 6.1; male patients significantly outnumber the female patients. The distribution of age is shown in figure 6.2. It is skewed towards the right, with most of the mass located at 40 years and up and the mean being at approximately 60 years of age.

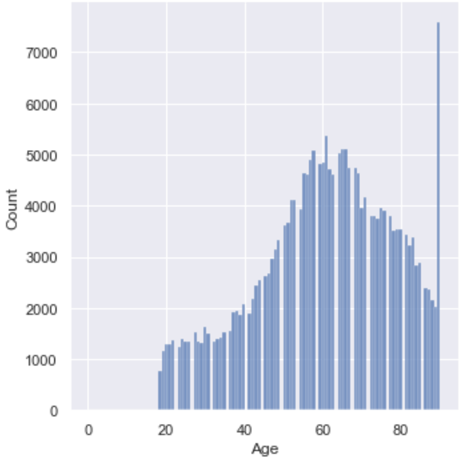


Figure 6.2: Histogram of ages in CheXpert

Moving on from patient-specific attributes to look at more general structural aspects such as image type and the prevalence of different uncertainty labels. The images themselves are made up of both frontal and lateral views, but not for all patients equally.

View type	**Amount**
Frontal	191229
Lateral	32419

Table 6.2: Table showing the distribution of x-ray views in the CheXpert data

Table 6.2 shows that the number of frontal views dwarfs the number of lateral views. They will be filtered out in pre-processing to maintain comparability to the state of the art.

Table 6.3 splits the frontal and lateral views by the relevant pathologies. We can see that there is a large imbalance between the pathologies. Lung lesions and fractures are the two pathologies with the lowest prevalences, with each at around 8000 frontal images. Cardiomegaly appears around 4×, and pleural effusion over 10× more often, respectively. Considering this, the pre-processing of the data and evaluation of the results will need to be performed such that the imbalance is taken into account. For instance: after randomisation, there must still be enough images of each category in the training data to train the model on.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Pathology	**Frontal**	**Lateral**
No Finding	17000.0	5419.0
Cardiomegaly	30158.0	4997.0
Lung Lesion	8149.0	2526.0
Pleural Effusion	86541.0	11341.0
Fracture	7935.0	1747.0

Table 6.3: Table showing the occurrence of pathologies, split by the type of view (frontal/lateral)

Table 6.4 shows the prevalence of the "uncertain" ($= -1.0$) labels in the data before imputation. As described in the previous section, "uncertain" labels will be changed to the "present" label. This means we remain conservative, which makes intuitive sense in a sensitive environment. In addition, we also remain comparable to the literature, that also employed this data preprocessing step Irvin et al. (2019) Baltruschat et al. (2019).

Pathology	**-1.0**	**0.0**	**1.0**	**prevalence (-1)**
No Finding	–	201229.0	22419.0	0.0
Cardiomegaly	8087.0	188493.0	27068.0	3.6
Lung Lesion	1488.0	212973.0	9187.0	0.7
Pleural Effusion	11628.0	125766.0	86254.0	5.2
Fracture	642.0	213966.0	9040.0	0.3

Table 6.4: Table showing the prevalence of the "uncertain" label in the data before imputation

Figure 6.3 shows a small subsample of images from each illness, that is **cardiomegaly, lung lesion, fracture and pleural effusion**, the 'No Finding' category is also included.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

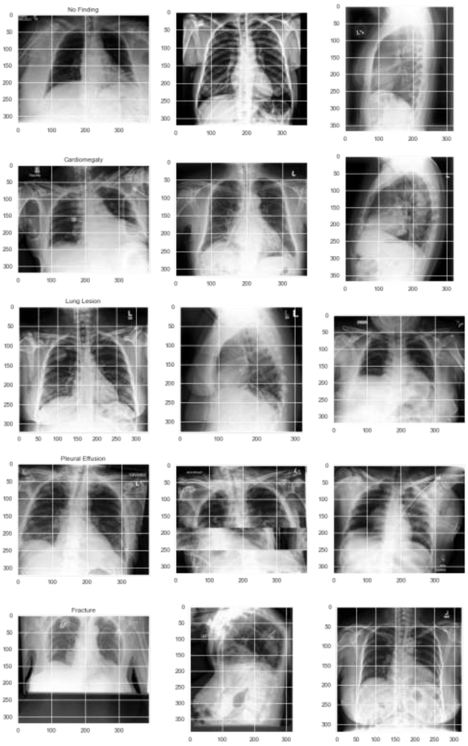


Figure 6.3: Grid of images showing sample images of each illness

It can be seen that some of the images contain imperfections, which can happen when the patient moves or the technician does not carry out the x-ray properly. If these images are too contaminated, they could be considered as anomalies. The following section will delve into another important structural aspect of data exploration: outlier detection.

6.2.2 Outlier Detection

Real world data almost always contains imperfections **?**, that is it contains a subset of observations that appear to be inconsistent compared to the majority of the remaining data Rousseeuw and Leroy (2005). These elements, in this case images, are known as outliers Barnett and Lewis (1984). Formally, observations are considered as outliers if their values are outside the range of variance estimated from the data Huber (2004). For chest x-rays, outliers often come in the form bad quality images as is seen in 6.3. Any learning model will be affected by outliers in some way Rousseeuw and Leroy (2005). How and to what extent depends on the model **?**. Neural networks are universal approximators

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

3.4.2 and may be at risk of overfitting by learning from outliers. For instance, a feature has a variance orders of magnitude larger than others, it might dominate the objective function, meaning it is unable to learn from other features LeCun et al. (2012). Thus, in order to better understand the data, it is important find and analyse outliers. In this thesis we will be using one of the most widely used statistical tools to identify outliers: Principal Component analysis (PCA). Due to PCA's sensitivity to very strong outliers Huber (2004), its robust version of PCA is also used Jolliffe (1986), Hotelling (1933), Eckart and Young (1936).

Given a data matrix $M \in \mathbb{R}^{n \times n}$ and its decomposition: $M = L_0 + S_0$ into a low-rank matrix L_0 and (sparse) matrix of perturbations S_0, PCA seeks to find the best rank-k estimate of L_0 by solving

$$\min_{L} \quad ||M - L|| \text{subject to} \quad rank(L) \leq k$$

Where $|| \cdot ||$ denotes the Frobenius Norm. This problem can be efficiently solved via the singular value decomposition (SVD) Candès et al. (2011).

Applying PCA reduces the dimensionality of the data down to the chosen number of principal components. The components can then be used in the inverse transform to re-transform the data from the PCA space back into the data space. Using the inverse transform inevitably causes information from the remaining dimensions to be lost Jackson (2005). In the context of outlier detection, this can be interpreted as applying a filter to find those images with the noisiest features. The pixel-wise MSE between each image in the original data and the inverse-transformed data allows us to find those images in the original dataset that show the highest inconsistency Singh and Kumar (2016). Carrying out PCA on the CheXpert dataset for outlier detection was carried out in Python by:

- calculating the PCA transform for a given number of principal components; this was done by converting the images to their matrix form and linearising them row by row, such that the final dataset before applying PCA contained one image per row with 224 × 224 columns.

- applying the inverse transform on the transformation back to the original space.

- calculating the pixel-wise MSE between each image in the original and the inverse-transformed data

This method yields the outliers shown in figure 6.4

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

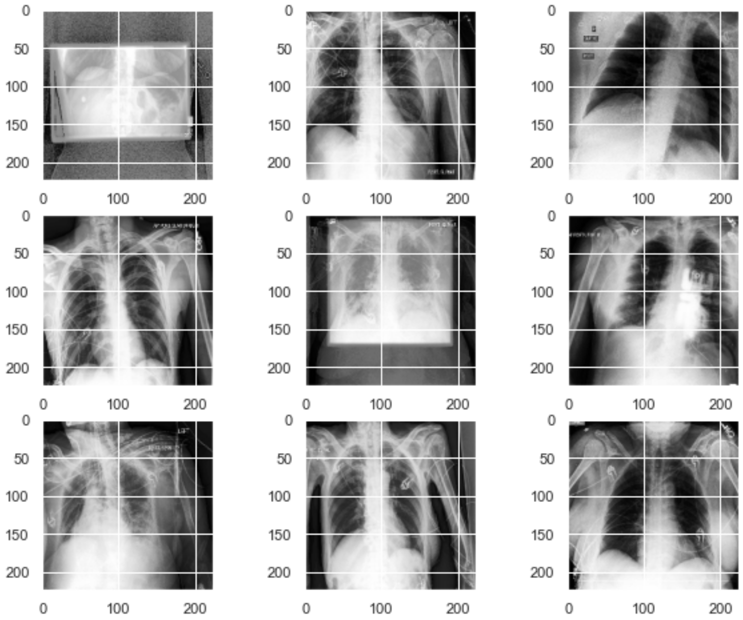


Figure 6.4: A grid of the top 9 outlying images in the chexpert dataset as determined by PCA

However, PCA itself is sensitive to grossly miscalculated observations Candès et al. (2011). This brings up the need to robustify PCA. Following Candès et al. (2011), robust PCA is achieved by solving the following convex optimisation problem:

$$\begin{aligned} \underset{L,S}{\text{minimize}} \quad & ||L||_* + \lambda ||S||_1 \\ \text{subject to} \quad & M = L + S \end{aligned}$$

The above problem is called Principal Component Pursuit (PCP) and can exactly determine L and S Candès et al. (2011). $||\cdot||_*$ is the sum of singular values (the nuclear norm) and $||\cdot||_1$ is the ℓ_1 norm. That is, the maximum (absolute value) column sum of a matrix. The PCP problem can be solved efficiently, at a cost similar to classical PCA. Empirically, Candès et al. (2011) has shown that this works under broad conditions for many types of real data. Using the same methodology to find outliers as like in the classical case, applying PCP to CheXpert yields the outliers shown in figure 6.5.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

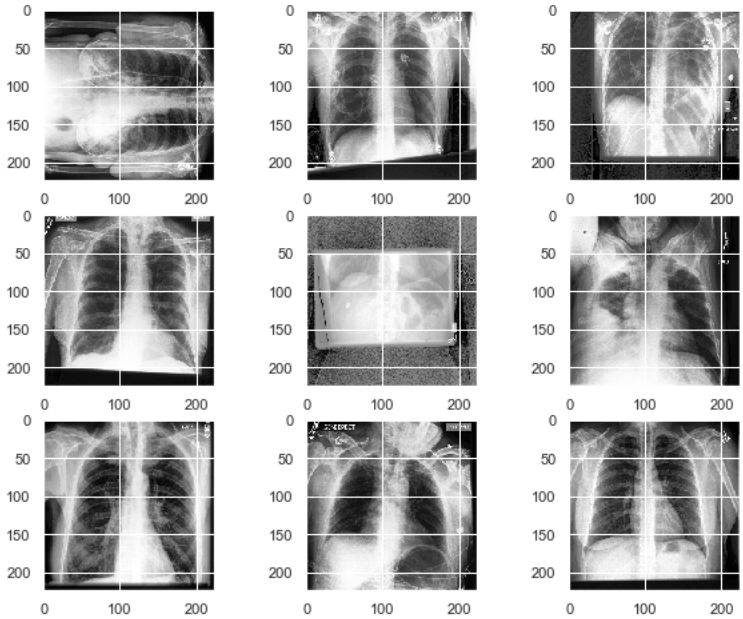


Figure 6.5: A grid of the top 9 outlying images in the chexpert dataset as determined by robust PCA

Looking beyond the finding of individual outliers, (robust) PCA can also be used to compare datasets. By plotting the distributions of the MSEs that occur when the inverse transform, the reconstruction of the dataset based only on the principal components, is compared to other image datasets, PCA's ability of detecting different types of datasets can be visualized. This implies, that PCA can, in a way, also be seen as a simple, non-parametric, non-modelling tool for OOD. It is an interesting theoretical approach to consider alongside the more sophisticated modelling approaches.

Overall, PCA and robust PCA was used to extract and visualise some of the most outlying elements.

6.2.3 ChestXray14

This dataset, published by the U.S. Department of Health and human Services, also contains chest x-rays. It contains over 100.000 anonymized chest x-rays from more than 30,000 patients across 14 different illnesses. The scans were collected at the NIH clinical center, the US's largest hospital devoted entirely to clinical research. The x-rays images are directly extracted from the DICOM file and resized as 1024×1024 bitmap images.

This dataset was used as part of the out of distribution experiments. Although this dataset also contains chest x-rays, these images were created under completely different circumstances as the CheXpert data, thus this dataset constitutes a separate distribution.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The differences can be technological or epidemiological, meaning the machines that took the images, the post processing, the population under consideration, all have an effect that distinguishes it from CheXpert. The rationale behind using this dataset is to observe whether the models are sensitive enough to distinguish between these different types of chest radiograph datasets and pick up on these differentiating factors that are not visible to the naked eye.

6.2.4 INbreast

The INbreast dataset contains mammographic scans from a breast center located in a university hospital in Portugal. The data includes 115 cases (410 images) of breast cancer of various degrees of severity, as well as mastectomy patients. The scans were made with a MammoNovation Siemens full-field digital mammography machine. The image sizes were 3328 × 4084 or 2560 × 3328 pixels, depending on the breast size of the patient. Images were originally saved in the DICOM format, but later saved as the NumPy arrays for more efficient data loading.

This dataset was used as part of the out of distribution experiments. The rationale behind using this dataset is that, although the domain is once again x-rays and, in general, a medical one, mammograms and chest x-rays are completely different types of medical images. We expect to see that the models are able to distinguish these from the CheXpert chest x-rays.

6.2.5 ImageNet

ImageNet is an image database of over 14 million images in over 20.000 categories that has been instrumental in advancing computer vision and deep learning research. The images are annotated by humans and each category can have thousands of images under it. ImageNet contains high quality, full resolution images with an average size of around 400 × 350 pixels, saved in the .jpeg format.

This dataset was used as part of the out of distribution experiments. The expectation behind giving the models images so far away from the medical domain is that they should easily be able to separate the datasets.

6.3 Implementation

6.3.1 Environment

The programming language used for all coding work was Python, specifically Python version 3.9. A lot of advanced and high-performance frameworks for deep learning and image processing are available in the environment. The machine learning framework used in this thesis was PyTorch Paszke et al. (2017), an open source machine learning framework based on the Torch library. It is an optimized tensor library for deep learning

TU Bibliothek, WIEN Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

using GPUs and CPUs used for applications such as computer vision and natural language processing, primarily developed by Meta AI.

The network training was carried out on two GPUs; one NVIDIA GeForce GTX1080 Ti GPU with 11 GB and CUDA version 11 and one with an NVIDIA A-100 GPU with 40 GB memory and CUDA version 11.

For the data pre-processing functionalities within PyTorch were used, as well as data processing packages NumPy and Pandas. In order to deal with the image data, the Python Imaging Library (PIL) was used, as well as OpenCV.
The following section will describe how the models in question were implemented. The training procedures can be broken down into distinct phases, where some are shared among the different approaches and some are relevant only to a specific framework.

- **Data Pre-processing**: This covers loading and augmenting the images and gathering the image-label pairs in preparation to be fed into the model. The same steps are required for moth modelling approaches
- **Model Training**: Training is done on CheXpert images following 1. state of the art deep learning and 2. EBM methodologies for the classes pleural effusion, fracture, lung lesion, cardiomegaly.
- **Evaluation**: the models' discriminative power, calibration and out of distribution detection is evaluated by calculating various metrics.

6.3.2 Data Pre-Processing

For all types of statistical learning and all types of data, the quality of the training data significantly influences the performance of the model. Thus, model quality is highly dependant on data preparation Litjens et al. (2017). By adding pre-processing steps, data inconsistencies such as outliers and noise can be corrected, as well to improve generalization and reduce over-fitting Kotsiantis et al. (2006). Based on the data exploration in 6.2, it is evident that the dataset is mainly made up of frontal x-ray images. In order to be comparable with the state of the art Irvin et al. (2019), lateral images will be filtered out in pre-processing.

Data Augmentation

Formally, data augmentation is a technique where transformations are applied to the training data to artificially create new training examples Perez and Wang (2017). In the training process for each type of model, the data was augmented in exactly the same way. The images were expanded to three channels and randomly split into batches, each with a size of 32 images, based on Baltruschat et al. (2019), before being augmented the following way Baltruschat et al. (2019):

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

- **Horizontal Flipping**: images in a batch were randomly flipped along the horizontal axis with a probability of 0.5 Baltruschat et al. (2019).
- **Rotations**: images in a batch were randomly rotated between ±7° Baltruschat et al. (2019)
- **Re-Scaling and Normalisation**: Networks process their inputs using small weights meaning that inputs with large pixel values can disrupt the learning process leading to lower accuracies and prolong convergence during loss minimization or, in extreme cases, cause the loss to diverge. Furthermore, images are standardised to z-scores, subtraction of the mean and division by the standard deviation, to make training more stable. This is done on a per channel level Goodfellow et al. (2016).

6.3.3 Model Design

This section describes how the two modelling setups were designed and trained. This thesis compares the classic state of the art approach with the energy based approach, thus the following sections cover the training of both models with a specific focus on how they differ and what this implies for the resulting predictions.

Model Architecture

At the heart of both the classical deep learning approach and the energy based approach, is the Convolutional Neural Network. Multiple CNN architectures were setup and tested in different ways during experimentation, including DenseNet Huang et al. (2017), ResNet He et al. (2016) and WideResNet Zagoruyko and Komodakis (2016). As explored in 3.6, the overall benefit of these networks is that firstly, they are very good at recognising features in visual inputs, thus achieving very high accuracies. Secondly, they efficiently solve issues that can arise when adding more and more layers to a network.

Considering accuracy, efficiency, and model complexity, our experimentations showed that the WideResNet (WRN) architecture performed better than the other two in terms of discriminative power, confirming findings in Zagoruyko and Komodakis (2016). In addition, the WideResNet has also been successfully employed in an energy based context Grathwohl et al. (2019) Du and Mordatch (2019). Thus, the WRN will be trained as a state of the art deep learning classifier and as an EBM for our final experiments. The deep learning classifier was trained from scratch as well as using a pre-trained network, while the EBM was trained from scratch. Using a so-called pre-trained network means that the network weights from a network that was previously trained on a very large, general dataset are loaded and subsequently trained on the task-specific data, this is known as transfer learning . Using pre-trained networks is proven to greatly increase efficiency in training while also delivering highly accurate results in the specific domain by taking advantage of previously learned features Shen et al. (2017). In practice, networks are typically pre-trained on the ImageNet dataset Russakovsky et al. (2015) as it is

TU Wien Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

the largest, most general dataset currently published. This method as also successfully been used on classifiers trained on the chest x-ray data Baltruschat et al. (2019). This technique also helps solve the problem of too little data, something often faced in the medical domain and it makes the OOD comparisons with ImageNet very interesting.

The final network was WideResNet 50 × 2, meaning the network was 50 layers deep and 2 layers wide. For the DL model a WRN implementation from PyTorch is used. For the EBM model a custom network is created that realised the idea of the hybrid model, see figure 6.6.

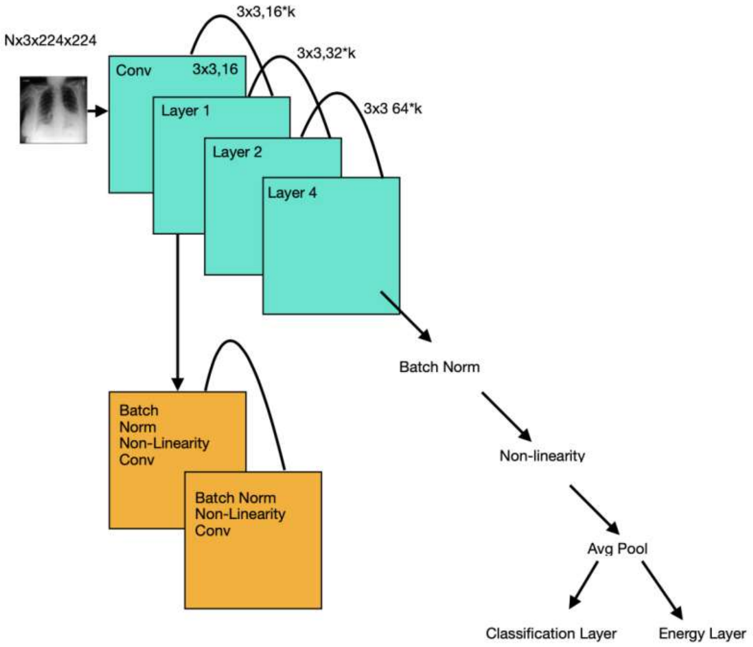


Figure 6.6: The hybrid model visualised schematically. The wide basic blocks each follow the structure shown in the detail diagram on the bottom left. The factor k defines how wide the network is. What sets this architecture apart is the hybrid output; the network has both an energy and classification output layer

The custom network has a WRN architecture with two output layers: one for energy output and one for class output. As the model improves its density estimate, the subsequent classification accuracy improves as well. The model definition is characterized by the widening factor k. If $k = 1$ the network is a standard ResNet.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

6.3.4 Model Training

Statistical Considerations

In order to ensure the validity of the experiments, various statistical considerations had to be made. The considerations encompass correct *preparation of the data* and *randomisation*. The relevant steps were applied to both modelling approaches and developed based on literature such as Baltruschat et al. (2019) and Hastie et al. (2009).

Preparation of The Data

The raw dataset is structured in a multi-label format meaning that an image may belong to multiple classes. In line with the literature used for the purposes of this thesis, the dataset is recast to a multiclass format **?**. Furthermore, the labels are encoded so that each category is represented by an integer that corresponds to the index of the respective class in the list of class names. In other words, the set of all responses is $\mathcal{Y} := \{y|y \in \{0, 1, 2, 3\}\}$. This is done for both approaches.

Data Randomisation

Neural network parameters are initialized randomly, this can result in increased variance of the resulting estimator which can impact the generalization capability of the model Hastie et al. (2009) Molinaro et al. (2005). In order to control for the inherent randomness, a resampling scheme is performed Baltruschat et al. (2019). The data is re-sampled 10-times; within each sample, the data is randomly partitioned into training (90%), validation (5%) and test (5%). This methodology was adapted from Baltruschat et al. (2019) and Ahn (2015). Furthermore, it is necessary to ensure that there is no overlap between between patients. This ensures that the model does not receive information that actually should not be available to it, also known as target leakage. The data is organized by patient IDs and subsequently randomised as described above.

There is no clear directive in the literature as to the how many re-sampling rounds and what size of splits to use, this is highly dependant on the data and the specific use case. For instance, in Baltruschat et al. (2019), the number of re-samples is only 5, with a smaller training set and larger test and validation sets. In principle, a higher number of re-sampling rounds typically implies that the results are more stable. Thus, the decision was made to re-sample and run the model 10-times, which was deemed as a sufficiently large number to guarantee stable results. However, training these big, complex models many times becomes very time consuming which makes running the models multiple times infeasible - this is case with the EBM where one Epoch could take up to a week to train. The complexity of the calculations needed to estimate the distribution, the many parallel markov chains generating samples, calculating gradients etc, are such that training an EBM multiple times becomes too time consuming and ultimately infeasible, especially for the scope of this thesis. However, while the calculations used in the approximations are computationally intensive, the MCMC methods used typically already deliver stable

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

results.

The size of the data partitions was chosen with respect to the prevalence of the individual pathologies. Table 6.3 shows how often the pathologies occur, categorised by the type of view. Looking at the number of frontal views, the only view under consideration in this work, the amount of lung lesion and fracture images is small compared to the other pathologies. In order to ensure that the network also sees enough examples of this type of illness, the training data is made up of 90% of the data.

Consideration of Dataset Size

The experiments were conducted in a two stage process: an initial proof of concept on a small subset of the data was conducted to pre-determine whether the hybrid EBM would deliver useful results on this type of image data and to observe how difficult the training would be, given that the existing literature often reported unstable training.

Resolution

The models are trained on three different image sizes in order to investigate what, if any, effect the image size has on accuracy and efficiency. The image sizes will become progressively smaller, starting with 224×224 pixels before getting reduced first to 128×128 and finally to a resolution of 64×64.

Training A SOTA discriminative Deep Learning Classifier

CNN Model

The architecture used for the training of the deep learning classifier is a WideResNet, originally implemented by PyTorch Zagoruyko and Komodakis (2016). The network is considered in two variants: pre-trained and "from scratch". For the pre-trained network, all layers re-trained, i.e. fine-tuned, and the final layer is adjusted to give a softmax output for each class. Previous experiments on chest x-ray data, such as in Baltruschat et al. (2019), have shown that the method fine tuning delivers highly accurate results and also makes intuitive sense, as the network needs to learn the specifics new domain.

The raw x-ray images only contain 1-channel, however the network is designed to expect 3-channel images, meaning either the network or the images must be modified to fit together. Ultimately, it was decided to modify the x-rays to be 3-channel images, mainly due to flexibility reasons. In the OOD evaluation phase, different types of datasets will be considered, including the 3-channel ImageNet dataset. In this setup, the network can easily be applied to the ImageNet data without larger manipulations to the images.

Loss Function

This thesis uses the negative loglikelihood loss for training the DL model, the defacto standard loss function for multi-class classification problems. As elaborated in section 3.5, minimizing NLL is equivalent to minimizing KL-divergence and crossentropy loss.

Given n sample pairs (I_i, y_i), $i = 1, \ldots, n$, with I_i being the $i-th$ image, y_i the corresponding label encoded as an integer, and weights w, the NLL loss J is expressed by:

$$\begin{aligned} J \colon \Omega &\rightarrow [0, +\infty] \\ (w) &\mapsto J(w) = -\mathbb{E}_{(I_i, y_i) \sim \hat{p}_{data}} \left[log \left(p_{model} \left(y_i \mid I_i; w \right) \right) \right] \end{aligned}$$

Optimiser

The optimization algorithm was chosen to be the popular Adam optimiser Kingma and Ba (2014), with an initial learning rate of 1e−4. In this optimization algorithm, running averages of both the gradients and the second moments of the gradients for every parameter are used. The best model with the lowest validation loss of all epochs is saved.

Training An Energy Based Model

There are very few best practices for EBM hyperparameter tuning in the existing literature, meaning they had to be monitored and adapted manually using a "trial and error" approach which could increase instability in the training process and cause the loss function, see below, to diverge.

Energy Function

In this thesis the energy function is parametrized by a CNN, specifically a WideResNet of the form shown in figure 6.6. The model is trained from scratch for 200 Epochs on all three aforementioned resolutions. Training networks from scratch, especially this type of generative model, takes up a lot of time; typically, the longer it trains the better. However, given that spending exorbitant amounts of time on training a single network is not feasible, there is a trade off: the number of epochs must be set such that training is long enough to ensure proper, but does not take too long and becomes inefficient. The network is designed so that it reflects the hybrid nature of the approach. This is done by creating two separate output layers, a classification output layer that gives conditional probabilities and a layer that gives likelihoods based on the estimated input data distribution.

Approximation and Random Sampling

Applying the methodology presented in Grathwohl et al. (2019), this thesis also uses stochastic gradient langevin dynamics to train the EBM Welling and Teh (2011). Generating samples happens via a series of Markov chains. For every step of the Markov Chain the following update rule is applied. The full algorithm is shown in 6.1.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

$$x_{t+1} = x_t - \frac{\alpha}{2}\nabla E(x_t) + \epsilon, \quad x_0 \sim \mathcal{U}(-1,1) \quad \epsilon \sim \mathcal{N}(0,\alpha) \tag{6.1}$$

where x_0 is sampled uniformly; ∇E is the energy's gradient; ϵ is normally-distributed noise; α is the step size and the standard deviation. In practice the step-size α and the standard deviation of ϵ is often chosen separately because it allows for faster training Grathwohl et al. (2019). The standard deviation is denoted as σ.

This is a method typically seen in Bayesian inference; it has as its posterior distribution the Boltzmann distribution, and is highly adept at efficiently approximating (the gradient of) the partition function as Grathwohl et al. (2019) and Du and Mordatch (2019) show. In every step the energy function's gradient is added to Gaussian noise. The energy function is represented by a CNN which means its gradient can be efficiently calculated using back-propagation.

Loss Function

Given the fact that the goal is to calculate probabilities, the loss function used for network training is again the negative log likelihood loss. This makes it possible to use maximum likelihood estimation, giving the set of parameters that maximises the likelihood under the distribution for given inputs. Furthermore, minimizing NLL equivalently minimizes the distance between the distribution estimated by the model, p_{model} and the true data distribution p_{data}. While it is tempting to assume that, because of their shared name, the loss functions for deep learning classifier in 6.3.4 and the energy based model have the same form, that would be a mistake, this can be seen in detail in 4.5. The negative log likelihood loss for the energy based model is based on the energy function and takes the following form:

$$\mathcal{L}_{nll}(w,\mathcal{S}) = \frac{1}{P}\sum_{i=1}^{P}\left(E_w(I_i,y_i) + \frac{1}{\beta}log\left(\int_{y\in\mathcal{Y}} exp(-\beta E_w(I_i,y))dI\right)\right)$$

The loss is decomposed into two key components, the *positive phase* and the *negative phase* Goodfellow et al. (2016):

- $\mathcal{S}$, $\mathcal{I}$ and $\mathcal{Y}$ defines as in 6.2.
- $E_w(I_i, y_i)$ the energy between the $i-th$ x-ray image and the $i-th$ label (positive phase).
- $log\left(\int_{y\in\mathcal{Y}} exp(-\beta E_w(I_i,y))dI\right)$ the log of the partition function, definition 4.6, also referred to as free energy (negative phase). This term contains the energies between the $i-th$ x-ray image and *all* possible responses/diagnoses (even the wrong ones).
- w are the weights, P is the size of the training data

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Optimiser

Similarly to the optimiser for the state of the art deep learning model, adaptive learning rate algorithms are used to update the weights. These usually generalise better than traditional optimisers like SGD, since they adapt parameters individually and automatically. Furthermore, traditional optimisers can be slow to converge and require manual effort to set the relevant hyperparameters. The EBM also uses the Adam Kingma and Ba (2014) algorithm, with an initial learning rate of 1e−4. Whenever the loss diverged mid training, learning rate and the number of steps the Markov chains take, had to be manually decreased and increased, respectively. The balance of these two hyperparameters was the deciding factor for stable training. The best model with the lowest validation loss of all epochs is saved.

Algorithm

The training algorithm used is based on the works of Grathwohl et al. (2019) and Du and Mordatch (2019):

Algorithm 6.1: EBM Training Algorithm Given network G_w, SGLD step-size α, SGLD std σ, replay buffer B, SGLD steps η, reinitialization frequency ρ

```
while not converged do
    /* Sample x and y from dataset */
    $L_{clf}(w) = loss(G_w(x), y)$
    Sample $\mathbf{x}_0 \sim B$ with probability $1 - \rho$, else $\mathbf{x}_0 \sim \mathcal{U}(-1, 1)$   // Initialize SGLD
    for $t \in [1, \ldots, \eta]$ do
        $\widehat{\mathbf{x}_{t+1}} \leftarrow \widehat{\mathbf{x}_t} + \alpha \cdot + \mathcal{N}(0, \sigma)$   // SGLD
    end
    $L_{gen} = LogSumExp_{y'}(G(\mathbf{x})[y']) - LogSumExp_{y'}(G(\widehat{\mathbf{x}_{t+1}})[y'])$   // Approximation derivative log likelihood
    $L(w) = L_{clf}(w) + L_gen(w)$
    Calculate gradient of $L(w)$ to minimize loss
    Add $\widehat{\mathbf{x}_{t+1}}$ to $B$
end
```

Where G_w is the CNN in question; α and σ are the step size and standard deviation for SGLD, respectively; B is the buffer of past samples; η is the number of steps to run the chain; ρ is the frequencey for resampling.

For the experiments, the number of SGLD steps, i.e. the steps for the Markov chains, are 40, buffer size, the number of states saved, is 10000, SGLD step-size 1 and SGLD noise 0.01. The reinitialization frequency ρ is the probability at which samples are drawn; it is set at 0.05 As an alternative, uniform sampling could also be used.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

6.3.5 Evaluation

The deep learning classifier and the EBM are compared to answer the initial research questions, see 1.3. The focus is on the following three main themes: comparing discriminative power for both modelling approaches and on all resolutions; analysing the out of distribution detection capabilities using three different out of distribution datasets; evaluating the ability to produce calibrated predictions.

The state of the art deep learning classifier and the Energy Based Model were both trained using a WideResNet architecture as a backbone. Both models undergo the same data generation and pre-processing steps. The deep learning classifier was trained using the resampling scheme described in 6.3.4, thus it is trained 10 times to ensure stable results. Given the complexity of the necessary calculations, training the EBM is highly time consuming, making training it more than once infeasible for the scope of this thesis. However, the MCMC methods used are constructed such that they deliver stable results.

Discriminative Power

The discriminative power for both the deep learning and energy based approach are assessed using the AUROC. The average across all four classes is calculated and subsequently compared to the benchmark for state of the art, as published in the original CheXpert paper Irvin et al. (2019). The average AUROC must be at least at 91%, with no individual class being under 90%. These values were chosen to be comparable with state of the art literature Irvin et al. (2019)

For each individual resampling round of the deep learning classifier, the average AUROC for all classes is calculated. After training all 10 models, the average AUROCs are once again averaged over all rounds to deliver the final result. Both a pre-trained and "from scratch" network are trained and compared to the EBM. This shows 1. the effect of pre-training on the classifier and 2. allows an evaluation of which "from scratch" model fares better out of the box and 3. whether the EBM can compete with pre-trained networks. All accuracy scores are delivered with 95% confidence intervals. For the state of the art deep learning classifier, these were calculated as bootstrap intervals by finding the relevant percentiles in the distribution of AUROC values. For the EBM, average accuracy was calculated based on the best 10 model checkpoints saved during training and a quantile-based confidence interval was calculated, again based on state of the art literature Irvin et al. (2019).

Out Of Distribution Detection

Trying to classify an input as out of distribution, meaning it was sampled from a different testing distribution compared to the training distribution, is known as Out of Distribution Detection (OOD). In other words, it is investigated how well the models can separate the data it knows, i.e. was trained on, from the data it does not know, i.e. was generated under a different distribution. The evaluation of OOD is structured according to two

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

main approaches: 1. PCA, a non-modelling approach, applied before training and 2. the state of the art deep learning classifier and EBM models applied after training.

Non-Model Approach

As expressed in 6.2.2, PCA can be seen as a simple, non-parametric tool for OOD that is an alternative to deep learning. By calculating the pixel-wise MSEs that occur when the inverse transform, the reconstruction of the dataset based only on the principal components, is compared to other image datasets, PCA's ability of detecting different types of datasets can be expressed. More specifically, the inverse-transform is used to calculate the pixel-wise MSEs compared to the original CheXpert dataset, and compared to the out of distribution datasets. Intuitively, this explores whether the strongest latent features found by PCA are truly representative of the CheXpert data. The distribution of both sets of errors are plotted and compared. Significant overlap of the distributions would imply that the features found by PCA are not representative of the data.

PCA can be seen as an interesting OOD baseline alongside which the modelling approaches can be viewed. However, while PCA can be used as an effective exploratory tool to get a fast understanding of the data *before modelling*, its main drawback in the context of the aims of this thesis is that cannot be used for downstream discriminative tasks. In addition, the principal components are the strongest features in the data; for image data, often very low-level features, that PCA would miss, can strongly contribute to an image. This is not the case for CNNs, that are very good at identifying these low-level features.

PCA and robust PCA were applied to the chexpert dataset, identifying the top 5 principle components. The data was transformed back into the original data space, based only on the 5 principle components. The original dataset and the dataset with reduced dimensionality were compared by aggregating the pixelwise mean squared error for each corresponding image from each dataset. A high mean squared error between images implies that this image was not adequately explained by the principle components that were identified as the most important, indicating a large difference to typical images in the dataset. This logic was used for an "internal" analysis to find outliers (see 6.2.2), or between datasets to test whether this method is capable of identifying datasets with different underlying features than the original one.

Model approach

One way to measure out of distribution detection once a model is fitted, is to consider OOD as its own binary classification problem applied to the predictions from the models. When training is complete, this thesis follows the work of Hendrycks and Gimpel (2016) to evaluate the OOD; however, the deep learning classifier and the EBM must be treated slightly differently.

For the deep learning classifier, the maximum softmax probability over the network

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

output is taken can be classified as in (= 1) or out (= 0) of distribution, based on a threshold of 0.5 in order to be comparable with existing literature.

The EBM can directly be used to express probabilities, thus the approach in Hendrycks and Gimpel (2016) is slightly modified by not using the softmax as a scoring function on the output. The output is again classified as in (= 1) or out (= 0) of distribution, based on a threshold of 0.5.

The classification results are then summarised by calculating the corresponding Area Under Receiver Operating Characteristic (AUROC) Davis and Goadrich (2006). A high AUROC indicates a successful classification, meaning that the model under investigation can successfully identify which images are sampled from the training distribution and which are not. Conversely, an AUROC of 0.5 indicates that the model under investigation *cannot* successfully identify which images are sampled from the training distribution and which are not.

OOD results are computed for every model, using each of the three OOD datasets described in 6.2 on each resolution. The datasets become progressively dissimliar to the original training data, giving insight into how well the model can discern *within* and *between* domains sensitive to model is at finding foreign inputs.

Calibration Metrics

Calibration was measured by calculating the *Expected Calibration Error* and *Maximum Calibration Error* 5.1.1. The Expected Calibration Error expresses the expected value of a classifier's confidence and accuracy. We can approximate 5.4 by binning the model's predictions into M equally spaced bins and taking a weighted average of the bin's accuracy/confidence difference. Maximum calibration error is not the weighted average but the maximum of the bin's accuracy/confidence difference.

It can be further visualized by using reliability plots. These diagrams plot expected sample accuracy as a function of confidence. They give a clear and transparent view of how confident a classifier is in its predictions. Predictions are grouped into bins based on the prediction confidence value and the accuracy for each bin is calculated and plotted. For perfect calibration, the diagram should plot the identity function. Any deviation from a perfect diagonal represents miscalibration.

Outlier Detection

We qualitatively analysed the scores of the two models to determine which of the frameworks shows more potential for correctly identifying outliers. For the DL classifier, the maximum softmax probability is once again used as an indicator of uncertainty. Ideally, these images would have produced low(-er) confidence predictions. In order to explore whether the EBM was able to identify outliers, we take the small sample

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

of outliers found in 6.2.2 and produced their corresponding scores. A high energy, in absolute terms, indicates low compatibility/probability and higher uncertainty, implying that the image may be an outlier.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

CHAPTER 7

Results And Discussion

This chapter presents the results of the conducted experiments; there will be a thorough analysis of the findings to assess whether the initial research questions, see 1.3, were answered. The results are structured into three main sections. Section 7.1 will report the findings on discriminative power for both modelling approaches and on all resolutions. Section 7.2 will elaborate on out of distribution detection; this section will also include the results on the outlier detection carried out during data exploration using PCA. Section 7.3 will detail the findings on the calibration experiments. The final section will showcase the samples that can be generated from the distribution estimated by the energy based model.

7.1 Discriminative Power

The discriminative power for both the deep learning and energy based approach are assessed using the AUROC on all three image resolution sizes. Table 7.1 summarises all of the discriminative results for all three different image sizes on the complete dataset. For the deep learning classifier, both the results of a pre-trained network and a network trained "from scratch" are reported and compared to the results for the EBM. This allows for a comparison of the frameworks on multiple levels.

Model	**64×64**	**128×128**	**224×224**
CNN (pretrained)	0.797 (0.768,0.813)	**0.923** (0.921,0.926)	**0.938** (0.936,0.941)
CNN (from scratch)	0.777 (0.762,0.781)	0.821 (0.815,0.826)	0.861 (0.858,0.865)
Hybrid EBM	**0.81** (0.76,0.83)	0.85 (0.80,0.88)	0.67* (0.61,0.69)

Table 7.1: Mean AUROC scores, including confidence intervals, across all four pathologies for all models, organized according to the three different resolutions. Both the results of a pre-trained network and a network trained "from scratch" are reported and compared to the EBM. The "winning" model is indicated in bold. *interim results gathered before full training run completed due to the fact that the full run would take around 3 months which was beyond the scope of this thesis

AUROC The initial experiments showed that the pre-trained DL model had the best discriminative performance on the two larger resolutions, and reaches the benchmark derived from the CheXpert results on the 224 × 224 resolution. The EBM achieved the highest AUROC on the smallest resolution and consistently improves as the resolutions rise. While at 90% it did not reach the CheXpert benchmark, it managed to achieve high, though not competitive, results compared to the state of the art. However, it is also important to remember that the EBM is not pre-trained; the EBM can clearly outperform the deep learning classifier that was trained "from scratch". This comparison directly shows us the positive effects of the EBM training scheme and how well the WideResNet architecture is utilised. The MCMC/Bayesian learning approach employed in the EBM captures the statistical, latent dependencies in the dataset vs the deep learning classifier trained via SGD where the network memorises the patterns in the images. The better EBM's approach leverages its knowledge of the data overall to deliver more accurate predictions.

The experiments on the full dataset, results shown in table 7.1, showed a similar picture. The pre-trained DL model had the best discriminative performance, indicating that it scales well to larger dataset sizes. At $\sim 94\%$ it reaches the benchmark derived from the CheXpert results on the 224 × 224 resolution. The DL classifier trained from scratch also showed a similar results compared to the initial experiments. It scales well in terms of training time, but at 86% it falls far below the CheXpert benchmark. The hybrid EBM on the other hand does not scale well. With a projected training duration of at least 3 months, not including time spent for manual adjustment of hyperparameters, the training time on the 224 × 224 resolution grew exponentially with the additional data. Since this goes beyond the scope of this thesis, interim results were generated and reported. In order to gain a more complete understanding of the training process and how final results could look, the trends in loss and AUROC between the initial run and the full run were compared. This comparison is shown in figureXXX. It can be seen that, while volatile, the behaviour for both training situations is very similar for both metrics; though it is a slow iterative process, loss continuously declines while AUROC steadily rises in both cases. It is therefore not unreasonable to assume that, given more time, the hybrid EBM on the full dataset would produce the same high quality discriminative results as it does for the smaller subset.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

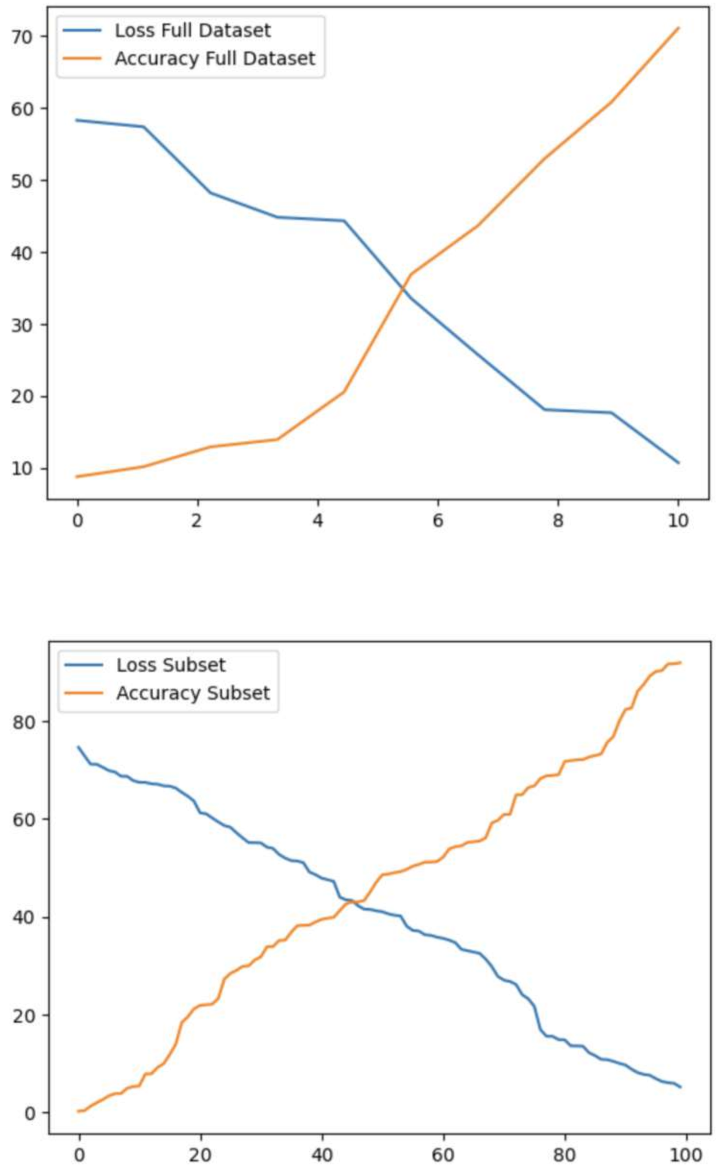
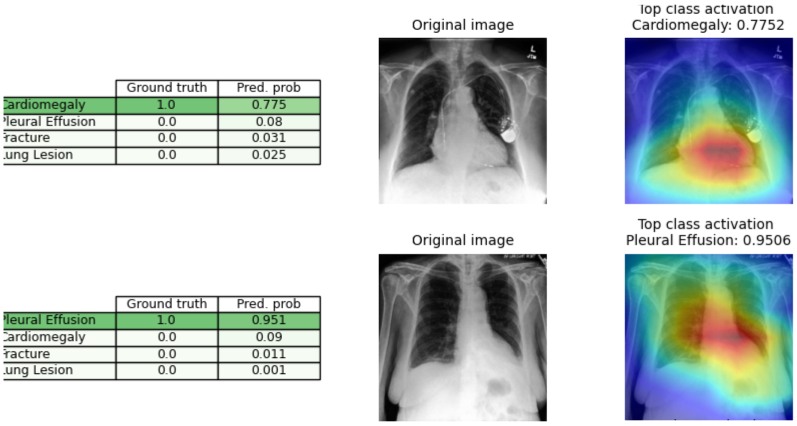


Figure 7.1: Comparison of loss and accuracy behaviour for the EBM trained on the full dataset (above) and the small subset (below). It can be seen that the behaviour of the two modelling runs are very similar. Thus, it is a reasonable assumption that, given more time, the model trained on the full dataset will achieve results similar to the model trained on the subset

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The AUROC estimate is surrounded by a very slim confidence interval of $(0.936, 0.941)$ for the pre-trained deep learning classifier, indicating high certainty. A similar picture is seen with the DL model trained from scratch, though it performs worse. The AUROC estimate for the EBM is surrounded by a slightly wider confidence interval of $(0.61, 0.69)$, indicating a slightly larger uncertainty attached to the estimate. This is in line with the expectation that the EBM has a more realistic sense of uncertainty compared to the deep learning classifier. The state of the art classifier clearly benefits from the pre-training on ImageNet. It uses its pre-existing knowledge of features to learn about the data very fast; it achieves the highest scores in the larger resolutions and clearly outperforms the EBM. However, the performance of the "from scratch" classifier shows that, when the pre-training is taken away, the EBM is has a higher auroc, showing that leveraging the generative capabilities of the EBM has a positive effect on classification. The MCMC methods running in the background work to deliver estimates that capture the latent connections which, in turn, clearly benefits the downstream classification task. The best performing discriminative model on the 224×224 resolution, the pretrained DL model, was used to create a visualization called GradCam Selvaraju et al. (2019), see 7.2. This is a type of 'visual explanation' to make the results more transparent; it is a matrix, also called a localization map, which highlights important regions in the image were particularly important for the prediction. It works by taking the un-normalised network output for an arbitrary class and differentiating it by backpropagation with respect to the feature maps produced by the final convolutional layer. The gradient matrices, one matrix for each feature map, are averaged to produce a weight for each feature map. Finally, the GradCam matrix/localization map is calculated as the weighted sum of the feature maps, using the average gradients as weights. The visualization can be seen in figure 7.2

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



	Ground truth	Pred. prob
ardiomegaly	1.0	0.775
leural Effusion	0.0	0.08
racture	0.0	0.031
ung Lesion	0.0	0.025

	Ground truth	Pred. prob
leural Effusion	1.0	0.951
ardiomegaly	0.0	0.09
racture	0.0	0.011
ung Lesion	0.0	0.001

Figure 7.2: Selection of GradCam images on a sample of x-rays using the DL model. The ground truth label, the predicted probability are shown in the left column. The middle column shows the original image and the right column shows the localization map overlayed onto the image, indicating which areas played an important role in the classification. The image on the top shows an example of a patient with Cardiomegaly. The image on the bottom shows a patient with pleural effusion. The model correctly classifies the images

GradCam can give an indication into the generalisability of the model and whether it is biased Selvaraju et al. (2019). If the localization map highlights a discriminative region that does not make intuitive sense, e.g. the radiograph contains a patient with heart problems but lights up in the background, we know that there must be a certain characteristic within the data that the model is learning and reproducing. The provenance of this phenomenon should then be searched for in the data. Figure 7.2 shows that the pre-trained deep learning classifier uses the intuitively correct areas of the image to make its prediction for these classes.

A Remark on Resolution The results show that the models trained on larger images deliver higher AUROCs. All models are the most accurate on the highest image resolution, 224 × 224. Table 7.2 shows how much accuracy is lost, in percent, as the image size is progressively reduced from 224 × 224 down to 64 × 64. This can be explained when considering that smaller images make the smaller details in the images harder for the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

network to discern, especially when the presence of a condition could depend only on individual pixels.

Model	**64×64**	**128×128**
CNN (pretrained)	**15%**	2%
CNN (from scratch)	9%	4%
Hybrid EBM	9%	**5%**

Table 7.2: Percent loss of accuracy as the models are trained on progressively lower image resolutions. The largest reduction for each image size is indicated in bold. All models have the highest accuracy on the highest image resolution, 224 × 224.

7.2 Out Of Distribution Detection

The evaluation of OOD, detailed in section 5.1.1, is structured according to two main approaches: 1. using PCA, a non-modelling approach, before training and 2. using the deep learning classifier and EBM models after training.

7.2.1 PCA

For our analysis we selected the top 9 most outlying images to give an indication of what outliers there could be. The calculations of the principle components can be very sensitive to strong outliers, thus robust PCA is also employed.

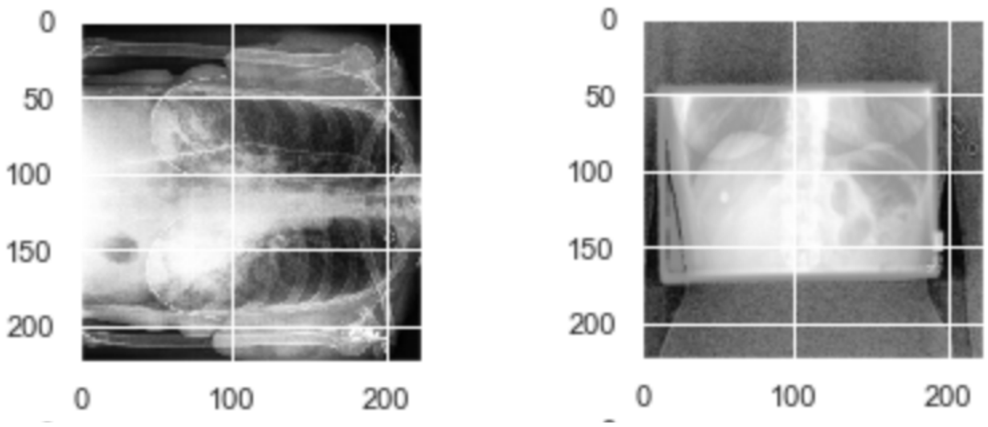


Figure 7.3: The top outliers in the CheXpert data found using PCA and robust PCA, respectively. The outliers are often images where the patient is not placed correctly or the image is cut off or blurred (a) Outlier found using PCA (b) The outliers found using robust PCA

It can be seen in figure 7.3, that the outliers are often images where the patient is not placed correctly or the image is cut off or blurred or rotated. The MSE can also be

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

used as a metric to compare the datasets. Using the inverse-transform, the pixel-wise MSE is calculated and compared to 1. the original CheXpert dataset, and 2. the out of distribution datasets to assess whether the strongest latent features found by PCA are truly representative of the CheXpert data. The inverse-transform of robust PCA is used and the distribution of both sets of errors are plotted and compared in 7.4. Significant overlap of the distributions would imply that the features found by PCA are not representative of the data.

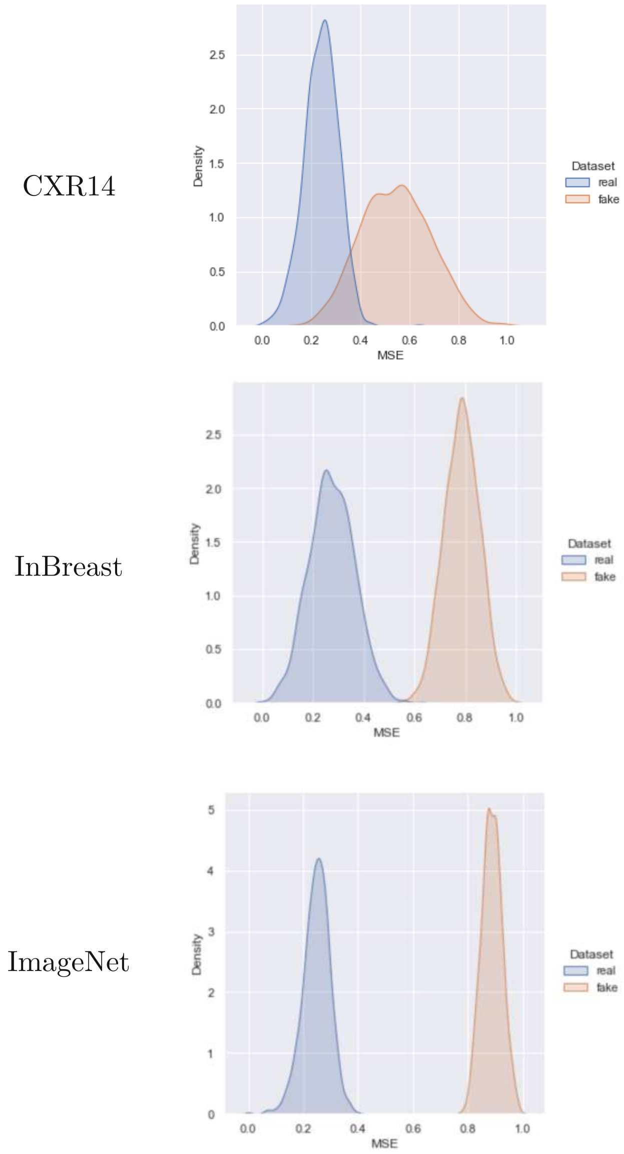


Figure 7.4: Distributions showing PCA's capability to find a specific representation of the CheXpert data. The x-axis represents the MSE

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The distributions show that there is some overlap between the MSE distributions when comparing the CheXpert with CXR14 which shows that PCA has some difficulty finding features that uniquely represent CheXpert when compared to a dataset that is very similar. However, as the datasets become less similar to CheXpert the distributions move further apart, indicating that the CheXpert features found by PCA are representative enough to recognise completely different domains. However, the representation is not sensitive enough to distinguish within a domain.

7.2.2 Modelling Approaches

One way to measure out of distribution detection once a model is fitted, is to consider OOD as its own binary classification problem applied to the predictions from the models, as detailed in 6.3.5. The deep learning classifier and the EBM must be treated slightly differently according to their scoring functions. The deep learning classifier softmax applied to its output while the EBM natively calculates probabilities. OOD scores are calculated for each model, at every resolution for each of the out of distribution datasets. Training was once again carried out in a two step approach, first on a small subset and then on the full dataset.

Model	**Dataset**	**64x64**	**128x128**	**224x224**
CNN (pretrained)	CXR14	0.5	0.61	0.63
	Inbreast	0.56	0.68	0.69
	Imagenet	0.61	0.72	0.74
CNN (from scratch)	CXR14	0.5	0.56	0.58
	Inbreast	0.53	0.56	0.58
	Imagenet	0.56	0.59	0.63
Hybrid EBM	CXR14	**0.94**	**0.96**	**0.98**
	Inbreast	**0.97**	**0.98**	**1.0**
	Imagenet	**1.0**	**1.0**	**1.0**

Table 7.3: AUROC scores for the Out of Distribution Detection classification. Each model is exposed to all three OOD datasets, at every resolution. A high AUROC expresses a successful classification of in and out of distribution elements; an AUROC of 0.5 indicates that the model was not able to detect out of distribution elements

Table 7.3 summarises the AUROC scores for the Out of Distribution Detection classification on the full dataset; a higher value indicates that the model is better at separating the individual datasets. The standard DL classifiers are both far inferior at OOD compared to the performance of the EBM, even though the training was cut short (explanation see aove under discriminative power). The scores show that distinguishing between the CheXpert data and another, separate chest x-ray dataset, i.e. within the domain, is not successful. As the out of distribution datasets become more and more different to the original dataset, the out of distribution performance slightly improves. In contrast, the EBM outperforms the other implementations in all datasets and on all resolutions. The

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

EBM can easily distinguish between the two chest x-ray datasets and can perfectly tell the other datasets apart. To further showcase this, figure 7.5 plots the score distributions of the two models.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

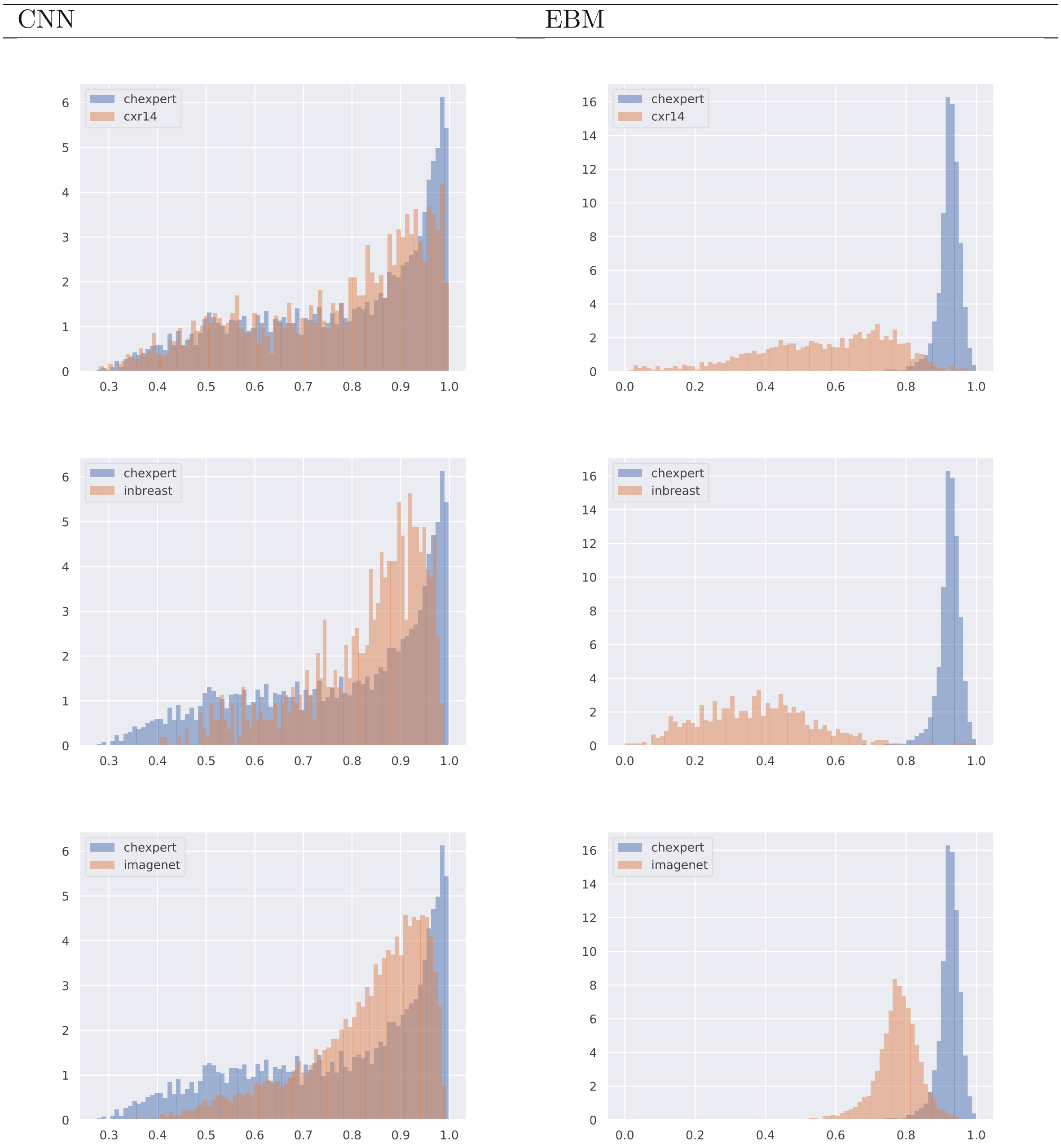


Figure 7.5: Visualisation of score distributions from the pre-trained DL model and the EBM. The first row shows the distributions of softmax (DL model) and energy-based (EBM model) probabilities between the CheXpert dataset and the CXR14 dataset; the second row the the distributions between the CheXpert dataset and the InBreast dataset and the third the distributions between the CheXpert dataset and the ImageNet dataset. The bottom row shows the distribution probabilities from the energy model between the CheXpert dataset and the out of distribution datasets. The degree of overlap is indicative of how well the respective model can find fake data. The x-axis is the score, softmax or energy, and the y-axis in the frequency

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Figure 7.5 plots the respective scores of the pre-trained DL model and the EBM models in order to compare how well each model can find the out of distribution data. The top row shows the distribution of softmax probabilities; it can be seen that there is significant overlap for all datasets, with high probabilities are assigned to images even though they are completely different to the CheXpert data.

In other words, this is interpreted as the model is saying that there is a high probability of, for instance, a dog being an element of the CheXpert dataset, a serious error. It may be considered correct that the DL model produces similar scores for the CXR14 data, the first image, in fact the distributions are almost completely the same. However, the DL model was only trained on four pathologies, while the CXR14 data contains 14. This strongly emphasises the importance of accurate OOD, especially in the medical field. Simply put: the deep learning classifier operates outside of its validated configuration, meaning that if a chest x-ray showing atelectasis is given to the network, it might confidently predict pleural effusion, a potentially fatal mistake.

The bottom row of figure 7.5 shows the distribution probabilities from the energy model trained on the full dataset. It can clearly be seen in all three images that the mass related to the CheXpert is located to the far right, meaning that high probabilities are assigned to these images. On the other hand, the out of distribution images are all assigned lower probabilites. Simply put, this means that the representation them model learns can effectively tell what is part of its distribution, the images assigned high probability, and what is foreign, the images assigned low probability. The model is even sensitive enough to filter out the CXR14 data.

7.3 Calibration

Producing calibrated probabilities is an essential requirement for discriminative tasks. Well calibrated predictions provide a valuable extra bit of information to establish trustworthiness with the user – especially for neural networks, whose classification decisions are often difficult to interpret Guo et al. (2017). Section 6.3.5 details how calibration was measured by calculating the **Expected Calibration Error** and **Maximum Calibration Error**, and subsequently visualized using reliability plots. Calibration was tested on various portions the dataset, to check whether there was any dependency on dataset size. Reliability plots, plot expected sample accuracy as a function of confidence. If the model is perfectly calibrated, the identity function should be plotted Guo et al. (2017).

Figure 7.6 visually summarises the calibration for the predictions of the different modelling set ups split by dataset sizes. The x-axis is the binned predictive confidence and the y-axis the corresponding accuracy. Ideally, highly confident predictions should also be highly accurate. The blue bars indicate the accuracy of the specific confidence bin, the red bars indicate the gap between the achieved accuracy and the ideal calibrated value. In general, all models provide miscalibrated, overconfident predictions. This was expected for the DL model, where the accuracy bars are at a consistently low level, or even slope downwards, indicating that high confidence predictions are not accurate. This confirms the hypothesis about the DL model behaviour.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

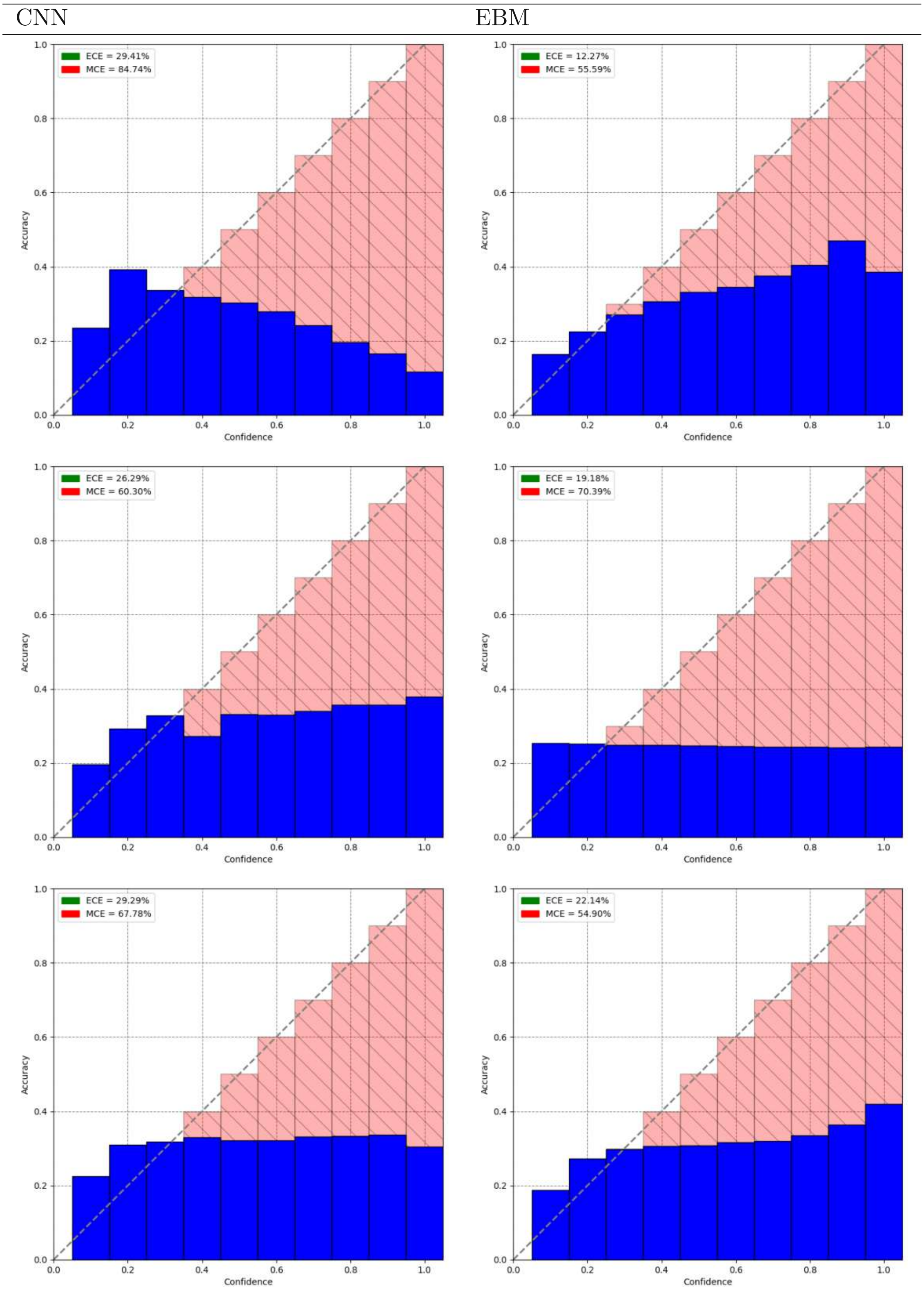


Figure 7.6: Reliability Plots on different dataset sizes for the pre-trained CNN and the EBM. The first row indicates the complete dataset, the second 60% of the complete dataset and the third 30%. The x-axis is the binned predictive confidence and the y-axis the corresponding accuracy. Ideally, highly confident predictions should also be highly accurate. The blue bars indicate the accuracy of the specific confidence bin, the red bars indicate the gap between the achieved accuracy and the ideal calibrated value.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

However, the performance of the EBM was unfortunately not as expected. Based on the literature, specifically Grathwohl et al. (2019), it was expected that the EBM would deliver well-calibrated predictions out of the box. There is a noticeable upwards trend, i.e. the bins are becoming more accurate with increasing confidence, however it is still far from ideal. The reason for this could be because EBMs need a lot of time to train to estimate a distribution that also delivers calibrated predictions, especially considering all the operations and approximations that need to be optimised in the course of training. The corresponding ECE and MCE are reported in the table 7.4.

Percentage of complete dataset	Model	ECE	MCE
100	CNN	29.41	84.74
	EBM	12.27	55.59
60	CNN	26.29	60.30
	EBM	19.18	70.39
30	CNN	29.29	67.78
	EBM	22.14	54.90

Table 7.4: Comparison of ECE and MCE between CNN and EBM

Figure 7.7 shows the development of accuracy and ECE during training. The average AUROC consistently increases as training progresses, while the calibration error consistently decreases. It stands to reason that with longer training times, the distribution estimate would further improve and deliver better calibration results.

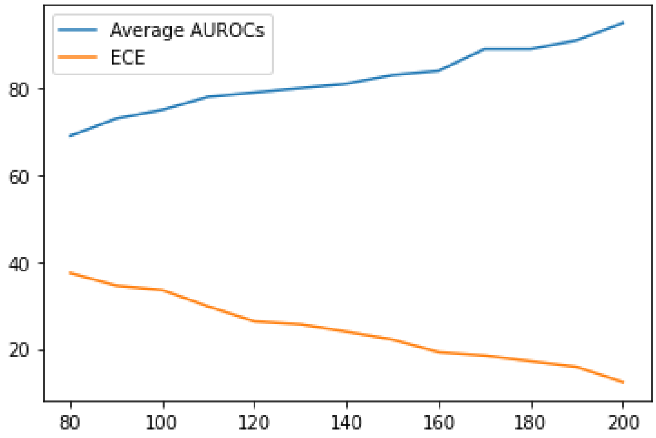


Figure 7.7: Diagram showing the trajectory of accuracy and ECE during training. There is a clear downward trend of the calibration error (orange) and a clear upward trend of the AUROC (blue). The x-axis shows the number of epochs.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

7.4 Outlier Detection

The analyses so far have largely been focused on comparing different datasets. Outlier detection is a task that is purely focused on comparisons within the dataset. In section 6.2.2, PCA and robust PCA was used to identify the most outlying points. The top 9 images found with robust PCA have been selected to observe whether the model-based approaches would classify these images as outliers as well.

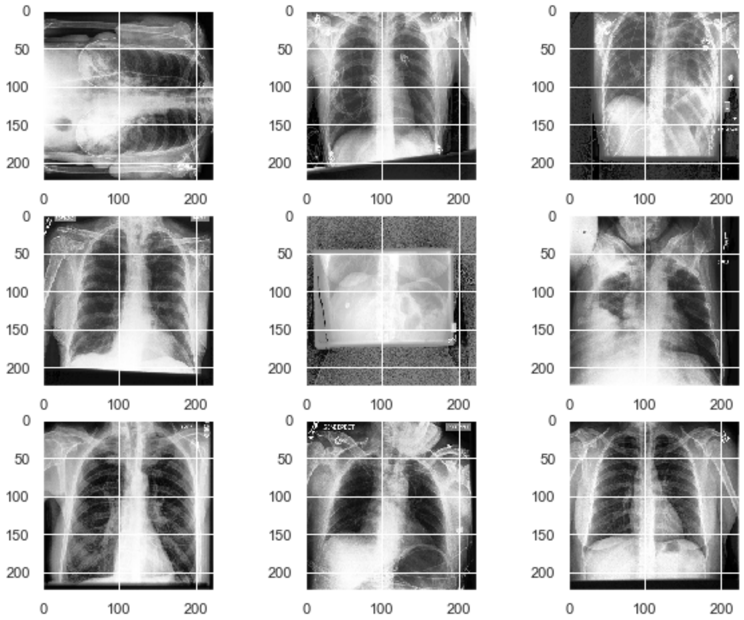


Figure 7.8: Top 9 most outlying images in CheXpert as found by robust PCA

The images in figure 7.8 were fed into the pre-trained DL model and EBM models in order to observe what kind of scores they would produce for these outlying images. We qualitatively analysed the respective outputs to determine which of the frameworks shows more potential for correctly identifying outliers. For the DL model, the maximum softmax probability, see section 5.1.1, is once again used as an indicator of uncertainty. The DL model produced very high scores. This indicates that the pre-trained DL model could not be used to filter out these outliers. Ideally, these images would have produced low(-er) confidence predictions. For the EBM, outliers were determined by looking at the probability output for the individual images. A high energy, in absolute terms, indicates low compatibility/probability and higher uncertainty, as elaborated in 2.3, which would imply that image be an outlier. The above 9 images all showed high energies, or equivalently low probabilities, implying that the EBM found something significant, but negative in these images. Using this quantification of uncertainty, the model could be used to identify outliers.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

7.5 Sample Generation

The EBM is a generative model, enabling us to estimate a distribution over the data. While generating random samples plays a pivotal role in the training of the model, see section 4.6.3, the resulting probability distribution can also be used to generate completely new random samples/synthetic data. To do this SGLD is again employed to generate the samples; Markov chains where the next step step is determined by the update rule 6.1 are run to produce a set of images that are based on the features the model learned during training. More specifically, this means that the model can be used to generate images chest radiographs, based on the characteristics it has learned from the training images.

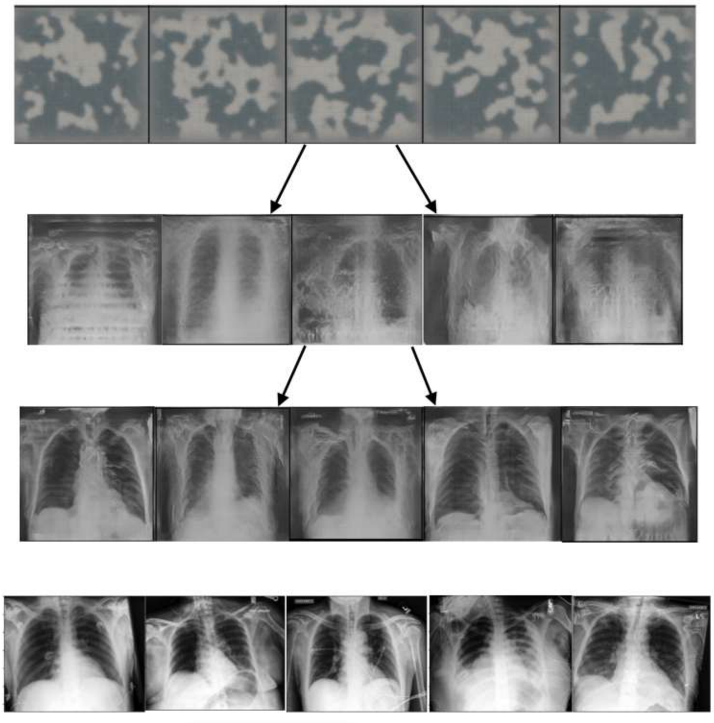


Figure 7.9: Random samples generated in the course of training the EBM. During training, samples were periodically generated from the data distribution at that point in time. The top row shows samples at the beginning of training and resemble random noise. The second row shows samples midway though the training; outlines of torsos can be discerned but the images are still very blurred, indicating that the distribution does not represent the data well yet. The third row shows the samples generated with the trained model and the final row shows real CheXpert samples; model has clearly learned the most important characteristics of chest radiographs and can reproduce them to generate compelling samples.

Figure 7.9 shows a series of samples generated throughout the training process. The

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

samples were taken at the start of training, in the middle and at the end and show how well the data distribution represents the data at that time. Finally, the samples are compared to real chest radiographs to visually assess their quality. The top row shows samples at the beginning of training and resemble random noise. The second row shows samples midway though the training; outlines of torsos can be discerned but the images are still very blurred, indicating that the distribution does not represent the data well yet. The third row shows the samples generated with the trained model and the final row is a samples of real chest radiograophs; the model has clearly learned the most important characteristics of chest radiographs and can reproduce them to generate compelling samples.

7.6 Discussion

The collected results have been analysed to assess whether the initial research questions have been successfully answered. The goal of this thesis is to study the suitability of a hybrid energy based model, as described in Grathwohl et al. (2019), for medical image pathology classification. The following questions were addressed:

1. Does training the classifier in a hybrid EBM schema reach comparable results to a standard discriminative training setup
2. Can the intrinsic generative model of the hybrid setup be utilised for OOD detection
3. How does EBM training affect the model discriminative calibration

Discriminiative Power The pre-trained deep learning classifier was the best performing model. The EBM was successfully used for a downstream discriminative task and performed with high accuracies; however, it could not compete with the pre-trained deep learning classifier. Compared to the deep learning classifier trained from scratch, the EBM was able to achieve far greater discriminative results, showing the benefits of exploiting the distributional properties of the EBM.

Out of Distribution Detection In a pre-modelling exploration, PCA and robust PCA were used to derive representations from the CheXpert data and compare their quality to different types of datasets, as type of indicative OOD method. The representations learned by PCA were effective when comparing the CheXpert data to different domains, such as ImageNet and INbreast. However, within a domain the features were not sophisticated enough to properly distinguish between CXR14 and CheXpert. Overall, considering the relative simplicity of the method, it performed well and makes for an interesting alternative next to the highly complex models, though it cannot be considered as a substitute for a diagnosis system. It has no discriminative abilities and the representations it learns are not sensitive enough to be considered for a high-risk domain such as medicine.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

When applying the trained models for OOD, the deep learning classifier was not able to distinguish between datasets in a useful way, assigning high probabilities to inputs that were not part of the original training data distribution. Specifically, when looking at the way the deep learning classifier compares the two chest x-ray datasets, the importance of accurate OOD, how classical deep learning classifiers fall short, and how significant it is in the medical setting, is emphasised. The EBM on the other hand, excelled at this task. The model clearly learned the latent dependencies very well and was able to leverage this to identify different inputs even within a domain. Its capabilities make it very useful for applications in sensitive areas and it hows that it is clearly superior to the deep learning classifier.

Model Calibration The EBM was not able to deliver, as expected, perfectly calibrated predictions. However, the predictions were better calibrated than the ones produced by the deep learning classifier. It is important to note that the EBM was trained "from scratch", with a highly complex training algorithm working in the background. Thus, possibly more training time would have improved the calibration. The predictions made by the pre-trained deep learning classifier delivered moderately calibrated, but if the pre-training was taken away, the model was grossly miscalibrated.

Additional Analyses Finally, the models were explored for their outlier detection abilities. A set of 9 outliers from the CheXpert data was selected and fed to the models in order to test whether their output would indicate that they were outliers. There was discernible change in the deep learning classifier's output that would indicate an outlier. In a larger sense, this indicated a failure to quantify uncertainty for a given image. The EBM's output showed strong signs of uncertainty for these images, implying that it realised these images something significant, but inadequate, was contained in these images.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 8

Conclusions

This chapter presents a thorough discussion of the results, contextualising them with the initial research questions, and drawing final conclusions to assess what was gained and what was lost when applying the energy based framework. Limitations of the experiments and of energy based modelling as a whole are also discussed, as well as a description of possible clinical usage.

This thesis investigated if the concept of an Energy Based Model is suitable for medical image diagnosis on posteroanterior chest radiographs. The work in this thesis is based in a large part on EBM research conducted by Will Grathwohl and Yann LeCun Grathwohl et al. (2019) LeCun et al. (2006). The conducted experiments address some major challenges for state of the art deep learning classifiers. While they have a proven track record of being highly accurate, both on medical data as well as in other domains, there are drawbacks that limit their large scale real world application. Some of these drawbacks include not delivering out of the box calibrated predictions, and not having the ability to identify when they are given an image that is significantly different to the training distribution, also known as Out of Distribution Detection. The root of this problem is that the training scheme of state of the art deep learning classifiers does not equip them to accurately quantify uncertainty. In other words, their output scores cannot be used to make statements/assumptions about how well an observation fits into the data overall. This is because deep learning classifiers are very good at learning the patterns that minimize a loss criterion to distinguish between chest pathologies, thus appearing to have some powerful, deeper understanding. But it does not have an understanding of what it really means for a patient to have this pathology in relation to the rest of the cases. It does not know what is does not know Gal (2016).

The work in this thesis is laid out as a comparison between the state of the art deep learning classifier and energy based modelling, with a special focus on the respective discriminative, out of distribution and calibration capabilities. A chosen Convolutional Neural Network is trained, establishing a benchmark to evaluate the performance. The connection between the EBM and the deep learning classifier, is that both methods are based on the same CNN architecture, but use different methods of training to fit a model. Both modelling approaches are trained on chest x-ray data that contains four different pathologies: cardiomegaly, pleural effusion, lung lesions and fractures. For the investigation into the fulfilment of the research goals, a selection of metrics and visualisations are used.

Discriminative power, summarised in table 7.1, is evaluated by comparing average AUROC values on three different resolutions. The EBM is pitted against both a pre-trained CNN and a CNN trained "from scratch". he experiments were conducted in a two stage process: an initial proof of concept on a small subset of the data was conducted to pre-determine whether the hybrid EBM would deliver useful results on this type of image data and to observe how difficult the training would be, given that the existing literature often reported unstable training. On the small subset of data, the EBM was successfully used for a downstream discriminative task and performed with high accuracies. However, it could not compete with the pre-trained CNN which dominated the experiments. When the pre-training is removed, the EBM was able to outperform the CNN trained "from scratch", showing the benefits of exploiting the distributional properties for discriminative tasks. On the complete dataset, both deep learning approaches scaled well and replicated their discriminative performances compared to the proof of concept on the small subset. The EBM on the other hand did not scale well. The additional data prolonged the duration of training to such an extent ($\sim$ 3 months for a complete training run, not considering a parameter search and divergence of the loss function) that final results could not be calculated in a feasible amount of time. In order to assess discriminative performance of EBM on the full dataset, loss and accuracy behaviour for the full and initial runs were compared. The comparison showed similar trends in both metrics for the two respective runs, which allows for the reasonable extrapolation that, given enough time, the model on the full dataset would generate similarly good discriminative results as on the smaller subset.

Out of Distribution Detection, summarised in table 7.3, was investigated in both a pre- and post modelling approach by using different datasets to test how well the approaches would be able to distinguish between in and out of distribution data. The out of distribution datasets were deliberately picked to progressively differ from the training dataset, thus showing how well the approaches were able to differentiate between domains and within a domain. The datasets under consideration were: a different chest x-ray dataset ChestXray14 Wang et al. (2017), a breast mammogram dataset INbreast Moreira et al. (2012) and a random subsample from the ImageNet dataset Russakovsky et al. (2015). In a pre-modelling exploration, robust PCA was used to derive representations from the CheXpert data and compare their quality to different types of datasets, as type of indicative OOD baseline method. The features learned by PCA were effective when

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

comparing the CheXpert data to different domains, such as ImageNet and INbreast. However, within a domain the features were not sophisticated enough to properly distinguish between CXR14 and CheXpert. Overall, considering the relative simplicity of the method, it performed well and makes for an interesting alternative next to the highly complex models, though it cannot be considered as a substitute for a diagnosis system. The post-modelling exploration was again carried out in a two step approach, i.e. the out of distribution datasets were first applied to the models were trained on a small subset of data as a proof of concept, and then to the model trained on the full dataset. In both cases the EBM far outmatched the DL classifier. The latter was not able to distinguish between datasets in a useful way, assigning high probabilities to inputs that were not part of the original training data distribution. The importance of accurate OOD is emphasised when looking at the way the DL models compare the two chest x-ray datasets. The model assigns high probabilities to chest x-rays with completely different pathologies than present in the training data, a potentially fatal flaw. The EBM on the other hand, excelled in both cases. It clearly learned the latent dependencies very well and was able to leverage this to identify different inputs between domains and even within a domain. Its capabilities make it very useful for applications in sensitive areas and it hows that it is clearly superior to the CNN. Interestingly, the EBM on the full dataset was already able to achieve excellent OOD results, even though its training was cut short. This can be interpreted in the following way: the distribution is estimated very quickly in the first epochs with already with good quality. Figuratively speaking, this means the general shape of a fitting energy function is determined quite fast. The remaining training time goes into fine tuning this shape. This would also explain why the improvement of the discriminative performance is a long and iterative process, that becomes very evident with such a large dataset: the model uses the information learned from its generative side to then develop and improve its discriminative side, i.e. the discriminative will always lag behind the generative and only when the generative is good, will the discriminative start to improve. This also gives rise the phrase "downstream discriminative capabilities/tasks".

Both modelling approaches did not deliver well-calibrated predictions. While the EBM was observed to perform better than the CNN approach, it was still miscalibrated. Since the EBM was trained from scratch, it stands to reason that calibration would improve given longer training time. The deep learning classifier, though also not delivering well-calibrated predictions, behaved as expected: they deliver highly confident predictions that are not in line with ground-truth likelihoods observed.

The main contribution of this work is the investigation into whether energy based modelling, previously untrained on medical data, can competitively deliver accurate predictions while also supplying 'context', i.e. the user can get a sense of the uncertainty attached to a prediction. Simply put, the EBM output has a deeper meaning in relation to the data, something the deep learning classifier does not have. This is the result of the main benefit of these models: they estimate entire distributions and thus naturally output probabilities, whereas deep learning classifiers give point estimates for parameters that minimize a specific loss. The EBM was shown to use this to distinguish between nonsensical inputs, a huge benefit for the medical domain where wrongly diagnosed outputs can have fatal consequences. In a tightly controlled domain, for a very specfic case, the deep learning classifier may seem superior because it can achieve high accuracies. However in reality, diagnosing medical images is a complex task with many contributing factors and different points of view. The EBM sacrifices some discriminative power compared to the pre-trained deep learning classifier, but this is more than made up when seeing that it can model more areas of the diagnostic process.

Overall, the experiments have shown that, energy based modelling can be employed in medical image classification and improve some of the drawbacks attached to classical deep learning. In contrast to classical deep learning classifiers, they know what they do not know and they push the topic of explainability and interpretability of artificial intelligence forward. While the benefits of generating a distribution on medical images and being able to use it for a downstream discriminative tasks are significant, notable limitations could also be observed. Estimating a distribution is always hard Goodfellow et al. (2016); training the EBM often exhibited instability, requiring a lot of manual work to find appropriate settings to fit the model, especially considering the approximative tools needed to express the distribution. The way to efficiently train an EBM is an active field of research. Many of the published training methods, including work done by Grathwohl et al. (2019), rely on approximative tools like Monte Carlo Markov Chain. While these methods generally work, the calculations done "under the hood" are highly complex and are the reason for the instability. Moreover, the process of training was highly time consuming and inefficient, meaning that EBMs do not scale well. This was especially obvious when the EBM was trained on the full dataset and training time increased exponentially. In addition to the computational complexity, a key factor in training EBMs are the hardware requirements. Training the model was only possible on GPU with sufficient memory, especially using larger image sizes such as 224 × 224 or 128 × 128. Given a lot more training time, it can be hypothesized that the quality of samples generated from the distributions, as well as the discriminative power and calibration would improve. On the other hand, training the classical deep learning classifiers was comparatively easy, especially considering the benefit of pre-trained weights. With minimal manual effort, the models were trained in a short amount of time and were very stable, though they also required the GPU hardware.

Given the benefits observed in the experiments, EBMs do show signs of potential to be integrated into a real-world clinical application. For instance, an automatised diagnostic system based on an EBM could not only give high quality predictions, but its deeper understanding of uncertainty could be used to give an indication of faulty images, or if there are specific features present in the image that are highly highly pronounced compared to the rest of the population. This indication, more specifically a likelihood-based metric, could be combined with an uncertainty that, when breached automatically raises a flag signifying the need for human intervention. This could significantly support radiologists in analysing chest x-rays by reducing their workload, thus allowing them to

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

be more attentive to more serious cases. In addition, it gives the users the power to be as strict as they want by adjusting the uncertainty threshold. Furthermore, this tyoe of framework can be used to create a reliable ranking of patients relative to the severity of their illness, with high uncertainty indicating urgent cases and low uncertainty indicating non-urgent cases. These implementations would contribute to the safety of using an automatic system, promoting wider clinical use.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Bibliothek, Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Figures

3.1 The perceptron model visualized 19
3.2 The artificial neuron including a continuous, nonlinear activation function f. Inputs x_i are weighted by weights w_i and added together before being fed into the activation function. 20
3.3 The multilayer perceptron model 24
3.4 A sigmoidal function according to definition 3.3 Nielsen (2015) Petersen (2020) 24
3.5 visualisation of a local receptive field Nielsen (2015). The small window of input pixels is weighted and aggregated and passed to a single neuron in the second layer. 41
3.6 (a) Graphical representation of creating a feature map. (b) Application of a filter Goodfellow et al. (2016) 42
3.7 Illustration of a convolutional neural network Berner et al. (2021) 43
3.8 The architecture of the ResNet with the residual blocks bordered by the short cuts (loops). In practice there are different types of residual blocks depending on the dimensions of the convolutional layers within. Niu et al. (2019) 45

4.1 Comparison of the energy functions, i.e. the underlying neural networks, for the Boltzmann and restricted Boltzmann machines. The original Boltzmann machine (left) has every neuron within a layer and between layers connected to every other neuron. The restricted Boltzmann machine has no intra-layer connections. Every neuron in a layer is only connected to every other neuron in the subsequent layer. O'Connor et al. (2013) 50
4.2 Visualisation of how the energy function can be applied to a x-ray image for classification. The function's output can be understood as a measure of compatibility between the x-ray image and each of the possible diseases. A scalar energy for each of the combinations is calculated, with the lowest resulting energy representing the best combination. The disease that determined that the lowest energy, becomes the predicted class. 51

5.1 An example visualization of reliability curves. An alignment with the 45 degree line indicates perfect calibration. Under the the 45 degree line indicates over-confidence of the classifier. Over the 45 degree line indicates under-confidence of the classifier. (a) shows a very well-calibrated classifier and (b) a very poorly, over-confident classifier 70

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

5.2 Visualisation of OOD: the goal is to classify an input as out of distribution, meaning it was sampled from a different testing distribution compared to the training distribution. If a model is trained on dog breeds, can it successfully distinguish between the blueberry muffing and the chihuahua? 72

6.1 samples of the out of distribution datasets; top is the ChestXray14 data; middle is INbreast data; bottom is ImageNet data 83

6.2 Histogram of ages in CheXpert 85

6.3 Grid of images showing sample images of each illness 87

6.4 A grid of the top 9 outlying images in the chexpert dataset as determined by PCA 89

6.5 A grid of the top 9 outlying images in the chexpert dataset as determined by robust PCA 90

6.6 The hybrid model visualised schematically. The wide basic blocks each follow the structure shown in the detail diagram on the bottom left. The factor k defines how wide the network is. What sets this architecture apart is the hybrid output; the network has both an energy and classification output layer 94

7.1 Comparison of loss and accuracy behaviour for the EBM trained on the full dataset (above) and the small subset (below). It can be seen that the behaviour of the two modelling runs are very similar. Thus, it is a reasonable assumption that, given more time, the model trained on the full dataset will achieve results similar to the model trained on the subset 107

7.2 Selection of GradCam images on a sample of x-rays using the DL model. The ground truth label, the predicted probability are shown in the left column. The middle column shows the original image and the right column shows the localization map overlayed onto the image, indicating which areas played an important role in the classification. The image on the top shows an example of a patient with Cardiomegaly. The image on the bottom shows a patient with pleural effusion. The model correctly classifies the images 109

7.3 The top outliers in the CheXpert data found using PCA and robust PCA, respectively. The outliers are often images where the patient is not placed correctly or the image is cut off or blurred (a) Outlier found using PCA (b) The outliers found using robust PCA 110

7.4 Distributions showing PCA's capability to find a specific representation of the CheXpert data. The x-axis represents the MSE 111

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

7.5 Visualisation of score distributions from the pre-trained DL model and the EBM. The first row shows the distributions of softmax (DL model) and energy-based (EBM model) probabilities between the CheXpert dataset and the CXR14 dataset; the second row the the distributions between the CheXpert dataset and the InBreast dataset and the third the distributions between the CheXpert dataset and the ImageNet dataset. The bottom row shows the distribution probabilities from the energy model between the CheXpert dataset and the out of distribution datasets. The degree of overlap is indicative of how well the respective model can find fake data. The x-axis is the score, softmax or energy, and the y-axis in the frequency 114
7.6 Reliability Plots on different dataset sizes for the pre-trained CNN and the EBM. The first row indicates the complete dataset, the second 60% of the complete dataset and the third 30%. The x-axis is the binned predictive confidence and the y-axis the corresponding accuracy. Ideally, highly confident predictions should also be highly accurate. The blue bars indicate the accuracy of the specific confidence bin, the red bars indicate the gap between the achieved accuracy and the ideal calibrated value. . 117
7.7 Diagram showing the trajectory of accuracy and ECE during training. There is a clear downward trend of the calibration error (orange) and a clear upward trend of the AUROC (blue). The x-axis shows the number of epochs. 118
7.8 Top 9 most outlying images in CheXpert as found by robust PCA 119
7.9 Random samples generated in the course of training the EBM. During training, samples were periodically generated from the data distribution at that point in time. The top row shows samples at the beginning of training and resemble random noise. The second row shows samples midway though the training; outlines of torsos can be discerned but the images are still very blurred, indicating that the distribution does not represent the data well yet. The third row shows the samples generated with the trained model and the final row shows real CheXpert samples; model has clearly learned the most important characteristics of chest radiographs and can reproduce them to generate compelling samples. 120

List of Tables

5.1 Comparison of the deep learning and energy based frameworks 77

6.1 Table showing the distribution of gender in the CheXpert data 84
6.2 Table showing the distribution of x-ray views in the CheXpert data 85

6.3 Table showing the occurrence of pathologies, split by the type of view (frontal/lateral) 86
6.4 Table showing the prevalence of the "uncertain" label in the data before imputation 86

7.1 Mean AUROC scores, including confidence intervals, across all four pathologies for all models, organized according to the three different resolutions. Both the results of a pre-trained network and a network trained "from scratch" are reported and compared to the EBM. The "winning" model is indicated in bold. *interim results gathered before full training run completed due to the fact that the full run would take around 3 months which was beyond the scope of this thesis 106
7.2 Percent loss of accuracy as the models are trained on progressively lower image resolutions. The largest reduction for each image size is indicated in bold. All models have the highest accuracy on the highest image resolution, 224 × 224. 110
7.3 AUROC scores for the Out of Distribution Detection classification. Each model is exposed to all three OOD datasets, at every resolution. A high AUROC expresses a successful classification of in and out of distribution elements; an AUROC of 0.5 indicates that the model was not able to detect out of distribution elements 112
7.4 Comparison of ECE and MCE between CNN and EBM 118

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

List of Algorithms

3.1 Backpropagation algorithm for a Multilayer Perceptron. The computation yields the gradients with respect to the parameters of each layer, starting from the output layer and going backwards to the first layer. These gradients can be interpreted as an indication of how each layer's output should change to reduce error. 38

3.2 Stochastic Gradient Descent 39

6.1 EBM Training Algorithm Given network G_w, SGLD step-size α, SGLD std σ, replay buffer B, SGLD steps η, reinitialization frequency ρ 99

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Index

activation function, 23
Adaptive Moment Estimation, 39
artificial neural network, 20, 22
artificial neuron, 19, 22

backpropagation, 36
backpropagation algorithm, 36
black boxes, 66
Boltzmann machine, 49

calibration, 67
categorical distribution, 55
computer vision, 40
conditional probability, 31
confidence, 69
contrastive divergence, 61
convolution, 41
convolutional block, 43
convolutional layer, 41
Convolutional Neural Networks, 21
cost function, 30
cross entropy, 35

deep learning, 18
degradation, 44

energy function, 47
entropy, 35
epistemic uncertainty, 66
estimator, 33
expected sample accuracy, 69
explainable machine learning, 66

feature map, 41
feedforward networks, 22
filter, 41
Frobenius Norm, 88
fully connected architecture, 22
function approximation, 17

generative model, 48
Gibbs sampler, 60
gradient, 36
Gradient based methods, 36
Gradient descent, 37
graphical models, 56

Harmonium, 49
harmony, 49
Hessian Matrix, 38
hyperplane, 17

Joint Energy Based Model, 75

kernel, 41
Kullback-Leibler Divergence, 34

learning rate, 39
likelihood functions, 31
likelihoods, 31
local receptive field, 41
logits, 74

Markov chains, 60
Markov Networks, 56
Markov Property, 60
Markov Random Fields, 56
maximum softmax probability, 74
measure theory, 32
Metropolis-Hastings, 60

model distribution, 32
model uncertainty, 66
Monte Carlo Markov Chain, 59
multilayer perceptron, 19
multinomial distribution, 55

optimum, 36

partition function, 57
Persistent Contrastive Divergence, 61
Principal Component Analysis, 88
Principal Component Pursuit, 89

Radon–Nikodym derivative, 32
Radon–Nikodym Theorem, 31
regularizer, 52
reliability diagrams, 69
residual blocks, 44
residual mappings, 44
ResNet, 44
Restricted Boltzmann Machine, 49
Riesz Representation Theorem, 16
Rosenblatt Perceptron, 19

sensitivity/recall, 82
sigmoid function, 20
skip connection, 44
specificity, 82
statistical estimation, 30
Stochastic Gradient Langevin Dynamics, 62
stochastic processes, 60
structured probabilistic model, 56

transfer learning, 93
true positive rate, 82

undirected graphical models, 56
universal approximation, 25

vanishing gradient, 44

XOR function, 19

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acronyms

ADAM Adaptive Moment Estimation. 39, 40

ANN Artificial Neural Network. 20–23

AUROC Area Under Receiver Operating Characteristic. 8, 73, 80, 82, 102, 105, 106, 108, 112, 124, 132

CNN Convolutional Neural Network. 2, 4, 8, 21, 22, 40–43, 52, 53, 56, 65, 68, 74–76, 93, 97–99, 101, 124, 125

DL deep learning. 7, 8, 15, 18, 57, 65, 75, 76, 94, 96, 102, 106, 108, 109, 112, 114–116, 119, 125, 130, 131

EBM Energy Based Model. ix, 7, 9, 47–54, 56–59, 61–64, 74–77, 79, 80, 82–84, 92–97, 99–102, 105–108, 110, 112–115, 118–126, 130–132

ECE Expected Calibration Error. 8, 80

IID Independent Identically Distributed. 33, 34

JEM Joint Energy Based Model. 75

KL Kullback-Leibler Divergence. 34

MCMC Monte Carlo Markov Chain. 9, 48, 59–64, 76, 100, 106, 108, 126

ML Maximum Likelihood. 31, 33, 34

MLP Multilayer Perceptron. 19, 20, 22–25, 40

MSE mean squared error. 88, 90, 101, 110–112, 130

NLL Negative Log Likelihood. 31, 36, 57, 96–98

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

OOD Out of Distribution Detection. 72–74, 76, 80, 82, 90, 94, 96, 100–102, 110, 112, 115, 121–125, 130, 132

PCA Principal Component analysis. 88–90, 101, 105, 110–112, 119, 121, 124, 130, 131

PCD Persistent Contrastive Divergence. 61

PCP Principal Component Pursuit. 89

RBM Restricted Boltzmann Machine. 49, 50, 52

ROC Receiver Operating Characteristic. 73

SGD Stochastic Gradient Descent. 38–40, 44, 62, 99, 106

SGLD Stochastic Gradient Langevin Dynamics. 62, 99, 120

W.R.T. with respect to. 24, 31–33, 35, 37, 51, 53, 55

WRN Wide Residual Network. 45, 93, 94

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Bibliography

Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 2018.

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31, 2018.

Sungjin Ahn. *Stochastic Gradient MCMC: Algorithms and Applications*. University of California, Irvine, 2015.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific reports*, 9(1):1–10, 2019.

Vic Barnett and Toby Lewis. Outliers in statistical data. *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics*, 1984.

Robert G. Bartle. *The Elements of Integration and Lebesgue Measure*. Wiley, New York, 1995.

Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.

Julius Berner, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. The modern mathematics of deep learning. *arXiv preprint arXiv:2105.04026*, 2021.

D. Bertsimas and J.N. Tsitsiklis. *Introduction to linear optimization*. Athena Scientific, 1997.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*, 2018.

Monica Bianchini and Franco Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE transactions on neural networks and learning systems*, 25(8):1553–1565, 2014.

Patrick Billingsley. *Probability and Measure*. John Wiley and Sons, second edition, 1986.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Ludwig Boltzmann. Studien über das gleichgewicht der lebendigen kraft zwischen bewegten materiellen punkten. *Wiener Berichte*, 58:517–560, 1868.

Vivek S Borkar and Sanjoy K Mitter. A strong approximation theorem for stochastic recursive algorithms. *Journal of optimization theory and applications*, 100(3):499–513, 1999.

Nicolas Bourbaki. *Elements of Mathematics: General Topology*. Addison-Wesley, 1966.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72: 102125, 2021.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Tianshi Cao, Chin-Wei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250*, 2020.

Augustin Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Robust out-of-distribution detection for neural networks. *arXiv preprint arXiv:2003.09711*, 2020a.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020b.

Xiaoran Chen, Nick Pawlowski, Martin Rajchl, Ben Glocker, and Ender Konukoglu. Deep generative models in the real-world: an open challenge from medical imaging. *arXiv preprint arXiv:1806.05452*, 2018.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Patrick Ferdinand Christ, Florian Ettlinger, Felix Grün, Mohamed Ezzeldin A Elshaera, Jana Lipkova, Sebastian Schlecht, Freba Ahmaddy, Sunil Tatavarty, Marc Bickel, Patrick Bilic, et al. Automatic liver and tumor segmentation of ct and mri volumes using cascaded fully convolutional neural networks. *arXiv preprint arXiv:1702.05970*, 2017.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Harald Cramér. *Mathematical methods of statistics*, volume 43. Princeton university press, 1999.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

Guillaume de L'Hospital. *Analyse des infiniment petits, pour l'intelligence des lignes courbes.* Number 145-146. Imprimerie royale, Paris, 1696.

M. Degroot and S. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32:12–22, 1983.

Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, and Ludolf Erwin Meester. *A Modern Introduction to Probability and Statistics: Understanding why and how*, volume 488. Springer, 2005.

Louke Delrue, Robert Gosselin, Bart Ilsen, An Van Landeghem, Johan de Mey, and Philippe Duyck. Difficulties in the interpretation of chest radiography. In *Comparative interpretation of CT and standard radiography of the chest*, pages 27–49. Springer, 2011.

Richard Drake, A Wayne Vogl, and Adam WM Mitchell. *Gray's anatomy for students E-book*. Elsevier Health Sciences, fourth edition. edition, 2009.

Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. *Advances in Neural Information Processing Systems*, 30, 2017.

Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019.

Jacques Dutka. The early history of the factorial function. *Archive for History of Exact Sciences*, 43:225–249, 1991.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Eurostat. Health in the european union – facts and figures. Technical report, Eurostat, 2020.

R. A. Fisher. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700–725, 1925.

Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer, 1992.

David Foster. *Generative deep learning: teaching machines to paint, write, compose, and play*. O'Reilly Media, 2019.

Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and Cooperation in Neural Nets*, pages 267–285. Springer, 1982.

Yarin Gal. *Uncertainty in Deep Learning*. P.h.d, 2016.

Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

Josiah Willard Gibbs. *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundations of Thermodynamics*. C. Scribner's sons, 1902.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE transactions on medical imaging*, 35(5):1153–1159, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

Jun Han and Claudio Moraga. The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International workshop on artificial neural networks*, pages 195–201. Springer, 1995.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Geoffrey Hinton, Simon Osindero, Max Welling, and Yee-Whye Teh. Unsupervised discovery of nonlinear structure using contrastive backpropagation. *Cognitive Science*, 30(4):725–731, 2006a.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Geoffrey E Hinton, Terrence J Sejnowski, and David H Ackley. *Boltzmann machines: Constraint satisfaction networks that learn.* Carnegie-Mellon University, Department of Computer Science Pittsburgh, PA, 1984.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006b.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997a.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997b.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960, 2020.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

Dina M Ibrahim, Nada M Elshennawy, and Amany M Sarhan. Deep-chest: Multi-classification deep learning model for diagnosing covid-19, pneumonia, and lung cancer chest diseases. *Computers in biology and medicine*, 132:104348, 2021.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

J Edward Jackson. *A user's guide to principal components*. John Wiley & Sons, 2005.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, 2012.

IT Jolliffe. Principal component analysis. *Springer Series in Statistics*, 1986.

José Raniery Ferreira Junior, Diego Armando Cardona Cardenas, Ramon Alfredo Moreno, Marina de Fátima de Sá Rebelo, José Eduardo Krieger, and Marco Antonio Gutierrez. A general fully automated deep-learning method to detect cardiomegaly in chest x-rays. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, pages 537–542. SPIE, 2021.

S Kesavan. *Measure and Integration*. Springer, 2019.

Ross P Kindermann and J Laurie Snell. On the relation between markov random fields and social networks. *Journal of Mathematical Sociology*, 7(1):1–13, 1980.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Anton Kolesov, Dmitry Kamyshenkov, Maria Litovchenko, Elena Smekalova, Alexey Golovizin, and Alex Zhavoronkov. On multilabel classification methods of incompletely labeled biomedical text data. *Computational and Mathematical Methods in Medicine*, 2014, 2014.

Sotiris B Kotsiantis, Dimitris Kanellopoulos, and Panagiotis E Pintelas. Data preprocessing for supervised leaning. *International journal of computer science*, 1(2):111–117, 2006.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Dirk P. Kroese, Tim J. Brereton, Thomas Taimre, and Zdravko I. Botev. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6, 2014.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

P. Langevin. On the theory of brownian motion" ["sur la théorie du mouvement brownien,"]. *C. R. Acad. Sci. (Paris)*, 146:530–533, 1908.

Yann LeCun. *PhD thesis: Modeles connexionnistes de l'apprentissage (connectionist learning models)*. Universite P. et M. Curie (Paris 6), June 1987.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 1(0), 2006.

Yann LeCun et al. Generalization and network design strategies. *Connectionism in Perspective*, 19:143–155, 1989.

Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 31, 2018.

Gottfried Wilhelm Leibniz. *The early mathematical manuscripts of Leibniz.* Courier Corporation, 2012.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.

S-CB Lo, S-LA Lou, Jyh-Shyan Lin, Matthew T Freedman, Minze V Chien, and Seong Ki Mun. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE transactions on medical imaging*, 14(4):711–718, 1995.

Jun Lu. Gradient descent, stochastic optimization, and other tales. *arXiv preprint arXiv:2205.00832*, 2022.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. *Advances in Neural Information Processing Systems*, 30, 2017.

Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6437–6446, 2022.

David JC MacKay, David JC Mac Kay, et al. *Information theory, inference and learning algorithms.* Cambridge university press, 2003.

Frederik Maes, David Robben, Dirk Vandermeulen, and Paul Suetens. The role of medical image computing and machine learning in healthcare. In *Artificial Intelligence in Medical Imaging*, pages 9–23. Springer, 2019.

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

Jaime Melendez, Rick Philipsen, P. Chanda-Kapata, V. Sunkutu, Nathan Kapata, and B. Ginneken. Automatic versus human reading of chest x-rays in the zambia national tuberculosis prevalence survey. *International Journal of Tuberculosis and Lung Disease*, 21:880–886, 08 2017. doi: 10.5588/ijtld.16.0851.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021.

Tom M Mitchell and Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

Andriy Mnih and Geoffrey Hinton. Learning nonlinear constraints with contrastive backpropagation. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 1302–1307. IEEE, 2005.

Annette M Molinaro, Richard Simon, and Ruth M Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307, 2005.

Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

Michael A Nielsen. *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, USA, 2015.

Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021.

Wei Niu, Xiaolong Ma, Yanzhi Wang, and Bin Ren. 26ms inference time for resnet-50: Towards real-time execution of all dnns on smartphone. *arXiv preprint arXiv:1905.00571*, 2019.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2, 2019.

James R Norris and James Robert Norris. *Markov chains*. Cambridge university press, 1998.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Alexey A Novikov, Dimitrios Lenis, David Major, Jiri Hladuvka, Maria Wimmer, and Katja Bühler. Fully convolutional architectures for multiclass segmentation in chest radiographs. *IEEE transactions on medical imaging*, 37(8):1865–1876, 2018.

Peter O'Connor, Dan Neil, Shih-Chii Liu, Tobi Delbruck, and Michael Pfeiffer. Real-time classification and sensor fusion with a spiking deep belief network. *Frontiers in neuroscience*, 7:178, 10 2013. doi: 10.3389/fnins.2013.00178.

Kenta Oono and Taiji Suzuki. Approximation and non-parametric estimation of resnet-type convolutional neural networks. In *International Conference on Machine Learning*, pages 4922–4931. PMLR, 2019.

Harald Ostensen. Diagnostic imaging: what is it? when and how to use it where resources are limited? Technical report, World Health Organization, 2001.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Workshop on Autodiff*, 2017. URL `https://openreview.net/forum?id=BJJsrmfCZ`.

Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

Philipp Christian Petersen. Neural network theory. *University of Vienna*, 2020.

Eduardo HP Pooch, Pedro L Ballester, and Rodrigo C Barros. Can we trust deep learning models diagnosis? the impact of domain shift in chest radiograph classification. *arXiv preprint arXiv:1909.01940*, 2019.

Lorentz GBA Quekel, Alphons GH Kessels, Reginald Goei, and Joseph MA van Engelshoven. Miss rate of lung cancer on the chest radiograph in clinical practice. *Chest*, 115(3):720–724, 1999.

Sivaramakrishnan Rajaraman, Prasanth Ganesan, and Sameer Antani. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. *PloS one*, 17(1):e0262838, 2022.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Pranav Rajpurkar, Jeremy Irvin, Robyn Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Bhavik Patel, Kristen Yeom, Katie Shpanskaya, Francis Blankenberg, Jayne Seekins, Timothy Amrhein, David Mong, Safwan Halabi, Evan Zucker, and Matthew Lungren. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

practicing radiologists. *PLOS Medicine*, 15:e1002686, 11 2018. doi: 10.1371/journal.pmed.1002686.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*. John wiley & sons, 2005.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, 2019. doi: 10.1038/s42256-019-0048-x.

W. Rudin. *Functional Analysis*. International series in pure and applied mathematics. Tata McGraw-Hill, 1974. ISBN 9780070995581.

Walter Rudin. *Real and Complex Analysis, 3rd Ed.* McGraw-Hill, Inc., USA, 1987. ISBN 0070542341.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

Wilhelm Röntgen. üeber eine neue art von strahlen. *Aus den Sitzungsberichten der Würzburger Physik.-medic. Gesellschaft Würzburg*, page 137–147, 1895.

Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 448–455, Hilton Clearwater Beach Resort, Clearwater Beach, Florida

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

USA, 16–18 Apr 2009. PMLR. URL `https://proceedings.mlr.press/v5/salakhutdinov09a.html`.

Ravi K Samala, Lubomir Hadjiiski, Heang-Ping Chan, Chuan Zhou, Jadranka Stojanovska, Prachi Agarwal, and Christopher Fung. Severity assessment of covid-19 using imaging descriptors: a deep-learning transfer learning approach from non-covid-19 pneumonia. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, pages 426–430. SPIE, 2021.

Vikash Sehwag, Arjun Nitin Bhagoji, Liwei Song, Chawin Sitawarin, Daniel Cullina, Mung Chiang, and Prateek Mittal. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pages 105–116, 2019.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2): 336–359, Oct 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL `http://dx.doi.org/10.1007/s11263-019-01228-7`.

N Shah, K Everton, A Nordvig, B Nguyen, N Khandelwal, and D Mollura. Increasing access to diagnostic imaging in developing countries: the asha jyoti mobile clinic. *World Health Organization. Available at: http://www. who. int/medical_devices/global_forum I*, 3, 2010.

Jun Shao. *Mathematical statistics: exercises and solutions*. Springer Science & Business Media, 2006.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248, 2017.

Annpurna Singh and Shashank Kumar. An overview of pca and jpeg image compression scheme, 2016.

Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, Colorado Univ at Boulder Dept of Computer Science, 1986.

David C Spelic, Richard V Kaczmarek, Mike C Hilohi, and Albert E Moyal. Nationwide surveys of chest, abdomen, lumbosacral spine radiography, and upper gastrointestinal fluoroscopy: a summary of findings. *Health Physics*, 98(3):498–514, 2010.

Michael Spivak. *Calculus*. Benjamin, New York, 1967.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

Paul Suetens. *Fundamentals of Medical Imaging*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511596803.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

Yee Whye Teh, Max Welling, Simon Osindero, and Geoffrey E Hinton. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4(Dec):1235–1260, 2003.

Bart ter Haar Romeny. *A Deeper Understanding of Deep Learning: Opportunities, Applications and Risks*, pages 25–38. Springer, 01 2019a. ISBN 978-3-319-94877-5. doi: 10.1007/978-3-319-94878-2_3.

Bart M ter Haar Romeny. A deeper understanding of deep learning. In *Artificial Intelligence in Medical Imaging*, pages 25–38. Springer, 2019b.

Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071, 2008.

John W. Tukey and Mosteller Frederick. *The collected works of John W. Tukey*, volume 4 of *V. 1-2: The Wadsworth statistics/probability series*. Wadsworth Advanced Books and Software, 1965 - 1986.

Edwin J. R. van Beek and John T. Murchison. *Artificial Intelligence and Computer-Assisted Evaluation of Chest Pathology*, pages 145–166. Spriner International Publishing, 2019. ISBN 978-3-319-94878-2. URL `https://doi.org/10.1007/978-3-319-94878-2_12`.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Yezhen Wang, Bo Li, Tong Che, Kaiyang Zhou, Ziwei Liu, and Dongsheng Li. Energy-based open-world uncertainty modeling for confidence calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9302–9311, 2021.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

Paul Werbos. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Science*. Ph.d., January 1974.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Charles S White, Thomas Flukinger, Jean Jeudy, and Joseph J Chen. Use of a computer-aided detection system to detect missed lung cancer at chest radiography. *Radiology*, 252(1):273–281, 2009.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *International Conference on Machine Learning*, pages 2635–2644. PMLR, 2016.

Dmitry Yarotsky. Universal approximations of invariant maps by neural networks. *Constructive Approximation*, 55(1):407–474, 2022.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 27, 2014.

Laurent Younes. On the convergence of markovian stochastic algorithms with rapidly decreasing ergodicity rates. *Stochastics: An International Journal of Probability and Stochastic Processes*, 65(3-4):177–228, 1999.

Alan L Yuille and Chenxi Liu. Deep nets: What have they ever done for vision? *International Journal of Computer Vision*, 129(3):781–802, 2021.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020.

Song Chun Zhu, Yingnian Wu, and David Mumford. Frame: filters, random fields, and minimax entropy towards a unified theory for texture modeling. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 686–693. IEEE, 1996.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109(3):467–492, 2020.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.