HAW HAMBURG

MASTERTHESIS
Nima Chizari

The Impact of Image Resolution and Model Scaling on Deep Learning based Automated Chest Radiograph Interpretation

FAKULTÄT TECHNIK UND INFORMATIK
Department Informatik

Faculty of Computer Science and Engineering
Department Computer Science

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Nima Chizari

The Impact of Image Resolution and Model Scaling on Deep Learning based Automated Chest Radiograph Interpretation

Masterarbeit eingereicht im Rahmen der Masterprüfung
im Studiengang *Master of Science Informatik*
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Kai von Luck
Zweitgutachter: Prof. Dr. Andreas Meisel

Eingereicht am: November 18, 2021

Nima Chizari

Title of Thesis

The Impact of Image Resolution and Model Scaling on Deep Learning based Automated Chest Radiograph Interpretation

Keywords

Machine Learning, Deep Learning, Computer Vision, Medical Imaging

Abstract

The use of deep learning for automated chest radiograph interpretation has been largely hindered by the absence of an annotated dataset with an appropriate size. This problem seems to be solved by the CheXpert [21] dataset, which has been released publicly as a competition task. Multiple entries [30, 51] have also shown that it is possible to train a deep learning model successfully on their dataset by comparing the detection rate of their model to the detection rate of 3 radiologists on a test set of 500 studies on 5 pathologies, where the models outperformed the radiologists in most tasks. The authors used downscaled versions (320×320 or below) of the chest screenings as input to their model. This work empirically examines the impact of 5 different image resolutions on the detection rate Area Under Receiver Operating Characteristic Curve (AUROC) on the 5 evaluation tasks using various ImageNet pretrained models. The results hint at the potential of higher detection rates mainly caused by higher image input resolutions. The benefits are task dependent. In 3 of 5 cases, the models trained on an input resolution above 320×320 show greater detection rates, while the remaining two cases show declining detection rates past this point.

Nima Chizari

Thema der Arbeit

Die Auswirkungen der Bildauflösung und Modellskalierung auf Deep Learning basierte automatisierte Röntgen-Thorax Interpretation

Stichworte

Maschinelles Lernen, Deep Learning, Bilderkennung, Medizinische Bildgebung

Kurzzusammenfassung

Der effektiven Nutzung Deep Learning basierter Verfahren für die automatisierte Röntgen-Thorax Interpretation fehlte es an einem annotierten Datensatz mit entsprechender Größe. Dieses Problem scheint durch den CheXpert Datensatz [21], welcher als Wettbewerbsaufgabe veröffentlicht wurde, gelöst zu sein. Mehrere Teilnehmer [30, 51] konnten bereits zeigen, dass man erfolgreich ein Deep Learning Modell auf diesem Datensatz trainieren kann, indem sie die Detektionsrate ihres Modells mit der Detektionsrate von 3 Radiologen auf 500 Studien bestehend aus 5 verschiedenen Pathologien verglichen. Die Modelle übertrafen die Detektionsrate der Radiologen im Großteil der Pathologien. Die Autoren nutzten die Röntgenbilder in runterskalierten Varianten (320×320 oder kleiner) als direkte Eingabe zu den Modellen. Diese Arbeit untersucht empirisch die Auswirkung von 5 verschiedenen Bildauflösungen, gemessen in der Metrik Area Under Receiver Operating Characteristic Curve (AUROC), an den gleichen 5 Pathologien mit verschiedenen auf ImageNet vor trainierten Modellen. Die Resultate deuten auf potenziell höhere Detektionsraten, hauptsächlich verursacht durch höhere Bildauflösungen. Dieses Verhalten ist allerdings abhängig von der Pathologie. In 3 von 5 Fällen profitierten die Detektionsraten der Modelle, die auf einer Bildauflösung über 320×320 trainiert wurden, während die verbleibenden zwei Pathologien abnehmende Detektionsraten aufzeigen.

Contents

List of Figures

List of Tables

1 Introduction

Due to technological advancements in the field of computer vision, such as image classification [14, 16] and semantic image segmentation [27, 34], increased usage of deep learning could be observed in many domains, including medical image interpretation systems. Detecting pathologies such as diabetic retinopathy [10], skin cancer [7], arrhythmia [11] or hemorrhages [9] at the level of clinical professionals using medical images as inputs was suddenly made possible by these systems.

Chest radiography is the most common imaging examination globally, critical for screening, diagnosis and management of many life threatening diseases. Automated chest radiograph interpretation at the level of practicing radiologists could provide substantial benefit in many medical settings, from improved workflow prioritization and clinical decision support to large-scale screening and global population health initiatives. [21] These systems should not replace a medical professional. They should rather accelerate the process by providing a representation of the findings, which can be used by the medical expert as a second opinion.

The use of deep learning for this domain has been largely hindered by the absence of an annotated dataset with an appropriate size. This problem seems to be solved by the CheXpert [21] and MIMIC-CXR [23] datasets. Both use the same automated labeling approach to extract observations from freetext radiology reports for their training dataset.

The CheXpert Dataset [21] is freely available and was released as a competition task. Besides the commonly used metric AUROC, a custom expert human performance metric is provided. The annotations of three board-certified radiologists are used as benchmark on the test dataset on five clinically relevant pathologies. The top entries of the leaderboard consist of models that have higher detection rates than the majority of the radiologists.

The Top 3 leaderboard approaches leveraged optimizations in the pretraining of the models [30] and usage of novel loss functions [51] beating almost all benchmarked radiologists‘

performance in all tasks. The radiographs were used in an image resolution of 320×320 pixels or even below and the authors provided no reason, besides the lack of resources [51] for their choice.

Filling the lack of an ablation study regarding this dataset, this work tries to answer the question of what the optimal image resolution for training a deep learning based automated chest radiograph interpretation is and how it impacts the detection rate of pathologies. Additionally, the role of model scaling in accordance to the image resolution is investigated.

1.1 Outline

The thesis is split into 4 chapters. In chapter 2, the field and problem of automated chest radiograph interpretation is introduced and relevant related work is presented. The chapter ends with the final formulation of the research question.

Chapter 3 then discusses the design and implementation of the experiments to investigate the research question. This includes the dataset class distribution, data preprocessing, model architectures and training procedure.

The experiments are evaluated in chapter 4. A quantitative evaluation of the general impact of image resolution on the detection rate and the role of model scaling are both investigated. Additionally, a brief qualitative evaluation is provided by visualizing certain model outputs and comparing them to bounding box ground truths provided by a board-certified radiologist.

The final chapter 5 concludes the results and summarizes the outcome. A brief outlook for the future of the field is discussed and ideas and concerns that have emerged while producing this work are presented.

2 Problem Analysis

In this chapter, the problem of automated chest radiograph interpretation is introduced by providing brief descriptions of digital medical imaging and x-rays, characterization of anomalies in chest radiographs and approaches to automating the detection of pathologies in section 2.1. Furthermore, related work is presented in the form of introductions to the AUROC metric in section 2.2.1 and the landscape of publicly available labeled chest radiograph datasets in section 2.2.2. Additionally, the CheXpert dataset is presented more thoroughly in section 2.2.3, while ImageNet pretraining in medical imaging problems and model scaling are discussed in section 2.2.4 and section 2.2.5 respectively. A related paper is presented, which also examined variations of performance for multiple image resolutions using another dataset in section 2.2.6 to finally conclude in the research question this work tries to answer in section 2.3.

2.1 Problem Introduction

According to [17], medical imaging is the study of human functions and anatomy through pictorial information. In order to generate this pictorial information, multidisciplinary knowledge ranging from biology to computer science is required. Two of the methods and procedures studied in this field, are the convertibility of a conventional medical image to a digital image and analysis of this digital image according to a specific application or clinical need. Medical image representations come in many forms but for this study only conventional 2 dimensional X-rays are relevant. A digital image $P(x, y)$ is defined as an integer function of two variables x, y such that

$$0 \leq P(x, y) \leq N \; where \; x \leq m, \; y \leq n \; and \; x, y, m, n, N \in \mathbb{N}^* \tag{2.1}$$

$P(x, y)$ is the gray-level value of a picture element, or *pixel* at (x, y). m, n describe the amount of pixels that are available in a digital image. This is referred to as image reso-

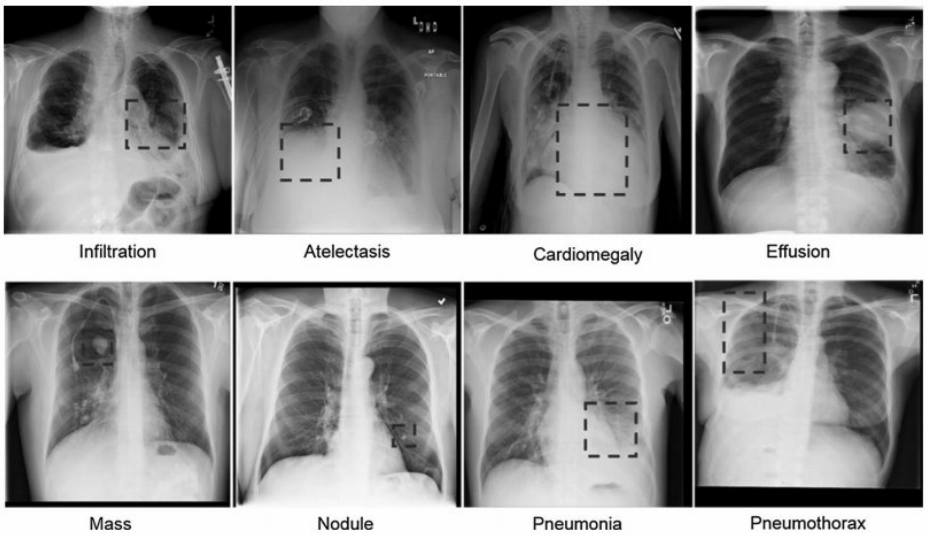


Figure 2.1: Eight common diseases detected in chest radiographs according to [32] (e.g. Infiltration, Atelectasis, Cardiomegaly, Effusion, Mass, Nodule, Pneumonia, Pneumothorax). Most chest radiograph datasets only provide *global* labels, mapping all of the pixels of a radiographic chest image to one or multiple labels. The red bounding boxes are not provided, and were only highlighted in this example.

lution. These parameters are responsible for the level of detail available and perceivable in an image. Digital images can also be represented as matrices or tables, where m could describe the amount of columns and n the number of rows. A conventional chest x-ray can have up to 4000×4000 pixels and each pixel can have a value ranging from 0 to 255.

[32] categorizes three main types of anomalies in chest radiographs: texture abnormalities, which are characterized by diffuse changes in the appearance and structure of the area, such as interstitial lesions; focal abnormalities, which are manifested as isolated changes in density, for instance pulmonary nodules; and abnormal shape, in which disease processes change the outline of the normal anatomy, namely cardiomegaly. This can be seen in image 2.1. Sometimes, the texture and shape of the chest changes at the same time as a certain disease, such as tuberculosis.

At the end of the 1990s, supervised techniques, where training data is used to develop a system, were becoming increasingly popular in medical image analysis. The concept of feature extraction and use of statistical classifiers (for computer-aided detection and diagnosis) was popularized. This pattern recognition or machine learning approach is still

very popular and forms the basis of many successful commercially available medical image analysis systems. Thus, we have seen a shift from systems that are completely designed by humans to systems that are trained by computers using example data from which feature vectors are extracted. Computer algorithms determine the optimal decision boundary in the high-dimensional feature space. A crucial step in the design of such systems is the extraction of discriminant features from the images. This process is still done by human researchers and, as such, one speaks of systems with *handcrafted* features.

The next logical step was to extract discriminant features that optimally represent the data for the problem at hand autonomously. This concept lies at the basis of many deep learning algorithms: models (networks) composed of many layers that transform input data (e.g. images) to outputs (e.g. disease present/absent) while learning increasingly higher level features. The most successful type of model for image analysis to date is the CNN.

In order to autonomously extract features by these powerful architectures, moderate to high amounts of labeled data is required. The output to each data point must be provided and this has proven itself as a problem in the field of medical imaging. The labeling can only be performed by radiologists. Each study must be labeled with the presence or absence of a pathology in order to successfully implement an automated medical interpretation system with supervised learning methods.

Most publicly available datasets only provide *global* labels, mapping all of the pixels of a radiographic chest image to one single label (e.g. disease present/absent). These types of labels only allow the development of a classifcation system. Localization of the findings is not possible directly by these systems. This would require segmentation information, where the affected area in the image would be annotated additionally. Segmentation information to accompany the assigned labels is more difficult to attain as this information is not captured in historic reports, and adds a great deal of expense in manual labeling approaches as manually drawing segmentations is typically a time intensive endeavor. [49]

2.2 Related Work

2.2.1 Performance measures for classification

In supervised learning, access to the data labels during the models training and validation stages is permitted and necessary. According to [42] classification falls into one of the following tasks, when data entries have to be assigned into predefined classes. First there is *binary* classification, where the input is to be classified into one, and only one, of two non-overlapping classes. Secondly there is *multi-class* classification where the input is to be classified into one, and only one, of l non over-lapping classes. In *multi-labelled* classification the input is to be classified into several of l non-overlapping classes. Lastly there is *hierarchical* classifcation, where the input is to be classified into one, and only one class which are themselves divided into subclasses or grouped into superclasses. The hierarchy is predefined and cannot be changed during classification.

The correctness of a classification can be evaluated by computing the number of correctly recognized class examples (true positives), the number of correctly recognized examples that do not belong to the class (true negatives), and examples that either were incorrectly assigned to the class (false positives) or that were not recognized as class examples (false negatives). These four counts constitute a confusion matrix shown in figure 2.2 for the case of binary classification.

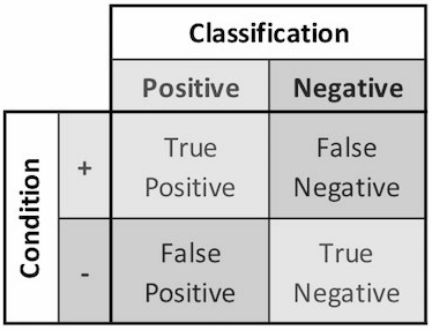


Figure 2.2: Confusion matrix in binary classification tasks. [48]

From the resulting confusion matrix, different performance measures or metrics can be derived. Among commonly used metrics are *Sensitivity* and *Specifity*. They are defined

as:

$$Sensitivity = \frac{tp}{tp + fn} \qquad Specifity = \frac{tn}{tn + fp} \tag{2.2}$$

These two performance measures are often combined and represented in the *Area Under the Curve (AUC)*, which captures a single point on the *Reception Operating Characteristic (ROC)* curve. AUC is defined as:

$$AUC = \frac{1}{2}(\frac{tn}{tn + fp} + \frac{tp}{tp + fn}) \tag{2.3}$$

When a classification model outputs probabilites between 0 and 1, the result can be controlled by a decision threshold dividing the outputs into the classes. Changing the threshold results in new confusion matrices and therefore new sensitivity and specifity values. This is leveraged in order to achieve the full ROC curve, which can be seen in figure 2.3.

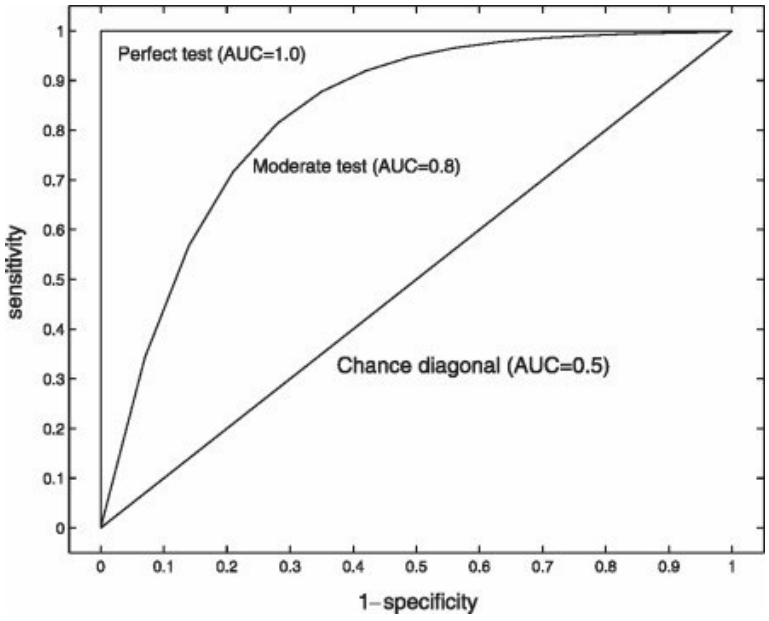


Figure 2.3: ROC curve for different cases. According to [6] AUROC serves as a well-established index of diagnostic accuracy. ROC following diagonal line results in AUC 0.5 (chance diagonal), whereas the maximum value of 1.0 corresponds to perfect assignment (unity sensitivity for all values of specificity)

According to [6] the ROC curve is a graphical presentation of sensitivity versus 1-specificity (or false positive rate) as the threshold is varying. Each point on the graph is generated by using a different threshold value. The AUROC gives a global measure of the classifier performance over a range of test thresholds. For example, as shown in figure 2.3 a test with an AUROC of 1.0 is perfectly accurate as the sensitivity is 1.0 when the specificity is 1.0 (perfect test). In contrast, a test with an AUROC of 0.0 is perfectly inaccurate. The line segment from (0,0) to (1,1) has an area of 0.5 and is called the chance diagonal. Tests with an AUC value larger than 0.5 have at least some discrimination ability. The closer the AUC reaches 1.0, the better the diagnostic test.

According to [29] ROC analysis was invented to counter the limitations of the *accuracy* performance measure, which is defined as:

$$Accuracy = \frac{tp + tn}{tp + fn + fp + tn} \tag{2.4}$$

Firstly, it does not consider class imbalance that often occurs in medical settings and applies equal cost to false positives and false negatives. Secondly, it is too generic and two diagnostic modalities can yield equal accuracies but perform differently with respect to the types of correct and incorrect decisions they provide; the incorrect diagnoses from one might be almost all false negative decisions (misses), while those from the other might be nearly all false positive decisions (false alarms).

2.2.2 Quantity & Quality of Labeled Chest Radiograph Datasets

Currently, publicly available chest radiograph datasets are plagued with a trade off between quality of labels and quantity of data. The choice is between large, multi-labeled datasets with risk of faulty labels [47, 21, 23] or a smaller dataset with less noisy labels where mostly, only one pathology is annotated [40, 22].

According to [49] the labels for datasets in this domain either rely on applying Natural Language Processing (NLP) techniques on historic medical reports or the use of networks of medical experts to prospectively read and annotate studies. Both approaches offer benefits and disadvantages. The usage of automated label extraction techniques on historic reports results in a overall larger dataset, but might introduce occasional noise and faulty labels. [47, 21, 23] Manual labeling by medical experts ensures correctness, but is economically expensive and therefore results in a smaller dataset size. [40, 22] These manual approaches often only cover one pathology, while the automated approaches

cover multiple pathologies. An overview of common publicly available chest radiograph datasets is given in table 2.1.

Name	Labeling Method	# Pathologies	# X-Rays
MC Dataset [22]	Manual Labeling	1	138
JSRT Dataset [40]	Manual Labeling	1	247
Shenzhen Dataset [22]	Manual Labeling	1	662
Indiana Dataset [5]	Manual Label Extraction	10	8.121
NIH Chest X-Ray14 Dataset [47]	Automated Label Extraction	14	108.948
CheXpert Dataset [21]	Automated Label Extraction	14	224.316
MIMIC-CXR Dataset [23]	Automated Label Extraction	14	371.920

Table 2.1: Overview of publicly available labeled chest radiograph datasets. Label extraction methods are based on historic radiology reports.

2.2.3 CheXpert Dataset

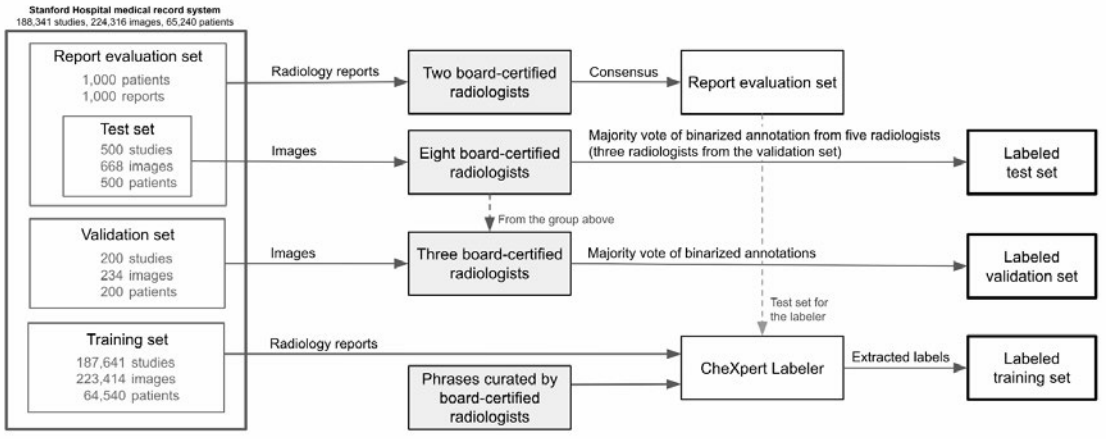


Figure 2.4: Sources for the training, validation, and test sets, highlighting the participation of the board-certified radiologists according to [8]

The CheXpert Dataset [21] consists of 224,316 chest radiographs taken of 65,240 patients. The radiographic examinations were collected from the Stanford Hospital and were performed between the time period of October 2002 and July 2017 in both inpatient and outpatient centers, along with their associated radiology reports.

Each report was labeled for the presence of 14 observations as positive, negative, or uncertain. In the training set, [21] decided on the 14 observations based on the prevalence in the reports and clinical relevance, conforming to the Fleischner Society's recommended glossary [12] whenever applicable. An automated rule-based labeler was developed to extract observations from the free text radiology reports to be used as structured labels for the images.

The labeler is set up in three distinct stages: mention extraction, mention classification, and mention aggregation. In the mention extraction stage, the labeler extracts mentions from a list of observations from the Impression section of radiology reports, which summarizes the key findings in the radiographic study. In the mention classification stage, mentions of observations are classified as negative, uncertain, or positive. In the mention aggregation stage, the classification for each mention of observations is used to arrive at a final label for the 14 observations (blank for unmentioned, 0 for negative, -1 for uncertain, and 1 for positive). This can be seen in figure 2.5.

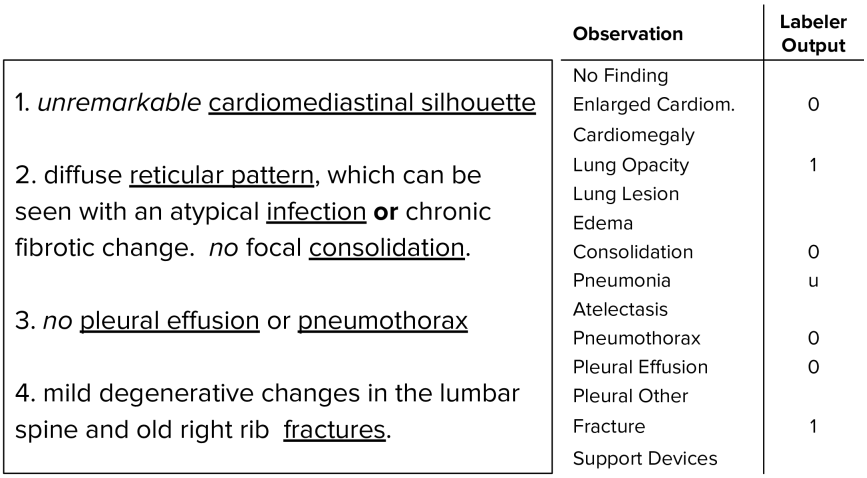


Figure 2.5: [21] automated rule-based labeler applied to extract observations from a free text radiology report

In the test and validation set however, [21] focuses on the evaluation of 5 observations which are called the competition tasks. These are selected based on clinical importance and prevalence: (a) Atelectasis, (b) Cardiomegaly, (c) Consolidation, (d) Edema, and (e) Pleural Effusion. The validation set consists of 200 studies on which the consensus of three radiologist annotations serves as ground truth.

The training and validation sets are comprised of 187,841 studies combined. 187,641 can be found in the training set. Each imaging study can pertain to one or more images. These can include multiple radiographs from a frontal or lateral view. The highest amount of concurrent images provided in a study is 3. Almost 82% of all studies consist of a single frontal x-ray image, followed by 16% which additionally provide a lateral image. Further information on the studies can be found in [3].

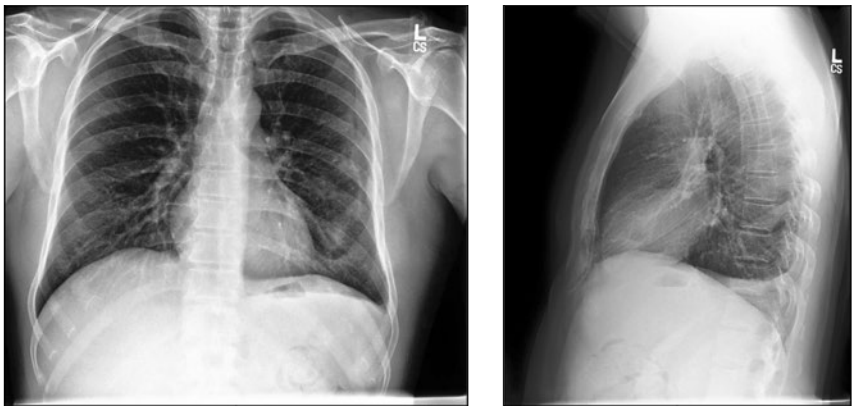


Figure 2.6: Left: frontal chest radiograph. Right: lateral chest radiograph

2.2.4 ImageNet Pretraining in Medical Image Interpretation Modeling

Most chest radiograph datasets presented in 2.2.2 are too small to train a CNN sufficiently. An effective technique to mitigate scarce data is *transfer learning*.

When initializing a neural network, there are two popular ways of setting up the learnable parameters. First there is random initialization, where the parameters are set up semi randomly with respect to some mathematical properties regarding the activation functions used. [13] Secondly there is transfer learning. In transfer learning, according to [50], a base network is trained on a base dataset and task, and afterwards the learned features are repurposed or transferred to a second target network to be trained on a target dataset and task.

The usual transfer learning approach is to train a base network and then copy its first n layers to the first n layers of a target network. The remaining layers of the target network are then randomly initialized and trained toward the target task. One can choose to backpropagate the errors from the new task into the base (copied) features

to fine-tune them to the new task, or the transferred feature layers can be left frozen, meaning that they do not change during training on the new task. The choice of whether or not to fine-tune the first n layers of the target network depends on the size of the target dataset and the number of parameters in the first n layers. If the target dataset is small and the number of parameters is large, fine-tuning may result in overfitting, so the features are often left frozen. On the other hand, if the target dataset is large or the number of parameters is small, so that overfitting is not a problem, the base features can be fine-tuned to the new task to improve performance. Of course, if the target dataset is very large, there would be little need to transfer because the lower level filters could just be learned from scratch on the target dataset.

Most base networks are trained on natural image datasets, usually *ImageNet* [35]. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) has been running annually (since 2010) and is the standard benchmark for large-scale object recognition. It consists of two components: (1) a publicly available dataset, and (2) an annual competition and corresponding workshop. The publicly released dataset contains a set of manually annotated training images. ILSVRC annotations fall into one of two categories: (1) image-level annotation of a binary label for the presence or absence of an object class in the image, e.g.,"there are cars in this image" but "there are no tigers," and (2) object-level annotation of a tight bounding box and class label around an object instance in the image, e.g., "there is a screwdriver centered at position (20,25) with width of 50 pixels and height of 30 pixels". The dataset comprises over 14 million natural images of more than 20.000 classes. These classes range from fruits to animals.

The base task therefore consists of classifying these natural objects in the image. Because of the domain gap between these natural images and medical images, the transferability of parameters of such base network for use in a medical imaging interpretation model remains an open research question.

[50] examines and quantifies the transferability of features from each layer of a neural network trained on ImageNet. The transferability is negatively affected by the specialization of higher layer features to the original task at the expense of performance on the target task. The transferability gap grows as the distance between tasks increases, particularly when transferring higher layers, but found that even features transferred from distant tasks are better than random weights. They also found that initializing with transferred features can improve generalization performance even after substantial fine-tuning on a

new task, which could be a generally useful technique for improving deep neural network performance.

[25] investigate the transferability of ImageNet performance to other computer vision tasks. They compare the performance of 16 classification networks on 12 image classification datasets. They find that, when networks are used as fixed feature extractors or fine-tuned, there is a strong correlation between ImageNet accuracy and transfer accuracy. On two small fine-grained image classification datasets, pretraining on ImageNet provides minimal benefits, indicating the learned features from ImageNet do not transfer well to fine-grained tasks. Their results show that ImageNet architectures generalize well across datasets, but ImageNet features are less general than previously suggested.

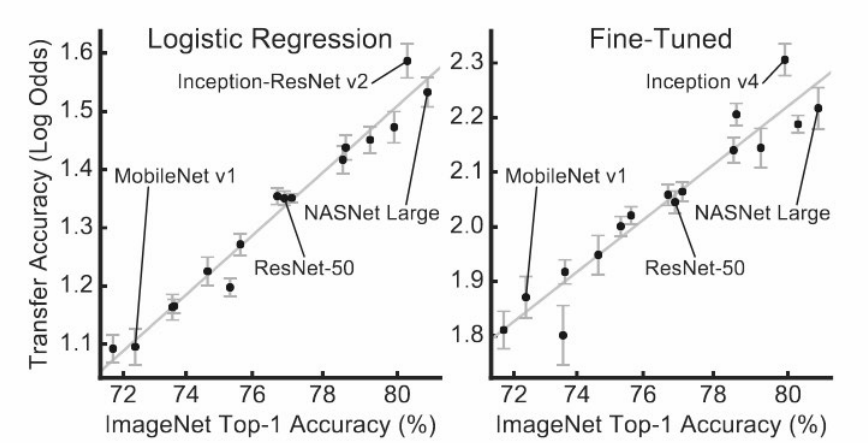


Figure 2.7: According to [25], transfer learning performance is highly correlated with ImageNet top-1 accuracy for fixed ImageNet features (left) and fine-tuning from ImageNet initialization (right). The 16 points in each plot represent transfer accuracy for 16 distinct CNN architectures, averaged across 12 datasets. Error bars measure variation in transfer accuracy across datasets.

[33] on the other hand, specifically investigated the use of ImageNet pre-training for medical imaging. They claim a performance evaluation on two large scale medical imaging tasks shows that surprisingly, transfer offers little benefit to performance, and simple, lightweight models can perform comparably to ImageNet architectures. Investigating the learned representations and features, they find that some of the differences from transfer learning are due to the over-parametrization of standard models rather than sophisticated feature reuse. Similar to the findings of [50], they also find that meaningful feature reuse is concentrated at the lowest layers.

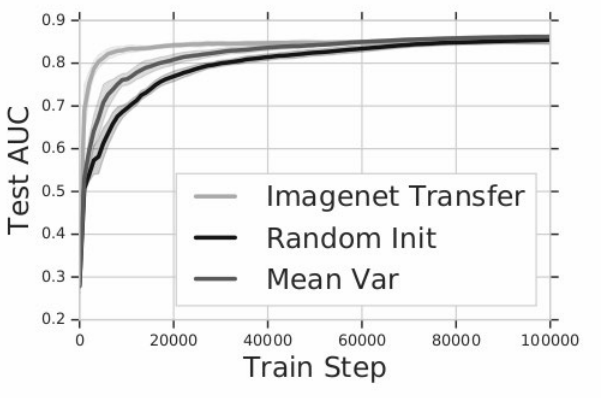


Figure 2.8: Convergence speed on the CheXpert consolidation pathology data and Resnet50 model architecture while using different parameter initialization methods. The figure compares ImageNet transfer learning and the Mean Var initialization scheme to random initialization. [50]

Among one other medical imaging dataset, they specifically examined the effectiveness of transfer learning for the CheXpert dataset. As can be seen in figure 2.8, they measured the convergence speed on three different initialization schemes. Among transfer learning and random initialization, the *Mean Var* initialization is used. This scheme initializes the parameters by using only the mean and variance of the pretrained weights, without using the pretrained parameters directly.

Contrary to their claims, using ImageNet pretraining offers faster model convergence than the other schemes. This can be seen in 2.8. Faster model convergence with transfer learning was also mentioned in [25]. Model convergence is achieved when the loss function has reached a minima and additional training will not improve the model. Most modern deep learning frameworks (e.g. PyTorch) offer ImageNet pretrained parameters for a wide variety of model architectures [1]. Usage of these parameters is therefore effortless and can be utilized easily.

2.2.5 Model Scaling in Deep Learning

As can be seen in image 2.1, chest pathologies can be very fine grained and nuanced. Scaling down the pixels of the image might blur or convolute the region of interest, decreasing the detection rate of a model. In order to process higher input dimensions, it might be necessary to scale other components of a CNN as well.

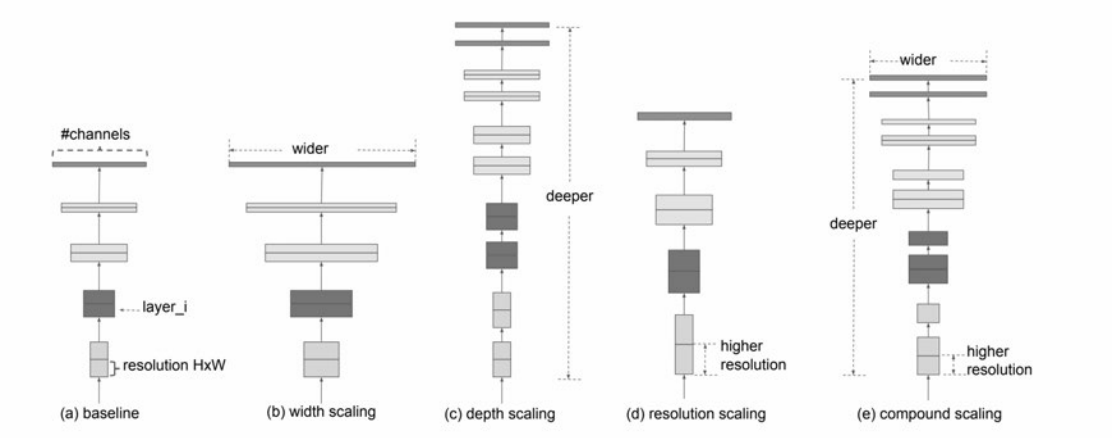


Figure 2.9: Model Scaling according to [46] (a) is a baseline network example; (b)-(d) are conventional scaling that only increases one dimension of network width, depth, or resolution. (e) proposed compound scaling method that uniformly scales all three dimensions with a fixed ratio.

According to [46] there are three scaling dimensions in a CNN: *depth*, *width* and *resolution*. Depth describes the amount of layers, width the number of neurons/channels per layer and resolution the amount of input pixels. The authors directly affiliate input resolution as being part of model scaling and therefore the model. We see them as separate components.

Scaling network depth is the most common way, used by many CNN architectures. [14, 16]. The intuition is that deeper networks can capture richer and more complex features, and generalize well on new tasks. However, these networks are also more difficult to train due to the vanishing gradient problem. [52] Although several techniques, such as skip connections [14] and batch normalization [20] alleviate the training problem, the accuracy gain of very deep network diminishes.

Scaling network width is commonly used for small size models. [15, 37, 45] Wider networks tend to be able to capture more fine-grained features and are easier to train. However, extremely wide but shallow networks tend to have difficulties in capturing higher level features.

With higher resolution input images, CNNs can potentially capture more fine-grained patterns. Starting from 224×224 pixels in early architectures, modern networks tend to use 299×299 [44] or 331×331 [53] for better accuracy. Recently, GPipe [18] achieved state-of-the-art ImageNet accuracy with 480×480 resolution.

[46] claim it is critical to balance all dimensions of network width/depth/resolution, and such balance can be achieved by simply scaling each of them with constant ratio. Based on this observation, they propose a *compound scaling method*. This can be seen in image 4.2. Based on a coefficient, a baseline network called *EfficientNet*, is scaled in all three dimensions.

2.2.6 Effect of Image Resolution on Automated Chest Radiograph Interpretation

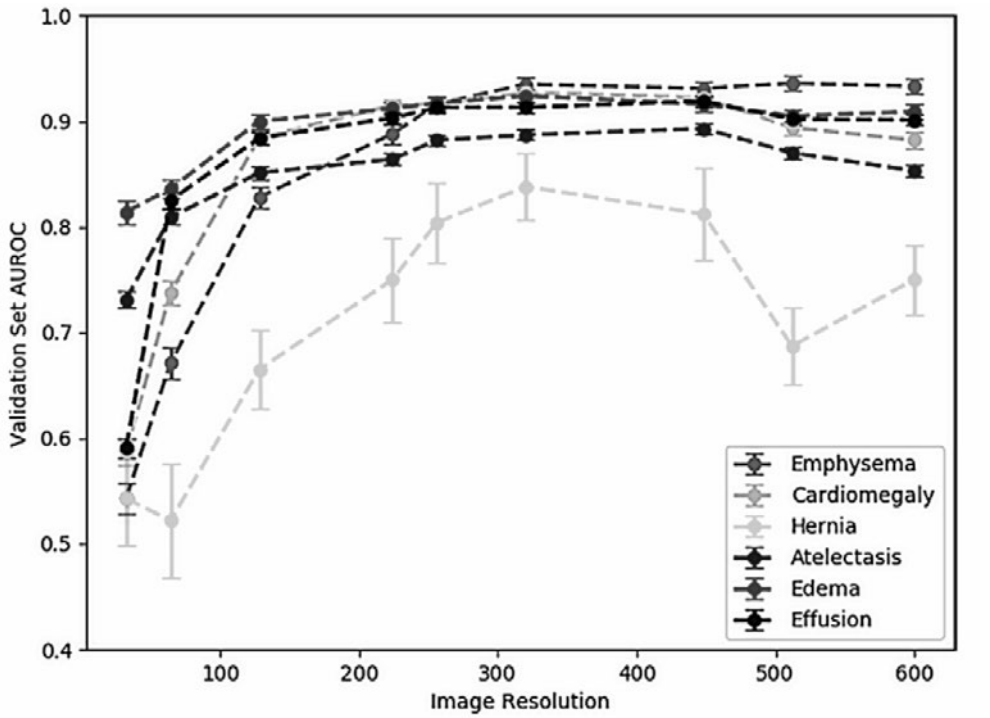


Figure 2.10: Validation set AUROC for six different diagnostic labels shows improved performance with increased image resolution and a plateau effect on performance improvement for resolutions higher than 224×224 pixels According to [36]. Models were trained with ResNet34 architecture. Resolutions shown are as follows: 32×32, 64×64, 128×128, 224×224, 256×256, 320×320, 448×448, 512×512, and 600×600 pixels. Error bars represent standard deviation of the AUROC calculated via the DeLong method.

[36] also examined variations of CNN performance for multiple chest radiograph diagnosis and image resolutions. They used the publicly available National Institute of Health (NIH) chest radiograph dataset (ChestX-ray14) [47] comprising 112.120 chest ra-

diographic images from 30.805 patients. The network architectures examined included ResNet34 [14] and DenseNet121 [16]. Image resolutions ranging from 32×32 to 600×600 pixels were investigated.

For this dataset and the chosen architectures, maximum AUROCs were achieved at image resolutions between 256×256 and 448×448 pixels for binary decision networks targeting different pathologies. Different diagnosis or image labels can have different model performance changes relative to increased image resolution (eg, pulmonary nodule detection benefits more from increased image resolution than thoracic mass detection).

One major shortcoming of their work is the usage of different batch sizes in their experiment configurations. Because of hardware limitations, they were forced to reduce the batch size when increasing the image resolution. The results of the experiments are therefore not comparable, because the differences in the detection rate could be a result of the differing batch size, not image resolution.

2.3 Research Question

The research questions revolve around one variable that was mostly overlooked in research regarding automated medical image interpretation, namely image resolution. Although not the focus of this work, researching this factor might reveal insights that push the current state-of-the-art in this field even further. This work therefore tries to answer the following questions:

1. What is the optimal image resolution to train a chest radiograph interpretation model in regards to the detection rate?
2. Is it necessary to scale other dimensions in a chest radiograph interpretation model in accordance to the input image resolution?

Presumably, as already mentioned in [36], optimal detection of individual pathologies will occur on different image resolutions. The detection rate of some disease classes will benefit more from higher input resolutions than others.

The methodology to examine these claims is by constructing empirical studies. An experimental setup with varying configurations is leveraged. ImageNet pretrained CNNs, namely ResNet [14], DenseNet [16] and EfficientNet [46] of varying scale are trained on five different downscaled image resolutions of the CheXpert dataset [21]. Two variants of

ResNet (ResNet-50, ResNet-152) and DenseNet (Densenet121, Densenet169) with varying scaling in depth and width will be used, while the suitable EfficientNet model based on the image resolution will be leveraged.

3 Design & Implementation

This section investigates the methodology and design of the experiments to investigate the research question from section 2.3. The CheXpert dataset will be leveraged. Although the training set of this dataset contains more classes, the validation set consists of only five classes, which are called the evaluation tasks. These are namely Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural Effusion. The experiments therefore will only concentrate on training models on these five classes.

Firstly, the class distribution of the evaluation tasks is investigated in section 3.1. For each class, multiple binary classification models in different configurations will be trained and evaluated. Image resolution and the preprocessing steps are discussed in section 3.2. The design of the model architectures is explained in section 3.3. Lastly, the training procedure and hyperparameters are discussed in 3.4.

3.1 Dataset Class Distribution & Uncertainty Approach

The uncertainty mapping approaches of [21] are leveraged. All of the uncertain labels of the classes Atelectasis, Edema and Pleural Effusion were mapped to positive examples, while labels of the Cardiomegaly class were mapped to the negative examples. The uncertain labels of the Consolidation class were ignored. This approach showed the best results on the validation set in [21]. Studies where a pathology is unmentioned are ignored and not mapped to negative cases.

This results in the class distribution in table 3.1. The number of training samples represents the number of studies meant for training the neural networks and optimizing the parameters. Each study consists of at least one x-ray image, but can have up to 3 concurrent images. The number of training samples range from 31,933 to almost 110,000 samples. The former represents the number of samples for the Consolidation class and the latter for Pleural Effusion. The validation sample size is 200 for every class.

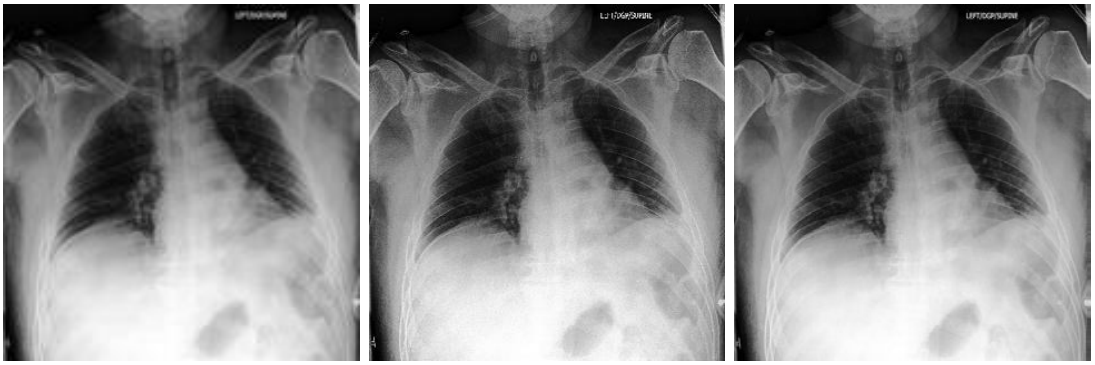
Pathology	Training Samples	Training Positive (%)	Training Negative (%)	Validation Positive (%)	Validation Negative (%)
Atelectasis	59717	58710 (98.31)	1007 (1.69)	75 (37.5)	125 (62.5)
Cardiomegaly	37311	23002 (61.65)	14309 (38.35)	66 (33.0)	134 (67.0)
Consolidation	31933	12730 (39.86)	19203 (60.14)	32 (16.0)	168 (84.0)
Edema	76143	60476 (79.42)	15667 (20.58)	42 (21.0)	158 (79.0)
Pleural Effusion	109947	85115 (77.41)	24832 (22.59)	64 (32.0)	136 (68.0)

Table 3.1: Class distribution after mapping uncertainty labels with the strategy of [21]. Studies where a pathology is unmentioned are ignored.

The class distribution for the training set are mostly biased towards positive examples with the consolidation set being the only exception, where the negative examples outweigh the positive. The opposite applies to the class distribution of the validation set, where the distribution of all classes is outweighed by the negative examples.

The Atelectasis class lacks negative examples and is extremely biased towards positive examples. This could result in inferior model performance.

3.2 Image Resolution & Preprocessing



(a) 160×160 (b) 320×320 (c) 480×480

Figure 3.1: Frontal radiographic image from the validation set of CheXpert [21] in three of five resized versions used for the experiments. X-ray taken of a 45 year old male patient with the atelectasis observation classified as positive by the consensus of three radiologists.

The x-ray images are provided in a resolution of up to 4000x4000 pixels. Each pixel is 8 bit encoded in a single channel. This means these images are grayscale and offer no color information. In a preprocessing step, all images are resized to the 5 image sizes 160×160,

240×240, 320×320, 400×400 and 480×480 using Lanczos algorithm implementation of the Python Image Library (PIL). The result can be seen in figure 3.1.

To optimally leverage transfer learning, further preprocessing of the images is necessary. The reason for this lies in the ImageNet dataset, which the pretrained models are trained with. Firstly, because the ImageNet data consists of colored RGB images, the pretrained model therefore expects three color dimensions as input. One way of mapping grayscale images for use in these pretrained models, is to simply duplicate the single color channel to three channels. This is leveraged in these experiments. Secondly, normalization of the pixel values can help boost model convergence. This is done with the standard and mean deviation of the ImageNet dataset, because the pretrained models were trained with this value distribution. The pixel values of the x-ray images are therefore divided by these values to achieve normalization.

Lastly, a suite of image augmentation is applied for regularization purposes. These consist of random rotations of up to 30°, random brightness/contrast fluctuations, black pixel padding and minimal perspective warping.

3.3 Model Architectures

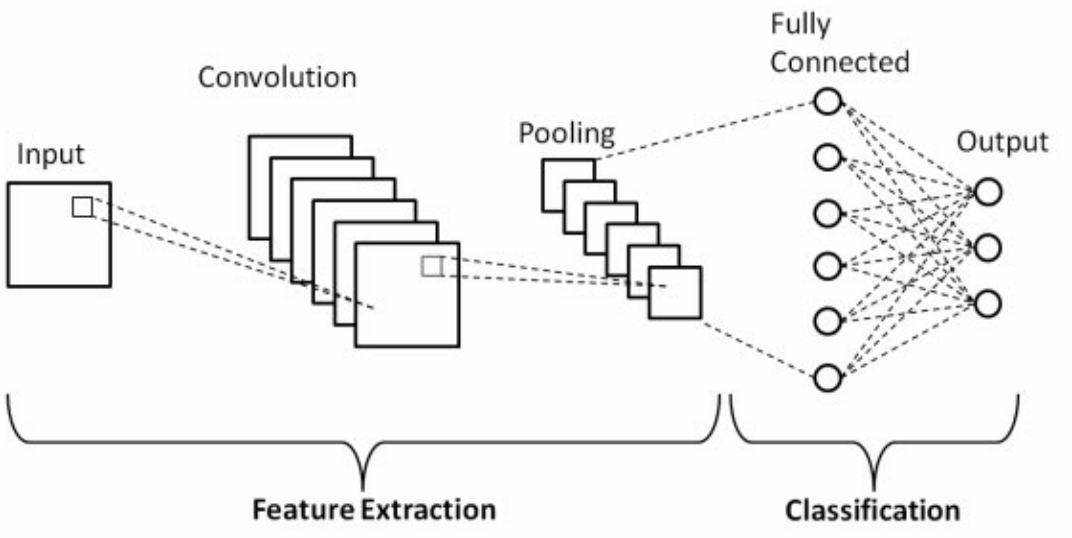


Figure 3.2: A CNN is composed of two basic parts of feature extraction and classification. Feature extraction includes several convolution layers followed by a pooling layer. The classifier usually consists of fully connected layers. [31]

All of the models used consist of two components. First, there is the *body* or *feature extractor* of the neural network. This part of the model is responsible for generating features by processing the pixel values of the preprocessed radiographic images. These features are then further processed in the second component of the network, referred to as the *head* or *classifier*. Based on the generated features of the model's body, this part of the neural network generates the final output. The general concept is visualized in figure 3.2. Because this is a binary classification problem, the output consists of a single value, which can be interpreted as the probability that the observation occurs in the study. When more than one view is available, the model outputs the maximum probability of the observation across the views.

The ResNet, DenseNet and EfficientNet architectures will be used as feature extractors of the neural networks. For each of the former two architectures, two ImageNet pretrained versions, namely ResNet50, ResNet152, DenseNet121 and DenseNet161 are leveraged. The different versions of the ResNet architecture only differ in the number of layers (depth) present in the model. In addition to the number of layers, the DenseNet121 and DenseNet161 differ in the growth factor k, which corresponds to the number of channels/feature maps (width) produced per dense-block. This also results in different amounts of output features. These pretrained models were trained on the image resolution of 224×224. For the EfficientNet models, the implementation of [28] was used. Each scaled EfficientNet model offers an ImageNet pretrained version which was trained on scaled versions of the dataset. While each ResNet and DenseNet model was trained on all 5 image resolutions for the experiments, each appropriate EfficientNet model was only trained on the next biggest (or equal) scaled image resolution version. This can be seen in Table 3.2.

Name	# Parameters	# Features	ImageNet Top-1 Acc.	Pretraining Resolution	Experiment Resolutions
ResNet50	23,508,032	2048	76.1	224×224	All
ResNet152	58,143,808	2048	78.3	224×224	All
DenseNet121	6,953,856	1024	74.4	224×224	All
DenseNet161	26,472,000	2208	77.1	224×224	All
EfficientNet-B0	4,007,548	1280	76.3	224×224	160×160
EfficientNet-B1	6,513,184	1280	78.8	240×240	240×240
EfficientNet-B3	10,696,232	1536	81.1	300×300	320×320
EfficientNet-B4	17,548,616	1792	82.6	380×380	400×400
EfficientNet-B5	28,340,784	2048	83.3	456×456	480×480

Table 3.2: Overview of ImageNet pretrained models used for feature extraction. Number of parameters only refers to the number of learnable parameters. Number of features refers to the number of output features.

The features produced by the body are then aggregated in a pooling layer and further processed in the head of the network. Other than the number of features taken as input into this pooling layer, the architecture of the head for each neural network is identical. The pooling layer consists of both averaging and using the maximum as aggregation strategies for the features. The results of these operations are concatenated, which doubles the number of features available. These features are further processed in two subsequent linear layers to produce a final output. The last layer is a sigmoid activation function which produces the probability (float value between 0 and 1) of a class occurring in the input image. Because these latter layers of the pretrained model are specialized for the base task, the weights can not be used. The parameters of these layers are initialized semi randomly, using the *Kaiming initialization* scheme [13].

The custom head combined with a ResNet50 feature extractor is visualized as a graph in figure 3.3. Outgoing channel dimensions of each layer are represented as a 4 dimensional tensor $batchsize \times channels \times height \times width$. The same model is represented with an input resolution of 240×240 and 480×480. The effect of using different input resolutions can be seen here. While the channel dimensions are equal, the height and width dimensions differ up to the interfacing point of the network between feature extractor and classifier, which is the pooling layer. While the smaller input resolution produces $2048 \times 8 \times 8$ features, the larger input resolution produces $2048 \times 15 \times 15$. This should result in more fine grained features. In order to further process the features without increasing the complexity/parameters of the network, the kernel sizes of both the max

pooling and average pooling layers are adapted to the height and width of the incoming features, resulting in identical output shapes. The number of trainable parameters of the head of the network is therefore only dependent on the number of output features from the feature extractor. The fully connected layers need to scale accordingly, in order to process the incoming features. The number of parameters for the classifier increases proportionately with the number of output features. This can be seen in table 3.3.

The ResNet50 and ResNet152 model variants with an identical input resolution of 480×480 are also visualized in figure 3.3. They output the same number of output features to the head of the network. This is achieved by only scaling the depth of the ResNet152 variant. This enables more computation on the input signal, which should have a positive effect on the detection rate. On the contrary, there is also a risk of overparametrization and therefore overfitting on the training data.

The different variants of the DenseNet model architecture, DenseNet121 and DenseNet161 are visualized in figure 3.4. Both graphs visualized use an input image resolution of 480×480. The variants differ in both width and depth. The prior can be seen by the number of Dense Layers. The latter is apparent by the outgoing channel dimensions, which is a result of differing growth factors k. The DenseNet121 variant uses a k value of 32, while the DenseNet161 variant uses a k value of 48. The channel growth is controlled by setting the output number of feature maps per Dense Layer to k and the usage of pooling in the transition layers.

The models, which use the EfficientNet model architecture as feature extractors, are visualized in figure 3.5. Because this architecture uses compound scaling it results in a unique model per image resolution. Both depth and width are scaled based on the number of input pixels. Each model variant expects a certain image resolution as input and the closest matching variant to our image resolution was chosen. This can be seen in table 3.2. For each model, the ImageNet pretrained variant was leveraged. One shortcoming of the models which use this model architecture as a feature extractor, is the usage of pooling in the interfacing part of the network between body and head. In the implementation, the model ends with an average pooling layer, which was overlooked by us. The custom head also starts with max and average pooling, which is now obsolete.

The choice of this set of models enables the examination of the different dimensions of model scaling described in section 2.2.5, namely depth, width and image resolution. Measuring the performance difference between ResNet50 and ResNet152 on scaled image resolutions enables the assessment of the dimensions depth and image resolution, while

Name	# Body Parameters	# Features	# Head Parameters	Sum Parameters
ResNet50	23,508,032	2048	2,106,880	25,614,912
ResNet152	58,143,808	2048	2,106,880	60,250,688
DenseNet121	6,953,856	1024	1,054,208	8,008,064
DenseNet161	26,472,000	2208	2,271,360	28,743,360
EfficientNet B0	4,007,548	1280	1,317,376	5,324,924
EfficientNet B1	6,513,184	1280	1,317,376	7,830,560
EfficientNet B3	10,696,232	1536	1,580,544	12,276,776
EfficientNet B4	17,548,616	1792	1,843,712	19,392,328
EfficientNet B5	28,340,784	2048	2,106,880	30,447,664

Table 3.3: Parameter counts of the ImageNet pretrained models and the resulting custom head. These two components were combined and used for the experiments. Number of head parameters scales according to the number of output features. Number of parameters only refers to the number of learnable parameters. Number of features refers to the output of the feature extractor.

DenseNet121 and DenseNet161 and the compound scaling of the EfficientNet models enables the examination for all three dimensions.

3.4 Training Procedure & Hyperparameters

Each pretrained model is initialized with the custom head formerly described and trained on each of the five image resolutions. A fine-tuning approach is leveraged. Therefore, the training procedure consists of two steps. Firstly, the ImageNet pretrained body of the model is frozen and only the head of the network is trained for one epoch. Secondly, the body of the model is unfrozen and both the head and body of the network are trained for 5 additional epochs. This is important because the head of the network is initialized semi randomly. The loss caused by the head of the network needs to be minimized before propagating the loss further into the body of the network and upgrading the weights.

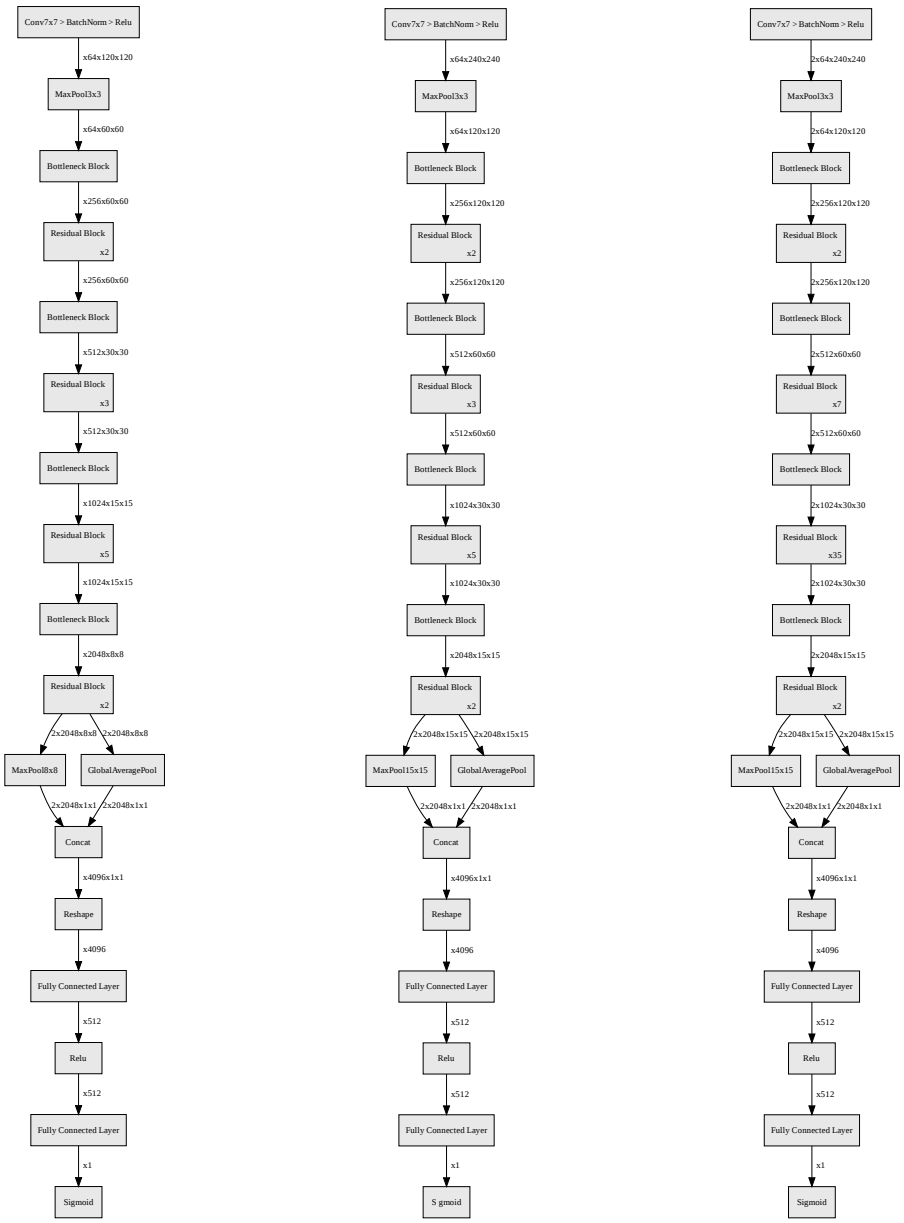
Binary cross entropy loss and a batch size of 40 was used for all experiments. One element of a batch contains one study, which can consist of up to 3 x-ray images. Adam [24], in conjunction with the 1-cycle learning rate [41] schedule were used as the optimizer with default parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$. The maximum learning rate was set to 5×10^{-3} for the first epoch. Discriminative learning rates were used afterwards. The

body of the network was trained with a maximum learning rate of 2.5×10^{-5} and the head with a maximum learning rate of 2.5×10^{-3}. Further regularization was introduced with Dropout [43] at a p-value of 0.6 in the linear layer of the head and Weight Decay, which was set to 1×10^{-1}. A checkpoint is saved after every epoch, if a higher AUC or lower loss value was achieved on the validation set in comparison to the earlier epochs.

These are fairly standard practically used hyperparameters with default parameters, which ensure good but not optimal results for most datasets. The focus of this work is not to train a new state of the art model for this dataset, which would probably require hyperparameter tuning among other optimizations. The more important aspect is to ensure that the changes in the detection rate (AUC) are caused by the differing image resolution. This is guaranteed by the usage of identical hyperparameters for each run and mitigation of randomness. Because the parameter initialization of the head is dependent on random factors, each experiment is repeated 5 times with differing seeds. Batch shuffling is disabled.

For each image resolution, the 5 models mentioned in section 3.3 are trained 5 times with differing seeds. This results in 125 experiments per image resolution. For the remainder of this work, one combination of image resolution, model architecture and seed will be referred to as one *run*. The choice in the extensive number of 6 epochs per run and training with five different model architecture ensures that the constellation of model and image resolution can reach its maximum potential and not underfit.

All experiments were performed in parallel on different virtual machines with varying amounts of NVIDIA Quadro P6000 24GB GPUs, kindly provisioned by [4]. Experiments were implemented using the Fastai library in conjunction with Pytorch. Further information on the infrastructure and software used can be found in my previous work [2].

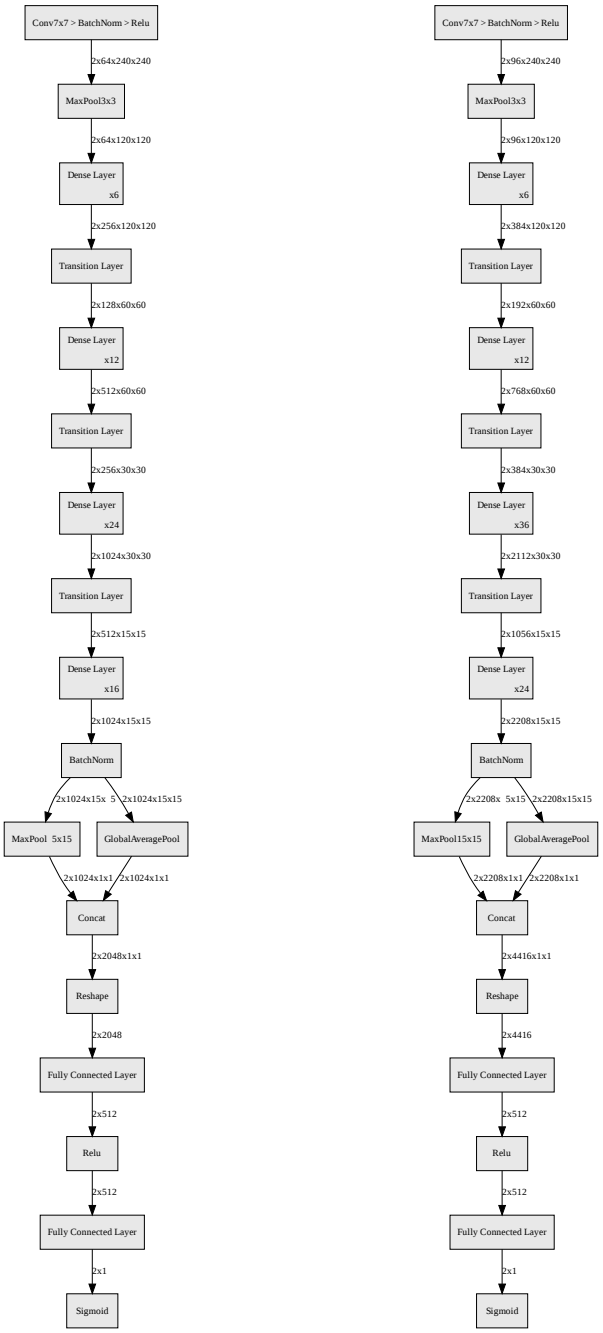


(a) ResNet50 with input resolution of 240×240

(b) ResNet50 with input resolution of 480×480

(c) ResNet152 with input resolution of 480×480

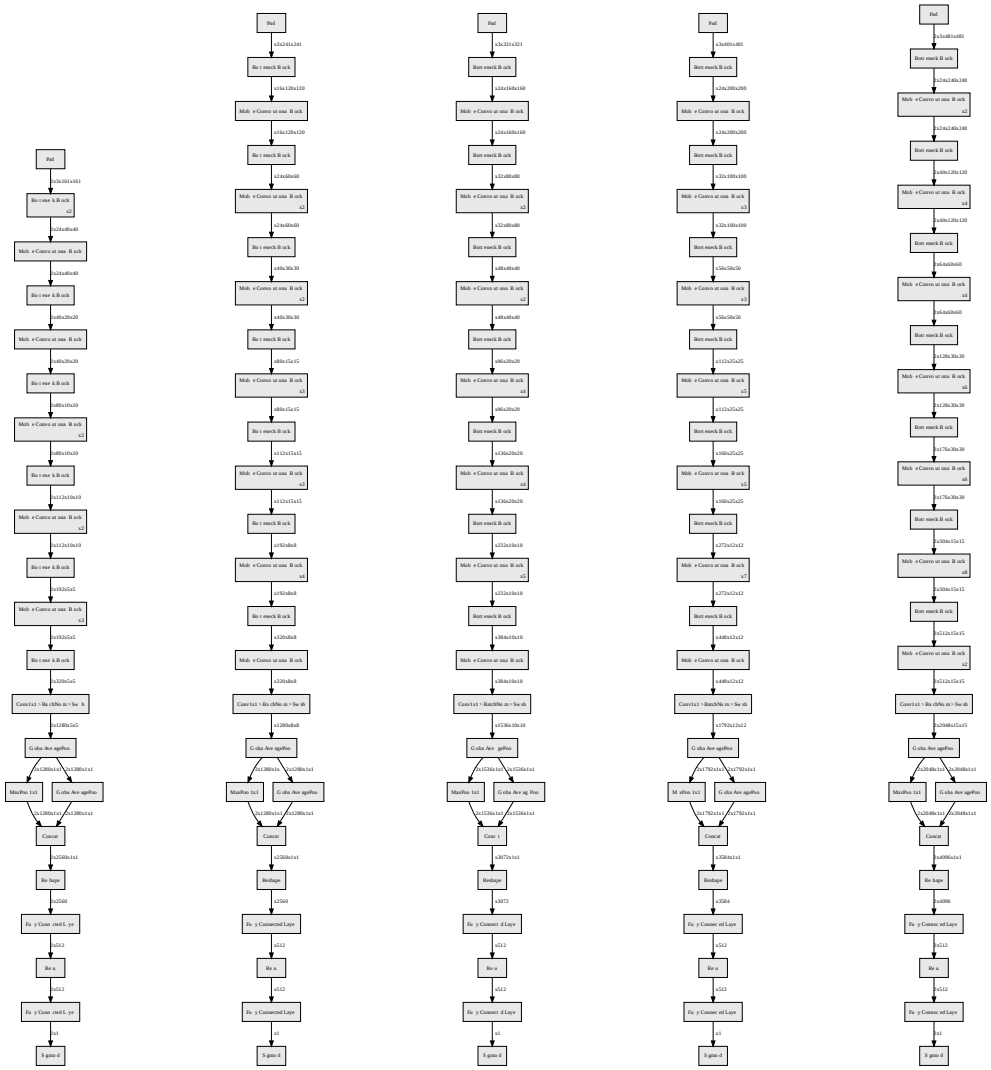
Figure 3.3: Graph visualization of ResNet50 and ResNet152 used as feature extractors with the custom classifier head for the experiments with different input resolutions. Outgoing channel dimensions of each layer represented as a 4 dimensional tensor $(batchsize) \times (channels) \times (height) \times (width)$.



(a) DenseNet121 with input resolution of 480×480

(b) DenseNet161 with input resolution of 480×480

Figure 3.4: Graph visualization of DenseNet121 and DenseNet161 feature extractors with the custom classifier head used for the experiments with identical input image resolutions. Feature extractor produces different number of features because of differing channel growth factors k.



(a) Efficientnet-b0 with input resolution of 160×160

(b) Efficientnet-b1 with input resolution of 240×240

(c) Efficientnet-b3 with input resolution of 320×320

(d) Efficientnet-b4 with input resolution of 400×400

(e) Efficientnet-b5 with input resolution of 480×480

Figure 3.5: Graph visualization of compound scaled ImageNet pretrained EfficientNet architectures used as feature extractors with the custom classifier head for the experiments with different input resolutions. Outgoing channel dimensions of each layer represented as a 4 dimensional tensor $(batchsize) \times (channels) \times (height) \times (width)$.

4 Evaluation

In this section an evaluation of the results is presented. As a means to measure the detection rate of the models, the AUROC metric is used. For this, the scikit-learn's implementation [38] of calculating the AUROC score from the prediction scores of each model on the validation set is leveraged. No test time augmentation was used. The evaluation consists of both quantitative and qualitative approaches.

For each of the 625 runs, the best performing epoch is chosen based on its validation AUROC score. In order to examine the experiments empirically, the results are grouped and aggregated in different ways and the resulting maximum and mean AUROC is evaluated. Firstly, the general relationship between image resolution and detection rate is examined in section 4.1.1. Secondly, the importance of model scaling in accordance to the image resolution is investigated in section 2.2.5.

For the qualitative evaluation, an explainable approach to certain model predictions is presented. Using the Grad-CAM method, visual explanations via gradient-based localization of the predictions are presented per image resolution from the best performing models based on the AUROC score and are compared to bounding box annotations made by a board-certified radiologist in section 4.2.

4.1 Quantitative Evaluation

4.1.1 Image Resolution Scaling

This section will focus on the relationship between general image resolution and detection rate, in this case measured in the metric AUROC score. The first hypothesis of this work will be investigated by analyzing and interpreting the results of 625 experiments.

Of each of the 625 runs, the best performing epoch is considered based on the AUROC score. Grouping the resulting scores on the image resolution and reporting the mean,

mean standard deviation and maximum AUROC score per image size results in figure 4.1. The performance of the different model architectures will be discussed in the next section.

As can be seen in figure 4.1, three of the five competition tasks scale well with increasing image resolution. The Edema class signals a high correlation between the variables, with both the mean of all runs and maximum scores increasing proportionately with the image resolution. Maximum scores also scale well for both the Atelectasis and Consolidation class, although the mean starts to settle or even worsen above an image size of 400×400 pixels. This could be the result of suboptimal hyperparameters or model architectures. The biggest AUROC score difference can be measured with the Atelectasis class, where a difference of up to 0.04 score points can be observed in the maximum. For all three classes the best performing model was achieved by using the x-rays in the maximum image resolution of 480×480 pixels used in the experiments.

Two of the five competition tasks do not benefit from increased image resolution above 320×320 pixels. Pleural Effusion is best detected by the models at an image resolution of 320×320 with the mean and maximum reaching the highest AUROC score values. The standard deviation is also smallest at this image resolution. This also applies to the Cardiomegaly class, where the best performing model was trained with an input image resolution of 320×320. For this class, no significant changes of performance can be measured with up scaling of input image resolutions up to 320×320 pixels, only a decline in performance afterwards.

4.1.2 Model Scaling

In order to investigate the role of model scaling in the performance gains shown in section 4.1.1, the highest AUROC score of each run is grouped by image resolution and model architecture. For each architecture type we will be defining Δ as

$$\Delta max(model\_type) = max(larger\_model) - max(smaller\_model) \tag{4.1}$$

$$\Delta mean(model\_type) = mean(larger\_model) - mean(smaller\_model) \tag{4.2}$$

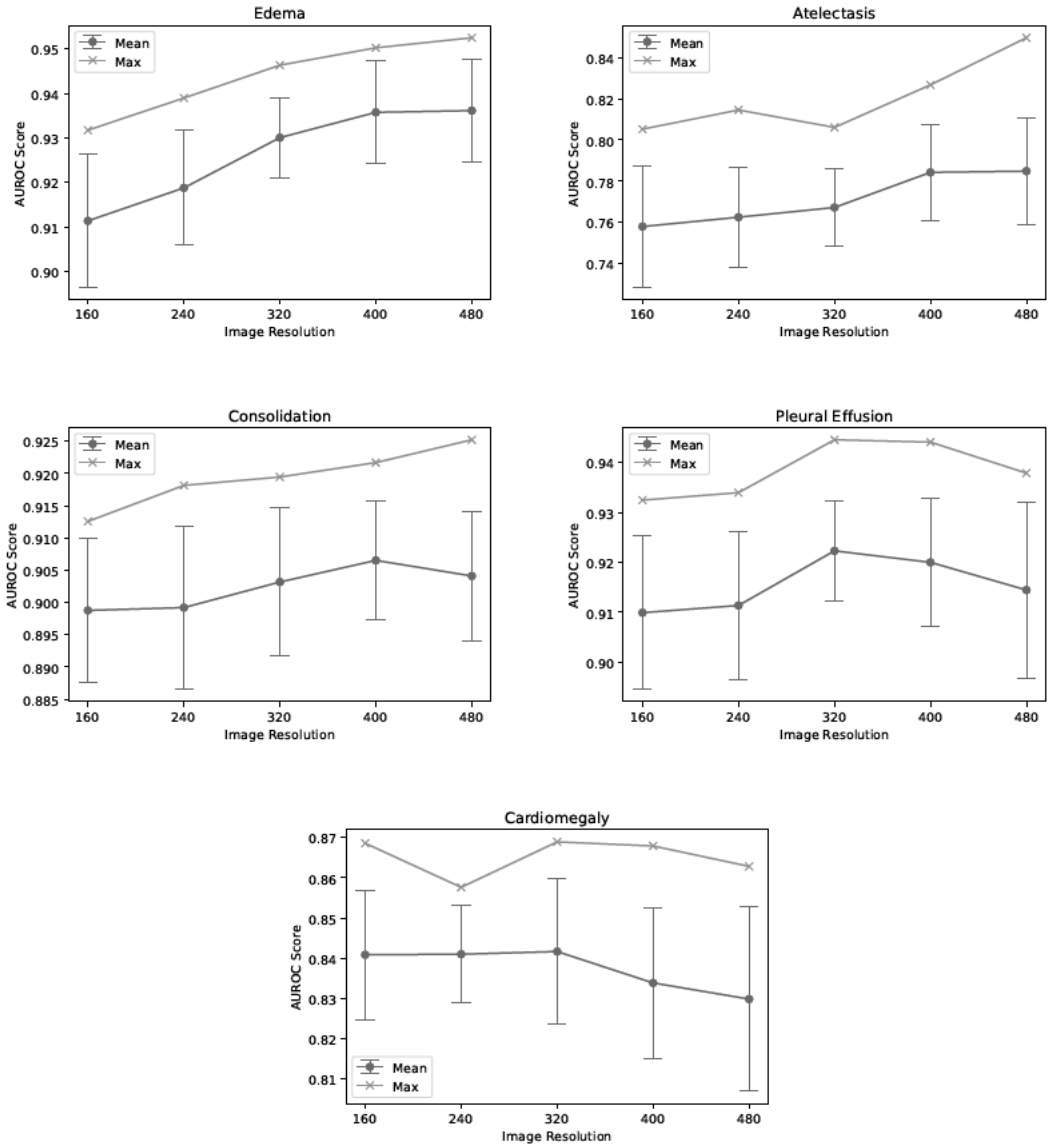


Figure 4.1: AUROC scores are grouped by image resolution and the mean, standard deviation and maximum per image resolution are reported. Three of the five competition tasks AUROC scores scale well with increasing image resolution. Performance differences of up to 0.04 could be measured. Two of the five competition tasks AUROC scores do not benefit from an image resolution above 320×320.

This results in the following equations for the ResNet and DenseNet model types

$$\Delta max(ResNet) = max(ResNet152) - max(ResNet50) \tag{4.3}$$

$$\Delta mean(ResNet) = mean(ResNet152) - mean(ResNet50) \tag{4.4}$$

$$\Delta max(DenseNet) = max(DenseNet161) - max(DenseNet121) \tag{4.5}$$

$$\Delta mean(DenseNet) = mean(DenseNet161) - mean(DenseNet121) \tag{4.6}$$

These metrics are visualized for each competition task per image resolution in figure 4.2. Blue marks show the score differences for the ResNet model type, while red marks symbolize the DenseNet model type. The circle symbol shows the mean, while the triangle symbolizes the max score differences. The green triangle shows the best overall performing model for each competition task. The 0 mark is visualized by a dashed line. Marks drawn above the dashed line therefore show that the bigger variant of a model type outperformed the smaller one and vice versa.

Firstly, the results of the best performing model, symbolized with a green triangle will be investigated. In only two cases, the larger model variant outperforms the smaller for the best performing models, namely Cardiomegaly and Atelectasis. In the former case, the performance difference is marginal. The best performing Cardiomegaly detection model was trained with the ResNet152 model variant and an input resolution of 320×320, outperforming its smaller ResNet50 variant with $\Delta = 0.0067$. But for the latter case, the performance difference is significant. For the Atelectasis class, the DenseNet161 model trained and evaluated with an input resolution of 480×480 outperformed all other models and its smaller DenseNet121 counterpart with a difference of $\Delta = 0.06$. For the detection of this class, this is an indication for the necessity of both scaling image resolution, model depth and width.

For the remaining three of the five competition tasks Edema, Consolidation and Pleural Effusion the smaller ResNet model variants were performing best. For the Edema class the best performing model was achieved with an image resolution of 480×480 with a marginal performance difference of $\Delta = -0.006$. For the Consolidation task, the same applies with a bigger performance difference of $\Delta = -0.016$, benefiting the most from the more shallow ResNet model variant. The Pleural Effusion class was best detected at an image resolution of 320×320 with a performance difference of $\Delta = -0.0095$.

Secondly, no clear trend can be observed between image resolution and scaling model depth and width for all of the competition tasks or model variants when looking at

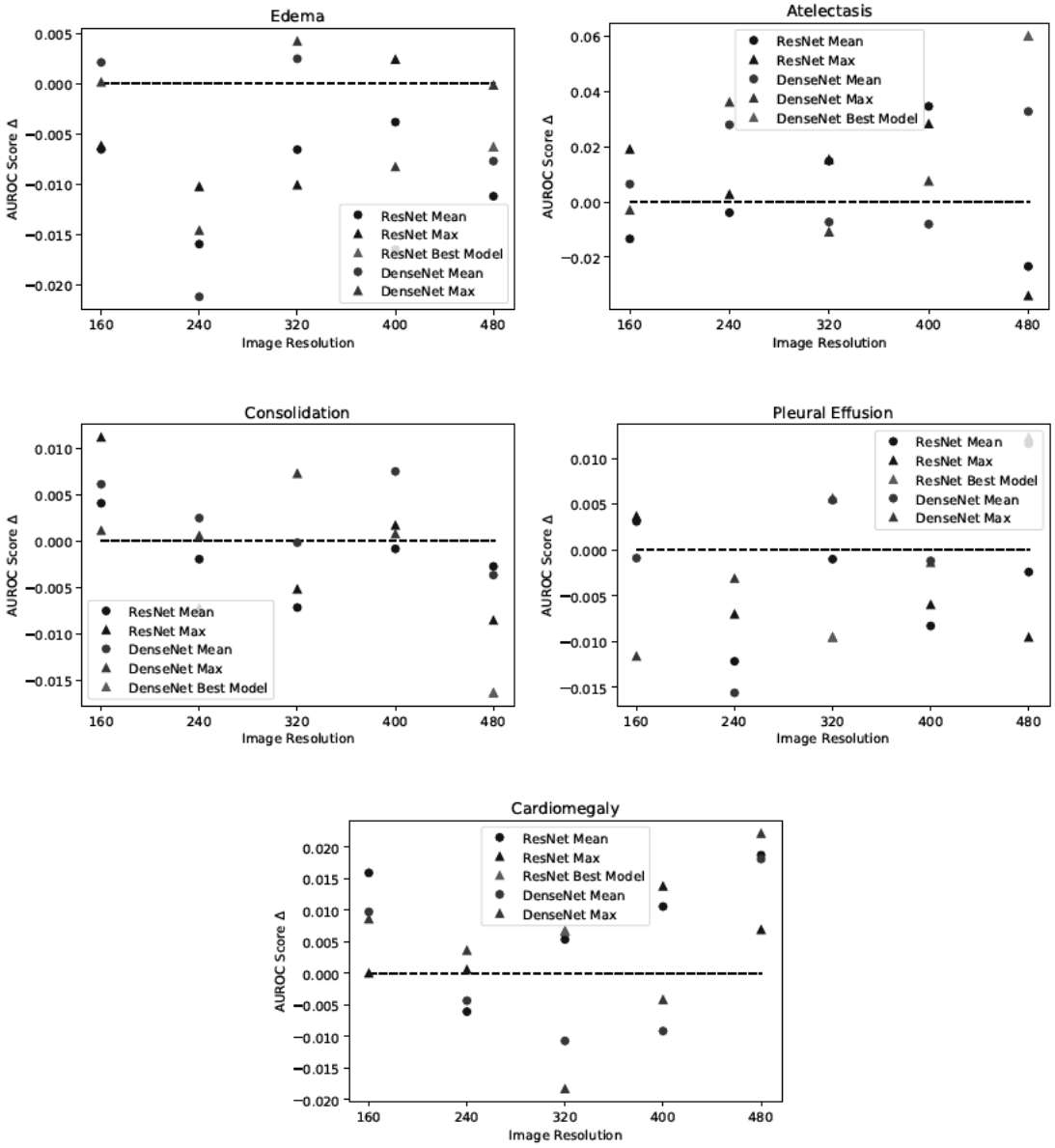


Figure 4.2: AUROC scores are grouped by image resolution and model architecture. The difference of mean and maximum between the two model types per image resolution is reported. Dot and triangle show the difference in mean and max of each architecture. Green triangle represents the best performing model for each evaluation task.

the mean and max marks. Even for classes like Edema and Atelectasis, which signaled greater detection rates in increased image resolution. No clear correlation between image resolution and model depth and width can be noticed. The prior task scales better with smaller model variants, while the latter does with larger models. The necessity for model scaling is therefore task dependent. Besides for the Atelectasis task, score differences in the model variants are generally marginal.

The impact of compound scaling via the EfficientNet architecture used as feature extractors can be seen in figure 4.3. These models generally performed worse than the overall average that can be seen in figure 4.1. The trends are similar. While the Edema and Atelectasis tasks show dependence on model and resolution scaling, tasks like Consolidation, Pleural Effusion decline sharply at an input resolution of 400×400 pixels. The Cardiomegaly class shows inverse scaling, where smaller image resolution and more shallow and narrow models outperform their bigger counterparts marginally.

4.2 Qualitative Evaluation

Gradient-weighted Class Activation Mapping (Grad-CAM) [39], uses the gradients of any classification prediction flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image responsible for the prediction. Convolutional layers naturally retain spatial information which is lost in fully-connected layers, so it is expected that the last convolutional layers have the best compromise between high-level semantics and detailed spatial information. The neurons in these layers look for semantic class-specific information in the image (say object parts). Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each neuron for a particular decision of interest.

Highlighting the important regions in a chest radiograph in order to assess model performance requires domain knowledge. In order to interpret the explanations generated by this method, knowledge of the way each pathology constitutes in a radiographic image is necessary. Optimally, this would be jointly assessed with a radiologist. Unfortunately, this is out of scope for this particular study.

This could be mitigated by leveraging bounding box annotations, which are not provided in the CheXpert dataset. The NIH Chest X-Ray14 [47] dataset however, offers a small number of images where hand labeled bounding boxes in addition to the pathology are

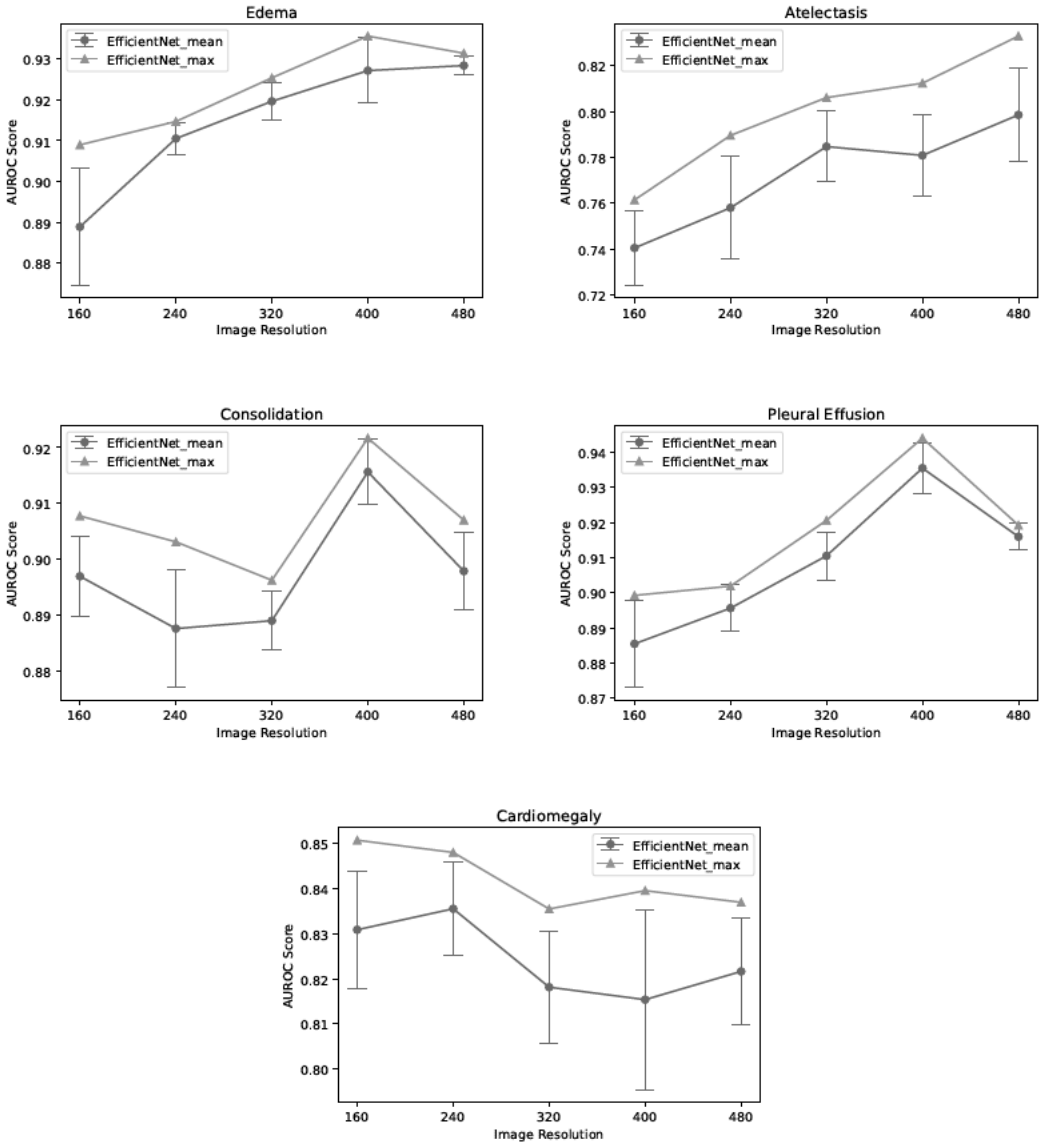


Figure 4.3: EfficientNet AUROC scores grouped by image resolution. The effects of compound scaling on the detection rate.

provided. These can be used as the ground truth to evaluate the disease localization performance. Given an image and a disease keyword, a board-certified radiologist identified only the corresponding disease instance in the image and annotated it with a bounding box.

The class labels of the NIH Chest X-Ray14 and the CheXpert dataset differ and there is only overlap in 3 of the 5 competition tasks that occur in the CheXpert dataset. These are namely Atelectasis, Cardiomegaly and Pleural Effusion, which is referred to as Effusion in the ChestX-ray8 dataset. Furthermore cross-dataset model performance can suffer from modality difference, caused by different X-ray equipment, various nationalities etc. [19]

This section therefore investigates the impact of image resolution on the localization performance of the globally labeled CheXpert trained models. This is enabled by comparing the bounding box annotated radiographs from the ChestX-ray8 to the coarse localization maps produced via the Grad-CAM method by inputting the same radiographs to the models. From each of the three overlapping classes, one image is chosen and used as input to the models. The images used can be seen in figure 4.4. The best performing models per task and image resolution based on the AUROC score are used.

(a) Pleural Effusion (b) Cardiomegaly (c) Atelectasis

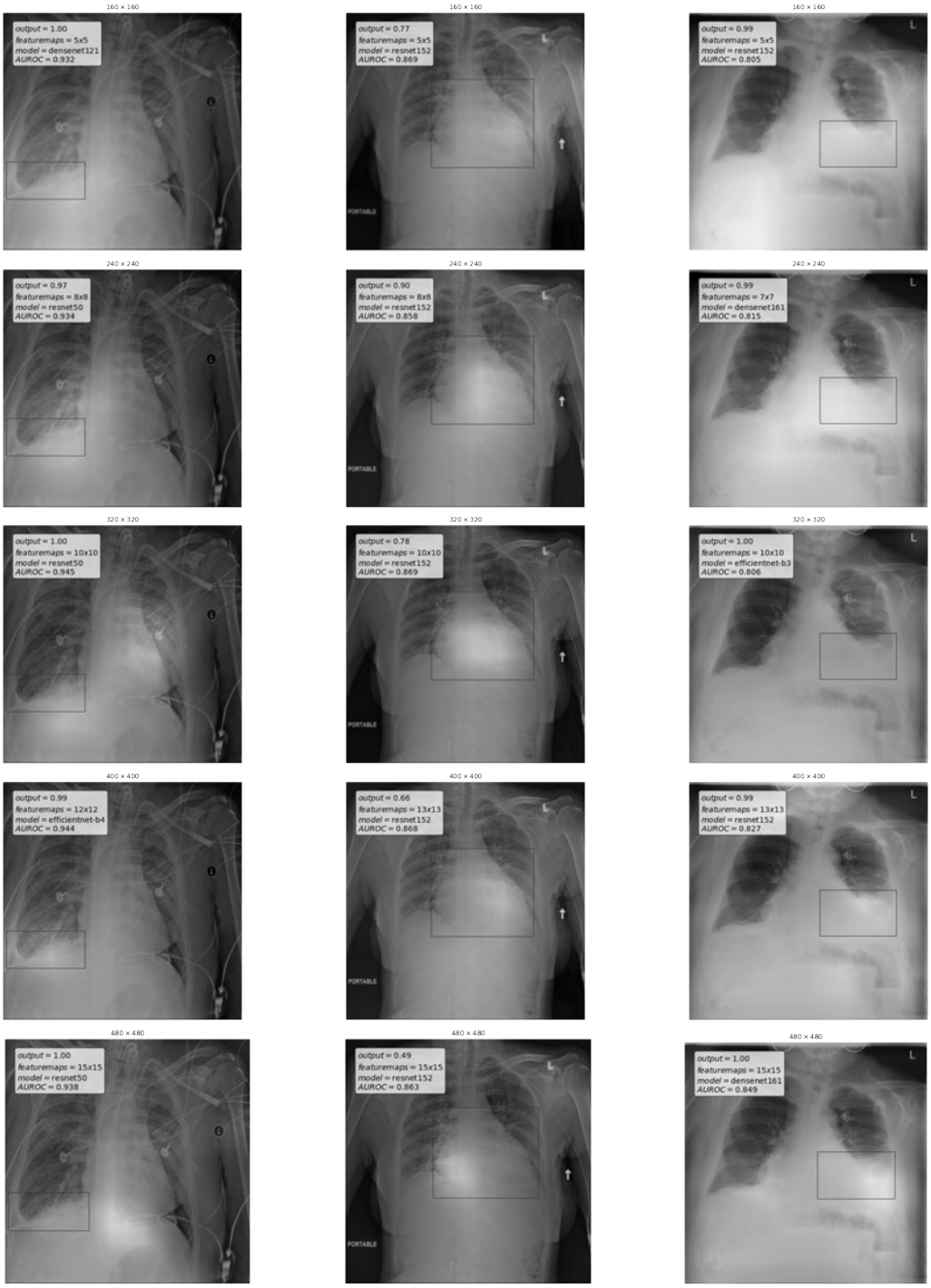
Figure 4.4: Three radiographic images from the ChestX-ray8 dataset labeled by a board certified radiologist. Different pathologies localized with bounding boxes. Images used for localization performance evaluation of the former trained models.

The granularity of the localization maps produced by the Grad-CAM method are based on the height and width of the output shape of the last convolutional layer for each model. The resolution of the maps therefore scale according to the input image resolution. These

range from 5×5 to 15×15. In order to retrieve a mask, which can be overlapped with the input image, an interpolation function is required. The bilinear function of the matplotlib library was leveraged.

The results can be seen in figure 4.5. Each column of the figure corresponds to a different pathology, while each row corresponds to the input image resolution. For each combination of input resolution and pathology, the best performing model based on the CheXpert validation AUROC was leveraged. The description box for each cell offers additional information such as output prediction probability from 0 to 1, the size of the feature maps of the final convolutional layer, the model architecture and the CheXpert validation AUROC.

The activated regions of the heatmap generally are more coarse for smaller input resolutions and more granular for higher input resolutions. The reason for this is the size of the feature maps, which are bound to the input image resolution. For smaller image resolutions 160×160 and 240×240 the highlighted regions seem to overlap with the bounding box ground truth, but are mostly larger and go beyond. Model outputs are close to 1 which signal high confidence in their correct predictions. For the 320×320 input resolutions, the highlighted regions mostly miss the bounding box for the Atelectasis and Pleural Effusion class, but outputs high values close to 1 which also signal high confidence. This could be the result of bias in the training dataset, where co-occurrences of a pathology and other visual indicators such as visual lines, tubes or wires are correlated. Another reason could be cross-dataset modality differences, such as different X-ray equipment, varying subject nationalities etc. For the Cardiomegaly class the highlighted area overlaps with most of the bounding box, but the output score is lower than with the former smaller input images. The best result for the Pleural Effusion class was achieved at an input resolution of 400×400, where the bounding box and highlighted area overlap almost perfectly and output score is also close to 1. For the Cardiomegaly and Atelectasis class, highlighted areas of the heatmap are inside the bounding box as well. For the Atelectasis class, the neck and head of the subject is also highlighted. Lastly, models trained with an input resolution of 480×480 start highlighting regions inside, but also a large region outside of the bounding box with a high confidence for the Pleural Effusion study. The smallest region of all input resolutions is highlighted inside the bounding box for the Cardiomegaly class. The model also outputted the smallest probability of pathology occurrence. For the Atelectasis class, the best overlap of heatmap and bounding box was achieved at this resolution.



(a) Pleural Effusion (b) Cardiomegaly (c) Atelectasis

Figure 4.5: Bilinear Interpolated Grad-CAM localization heatmaps overlapped with the input image. Bounding Box represents the ground truth of the pathology. Columns correspond to class and rows to resolution. Legend displays additional model information. output probability, feature-map size, model type and AUROC score on CheXpert validation set.

5 Summary & Outlook

In this thesis a broad overview of the current state of deep learning based automated chest radiograph interpretation was provided. Common datasets and their trade-off between quantity and quality were briefly introduced. A more thorough introduction to the CheXpert dataset [21] and the problem of ImageNet pre-training for medical applications was given to finally result in the research question of the importance of image resolution and model scaling in the detection rate of automated chest radiograph interpretation.

Extensive empirical studies were constructed. Three model architectures were leveraged in different variations, namely ResNet50, ResNet152 [14], DenseNet121, DenseNet161 [16] and five different EfficientNet [46] variations. The five competition tasks Edema, Atelectasis, Pleural Effusion, Cardiomegaly and Consolidation of the CheXpert dataset were used as input to the models. These were down-scaled from up to 4000×4000 pixels to 160×160, 240×240, 320×320, 400×400 and 480×480 respectively.

As a means to measure the detection rate of the models, the AUROC metric on the validation set was leveraged. Firstly, the general relationship between image resolution and detection rate was examined. It could be observed that the impact of image resolution on the detection rate is task dependent. The detection rate of Edema, Atelectasis and Consolidation scaled up with increased image resolution, while Pleural Effusion and Cardiomegaly reach the highest detection rate at 320×320 pixels. Lower input image resolutions than 320×320 seem to offer suboptimal results for all tasks except Cardiomegaly. Secondly, the role of model scaling in the detection rate gains was investigated. No clear trend could be observed between image resolution and scaling model depth and width across all competition tasks, when comparing larger model variants to their smaller counterparts. The detection rate of Edema was higher in the smaller model variants, while the contrary applies to the Atelectasis class. The detection rate of the Atelectasis class benefited the most from upscaling the DenseNet architecture with an AUROC score difference of $\Delta = 0.06$ on the best performing image resolution of 480×480. The necessity of model scaling is therefore also task dependent. Besides

for the Atelectasis task, score differences in the model variants for the remaining tasks were generally marginal. Additionally, no correlation between ImageNet accuracy and transfer accuracy could be observed. The AUROC scores did not scale according to the pretrained models ImageNet accuracy. ImageNet pre-training might not transfer well to this fine-grained task, as already mentioned in [25].

Additionally a brief qualitative evaluation was provided. By visualizing the gradients of predictions flowing into the final convolutional layer, coarse localization maps were produced via the Grad-CAM method. These saliency maps were compared to bounding box annotations from a board-certified radiologist to assess the model performance qualitatively on three studies in the aforementioned five image resolutions from the NIH Chest X-Ray14 [47] dataset. This dataset offers bounding box annotations, but only three of the five competition tasks overlap, namely Pleural Effusion, Cardiomegaly and Atelectasis. For each image resolution the best performing model based on the CheXpert validation set was evaluated. The activated regions of the heatmap generally are more coarse for smaller input resolutions and more granular for higher input resolutions. The highlighted regions in the smaller image resolutions of 160×160 and 240×240 have some overlap with the ground truth annotation, but are mostly larger and too coarse. Nonetheless, the representations seem to offer enough information for the model to correctly output high probabilities of pathology occurrence. The best alignment of ground truth and highlighted region for the Pleural Effusion study was achieved at an input resolution of 400×400. The same applies to the Atelectasis class at the highest investigated image resolution of 480×480. For the Cardiomegaly class, the very large size of the bounding box annotation causes even the coarser saliency maps of the lower image resolutions to align well.

In the following, we'd like to discuss the outlook for this field, in the form of ideas and concerns that have emerged while producing this work.

In hindsight, the quality and size of the CheXpert dataset is still questionable. In particular the size of the validation set consisting of 200 studies might not suffice to evaluate the quality of a model, even with the addition of the test set. The class distribution of the training set is also highly unbalanced, which is discussed in section 3.1. This is taken to the extreme for the Atelectasis class. One way of mitigating the issue would be to merge the MIMIC-CXR [23] and CheXpert dataset. Both use the same automated labeling approach on different databases, resulting in similar dataset structures and labels.

Unfortunately this was out of scope for this work, because access to the MIMIC-CXR dataset is restricted to credentialed and certified users of the physionet platform.

Another concern is the usage of ImageNet pretraining and transfer learning in medical imaging applications. Because of the domain gap between natural images and medical images, the benefits of using ImageNet pretraining for such applications are still uncertain. While it can accelerate model convergence, the end result might not be optimal. The creation of a general purpose medical imaging dataset like ImageNet for the medical domain could be beneficial. Transfer learning on models trained on this general purpose medical imaging dataset might offer accelerated model convergence, more confidence in an optimal end result and the option to train on even smaller, high quality medical datasets. The problem of producing high quality labeled datasets is still one major issue in this domain that could be further mitigated by this.

Working jointly with a medical professional could have brought more insight as to why certain pathologies seemed to be more detectable from the models trained on higher image resolutions than others. We assume the reason lies in the different ways the anomalies present themselves in chest radiographic images. Some pathologies might only manifest as very subtle anomalies that are convoluted by down-scaling the pixels of an image.

Bibliography

[1] : *torchvision.models — Torchvision 0.10.0 documentation.* – URL https://pytorch.org/vision/stable/models.html. – Zugriffsdatum: 2021-07-02

[2] Chizari, Nima: Deep-Learning Pipeline zur Erkennung von Anomalien in Thorax-Roentgenbildern.

[3] Chizari, Nima: Thoughts on Image Resolution and the Impact on the Detection Rate of Pathologies in Chest Radiographs based on the CheXpert Dataset using ImageNet Pretrained Deep Learning Models. (2021), S. 10

[4] CSTI: *CSTI Creative Space for Technical Innovations.* – URL https://csti.haw-hamburg.de/. – Zugriffsdatum: 2020-08-13

[5] Demner-Fushman, Dina ; Kohli, Marc D. ; Rosenman, Marc B. ; Shooshan, Sonya E. ; Rodriguez, Laritza ; Antani, Sameer ; Thoma, George R. ; McDonald, Clement J.: Preparing a collection of radiology examinations for distribution and retrieval. In: *Journal of the American Medical Informatics Association : JAMIA* 23 (2016), März, Nr. 2, S. 304–310. – URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5009925/. – Zugriffsdatum: 2021-06-30. – ISSN 1067-5027

[6] Devos, Andy. K. ; Huffel, Sabine van ; Simonetti, Arjan W. ; Graaf, Marinette van der ; Heerschap, Arend ; Buydens, Lutgarde M. C.: Chapter 11 - Classification of Brain Tumours by Pattern Recognition of Magnetic Resonance Imaging and Spectroscopic Data. In: Taktak, Azzam F. G. (Hrsg.) ; Fisher, Anthony C. (Hrsg.): *Outcome Prediction in Cancer.* Amsterdam : Elsevier, Januar 2007, S. 285–318. – URL https://www.sciencedirect.com/science/article/pii/B9780444528551500131. – Zugriffsdatum: 2021-08-18. – ISBN 978-0-444-52855-1

[7] Esteva, Andre ; Kuprel, Brett ; Novoa, Roberto A. ; Ko, Justin ; Swetter, Susan M. ; Blau, Helen M. ; Thrun, Sebastian: Dermatologist-level classification

of skin cancer with deep neural networks. In: *Nature* 542 (2017), Februar, Nr. 7639, S. 115–118. – URL https://www.nature.com/articles/nature21056. – Zugriffsdatum: 2020-04-16. – ISSN 1476-4687

[8] Garbin, Christian ; Rajpurkar, Pranav ; Irvin, Jeremy ; Lungren, Matthew P. ; Marques, Oge: Structured dataset documentation: a datasheet for CheXpert. In: *arXiv:2105.03020 [cs, eess]* (2021), Mai. – URL http://arxiv.org/abs/2105.03020. – Zugriffsdatum: 2021-08-25. – arXiv: 2105.03020

[9] Grewal, M. ; Srivastava, M. M. ; Kumar, P. ; Varadarajan, S.: RADnet: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, April 2018, S. 281–284. – ISSN: 1945-8452

[10] Gulshan, Varun ; Peng, Lily ; Coram, Marc ; Stumpe, Martin C. ; Wu, Derek ; Narayanaswamy, Arunachalam ; Venugopalan, Subhashini ; Widner, Kasumi ; Madams, Tom ; Cuadros, Jorge ; Kim, Ramasamy ; Raman, Rajiv ; Nelson, Philip C. ; Mega, Jessica L. ; Webster, Dale R.: Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. In: *JAMA* 316 (2016), Dezember, Nr. 22, S. 2402–2410. – URL https://jamanetwork.com/journals/jama/fullarticle/2588763. – Zugriffsdatum: 2020-04-16. – ISSN 0098-7484

[11] Hannun, Awni Y. ; Rajpurkar, Pranav ; Haghpanahi, Masoumeh ; Tison, Geoffrey H. ; Bourn, Codie ; Turakhia, Mintu P. ; Ng, Andrew Y.: Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. In: *Nature Medicine* 25 (2019), Januar, Nr. 1, S. 65–69. – URL https://www.nature.com/articles/s41591-018-0268-3. – Zugriffsdatum: 2020-04-16. – ISSN 1546-170X

[12] Hansell, David M. ; Bankier, Alexander A. ; MacMahon, Heber ; McLoud, Theresa C. ; Müller, Nestor L. ; Remy, Jacques: Fleischner Society: glossary of terms for thoracic imaging. In: *Radiology* 246 (2008), März, Nr. 3, S. 697–722. – ISSN 1527-1315

[13] He, Kaiming ; Zhang, Xiangyu ; Ren, Shaoqing ; Sun, Jian: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: *arXiv:1502.01852 [cs]* (2015), Februar. – URL http://arxiv.org/abs/1502.01852. – Zugriffsdatum: 2020-11-26. – arXiv: 1502.01852

[14] He, Kaiming ; Zhang, Xiangyu ; Ren, Shaoqing ; Sun, Jian: Deep Residual Learning for Image Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Juni 2016, S. 770–778. – ISSN: 1063-6919

[15] Howard, Andrew G. ; Zhu, Menglong ; Chen, Bo ; Kalenichenko, Dmitry ; Wang, Weijun ; Weyand, Tobias ; Andreetto, M. ; Adam, Hartwig: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. In: *ArXiv* (2017)

[16] Huang, Gao ; Liu, Zhuang ; Van Der Maaten, Laurens ; Weinberger, Kilian Q.: Densely Connected Convolutional Networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Juli 2017, S. 2261–2269. – ISSN: 1063-6919

[17] Huang, H. K.: Medical imaging. In: *Encyclopedia of Computer Science*. GBR : John Wiley and Sons Ltd., Januar 2003, S. 1118–1130. – ISBN 978-0-470-86412-8

[18] Huang, Yanping ; Cheng, Youlong ; Bapna, Ankur ; Firat, Orhan ; Chen, Dehao ; Chen, Mia ; Lee, HyoukJoong ; Ngiam, Jiquan ; Le, Quoc V. ; Wu, Yonghui ; Chen, zhifeng: GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism. In: Wallach, H. (Hrsg.) ; Larochelle, H. (Hrsg.) ; Beygelzimer, A. (Hrsg.) ; Alché-Buc, F. d. (Hrsg.) ; Fox, E. (Hrsg.) ; Garnett, R. (Hrsg.): *Advances in Neural Information Processing Systems* Bd. 32, Curran Associates, Inc., 2019. – URL https://proceedings.neurips.cc/paper/2019/file/093f65e080a295f8076b1c5722a46aa2-Paper.pdf

[19] Hwang, Sangheum ; Kim, Hyo-Eun ; M.d, Jihoon J. ; Kim, Hee-Jin: A novel approach for tuberculosis screening based on deep convolutional neural networks. In: *Medical Imaging 2016: Computer-Aided Diagnosis* Bd. 9785, SPIE, März 2016, S. 750–757. – URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9785/97852W/A-novel-approach-for-tuberculosis-screening-based-on-deep-convolutional/.full. – Zugriffsdatum: 2021-09-24

[20] Ioffe, Sergey ; Szegedy, Christian: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. Lille, France : JMLR.org, Juli 2015 (ICML'15), S. 448–456

[21] Irvin, Jeremy ; Rajpurkar, Pranav ; Ko, Michael ; Yu, Yifan ; Ciurea-Ilcus, Silviana ; Chute, Chris ; Marklund, Henrik ; Haghgoo, Behzad ; Ball, Robyn ; Shpanskaya, Katie ; Seekins, Jayne ; Mong, David ; Halabi, Safwan ; Sandberg, Jesse ; Jones, Ricky ; Larson, David ; Langlotz, Curtis ; Patel, Bhavik ; Lungren, Matthew ; Ng, Andrew: CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (2019), Juli, S. 590–597

[22] Jaeger, Stefan ; Candemir, Sema ; Antani, Sameer ; Wáng, Yì-Xiáng J. ; Lu, Pu-Xuan ; Thoma, George: Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. In: *Quantitative Imaging in Medicine and Surgery* 4 (2014), Dezember, Nr. 6, S. 475–477. – ISSN 2223-4292

[23] Johnson, Alistair E. W. ; Pollard, Tom J. ; Greenbaum, Nathaniel R. ; Lungren, Matthew P. ; Deng, Chih-ying ; Peng, Yifan ; Lu, Zhiyong ; Mark, Roger G. ; Berkowitz, Seth J. ; Horng, Steven: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. In: *arXiv:1901.07042 [cs, eess]* (2019), November. – URL http://arxiv.org/abs/1901.07042. – Zugriffsdatum: 2020-11-26. – arXiv: 1901.07042

[24] Kingma, Diederik ; Ba, Jimmy: Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations* (2014), Dezember

[25] Kornblith, Simon ; Shlens, Jonathon ; Le, Quoc V.: Do Better ImageNet Models Transfer Better? In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA : IEEE, Juni 2019, S. 2656–2666. – URL https://ieeexplore.ieee.org/document/8954384/. – Zugriffsdatum: 2021-08-04. – ISBN 978-1-72813-293-8

[26] Litjens, Geert ; Kooi, Thijs ; Bejnordi, Babak E. ; Setio, Arnaud Arindra A. ; Ciompi, Francesco ; Ghafoorian, Mohsen ; Laak, Jeroen A. W. M. van der ; Ginneken, Bram van ; Sánchez, Clara I.: A survey on deep learning in medical image analysis. In: *Medical Image Analysis* 42 (2017), Dezember, S. 60–88. – URL https://www.sciencedirect.com/science/article/pii/S1361841517301135. – Zugriffsdatum: 2021-06-11. – ISSN 1361-8415

[27] Long, Jonathan ; Shelhamer, Evan ; Darrell, Trevor: Fully convolutional networks for semantic segmentation. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Juni 2015, S. 3431–3440. – ISSN: 1063-6919

[28] Melas-Kyriazi, Luke: *lukemelas/EfficientNet-PyTorch*. Juli 2021. – URL https://github.com/lukemelas/EfficientNet-PyTorch. – Zugriffsdatum: 2021-07-20. – original-date: 2019-05-30T05:24:11Z

[29] Metz, Charles E.: Basic principles of ROC analysis. In: *Seminars in Nuclear Medicine* 8 (1978), Oktober, Nr. 4, S. 283–298. – URL https://www.sciencedirect.com/science/article/pii/S0001299878800142. – Zugriffsdatum: 2021-08-18. – ISSN 0001-2998

[30] Pham, Hieu H. ; Le, Tung T. ; Ngo, Dat T. ; Tran, Dat Q. ; Nguyen, Ha Q.: Interpreting Chest X-rays via CNNs that Exploit Hierarchical Disease Dependencies and Uncertainty Labels, URL https://openreview.net/forum?id=4o1GLIIHlh. – Zugriffsdatum: 2021-11-04, Januar 2020

[31] Phung, V.H. ; Rhee, E.J.: A deep learning approach for classification of cloud image patches on small datasets. In: *Journal of Information and Communication Convergence Engineering* 16 (2018), Januar, S. 173–178

[32] Qin, Chunli ; Yao, Demin ; Shi, Yonghong ; Song, Zhijian: Computer-aided detection in chest radiography based on artificial intelligence: a survey. In: *BioMedical Engineering OnLine* 17 (2018), Dezember, Nr. 1, S. 113. – URL https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-018-0544-y. – Zugriffsdatum: 2020-04-15. – ISSN 1475-925X

[33] Raghu, Maithra ; Zhang, Chiyuan ; Kleinberg, Jon ; Bengio, Samy: Transfusion: Understanding Transfer Learning for Medical Imaging. In: *arXiv:1902.07208 [cs, stat]* (2019), Oktober. – URL http://arxiv.org/abs/1902.07208. – Zugriffsdatum: 2020-11-10. – arXiv: 1902.07208 version: 3

[34] Ronneberger, Olaf ; Fischer, Philipp ; Brox, Thomas: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, Nassir (Hrsg.) ; Hornegger, Joachim (Hrsg.) ; Wells, William M. (Hrsg.) ; Frangi, Alejandro F. (Hrsg.): *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* Bd. 9351. Cham : Springer International Publishing, 2015, S. 234–241. – ISBN 978-3-319-24573-7 978-3-319-24574-4

[35] Russakovsky, Olga ; Deng, Jia ; Su, Hao ; Krause, Jonathan ; Satheesh, Sanjeev ; Ma, Sean ; Huang, Zhiheng ; Karpathy, Andrej ; Khosla, Aditya ; Bernstein, Michael ; Berg, Alexander C. ; Fei-Fei, Li: ImageNet Large Scale Visual

Recognition Challenge. In: *International Journal of Computer Vision* 115 (2015), Dezember, Nr. 3, S. 211–252. – URL https://doi.org/10.1007/s11263-015-0816-y. – Zugriffsdatum: 2020-08-19. – ISSN 1573-1405

[36] Sabottke, Carl ; Spieler, Bradley: The Effect of Image Resolution on Deep Learning in Radiography. In: *Radiology: Artificial Intelligence* 2 (2020), Januar, S. e190015

[37] Sandler, Mark ; Howard, Andrew ; Zhu, Menglong ; Zhmoginov, Andrey ; Chen, Liang-Chieh: MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Juni 2018, S. 4510–4520. – ISSN: 2575-7075

[38] scikit, learn: *sklearn.metrics.roc_ auc_ score — scikit-learn 0.23.2 documentation.* – URL https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html?highlight=roc#sklearn.metrics.roc_auc_score. – Zugriffsdatum: 2020-11-27

[39] Selvaraju, Ramprasaath R. ; Cogswell, Michael ; Das, Abhishek ; Vedantam, Ramakrishna ; Parikh, Devi ; Batra, Dhruv: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, Oktober 2017, S. 618–626. – ISSN: 2380-7504

[40] Shiraishi, J. ; Katsuragawa, S. ; Ikezoe, J. ; Matsumoto, T. ; Kobayashi, T. ; Komatsu, K. ; Matsui, M. ; Fujita, H. ; Kodera, Y. ; Doi, K.: Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. In: *AJR. American journal of roentgenology* 174 (2000), Januar, Nr. 1, S. 71–74. – ISSN 0361-803X

[41] Smith, Leslie N. ; Topin, Nicholay: Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. In: *arXiv:1708.07120 [cs, stat]* (2018), Mai. – URL http://arxiv.org/abs/1708.07120. – Zugriffsdatum: 2020-12-03. – arXiv: 1708.07120

[42] Sokolova, Marina ; Lapalme, Guy: A systematic analysis of performance measures for classification tasks. In: *Information Processing & Management* 45 (2009), Juli, Nr. 4, S. 427–437. – URL http://www.sciencedirect.com/science/

article/pii/S0306457309000259. – Zugriffsdatum: 2020-04-17. – ISSN 0306-4573

[43] Srivastava, Nitish ; Hinton, Geoffrey ; Krizhevsky, Alex ; Sutskever, Ilya ; Salakhutdinov, Ruslan: Dropout: a simple way to prevent neural networks from overfitting. In: *The Journal of Machine Learning Research* 15 (2014), Januar, Nr. 1, S. 1929–1958. – ISSN 1532-4435

[44] Szegedy, Christian ; Vanhoucke, V. ; Ioffe, S. ; Shlens, Jonathon ; Wojna, Z.: Rethinking the Inception Architecture for Computer Vision. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016

[45] Tan, Mingxing ; Chen, Bo ; Pang, Ruoming ; Vasudevan, Vijay ; Le, Quoc V.: MnasNet: Platform-Aware Neural Architecture Search for Mobile. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019

[46] Tan, Mingxing ; Le, Quoc: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: *International Conference on Machine Learning*, PMLR, Mai 2019, S. 6105–6114. – URL http://proceedings.mlr.press/v97/tan19a.html. – Zugriffsdatum: 2021-07-05. – ISSN: 2640-3498

[47] Wang, Xiaosong ; Peng, Yifan ; Lu, Le ; Lu, Zhiyong ; Bagheri, Mohammadhadi ; Summers, Ronald M.: ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Juli 2017, S. 3462–3471. – ISSN: 1063-6919

[48] Wingate, Ryan: *Binary Classifiers, ROC Curve, and the AUC.* Juli 2018. – URL https://ryanwingate.com/statistics/binary-classifiers/binary-classifiers/. – Zugriffsdatum: 2020-04-17

[49] Yao, Li ; Prosky, Jordan ; Poblenz, Eric ; Covington, Ben ; Lyman, Kevin: Weakly Supervised Medical Diagnosis and Localization from Multiple Resolutions. In: *arXiv:1803.07703 [cs]* (2018), März. – URL http://arxiv.org/abs/1803.07703. – Zugriffsdatum: 2020-04-25. – arXiv: 1803.07703

[50] Yosinski, Jason ; Clune, Jeff ; Bengio, Yoshua ; Lipson, Hod: How transferable are features in deep neural networks? In: *arXiv:1411.1792 [cs]* (2014), November. – URL http://arxiv.org/abs/1411.1792. – Zugriffsdatum: 2020-11-10. – arXiv: 1411.1792

[51] Yuan, Zhuoning ; Yan, Yan ; Sonka, Milan ; Yang, Tianbao: Large-Scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification, URL https://openaccess.thecvf.com/content/ICCV2021/html/Yuan_Large-Scale_Robust_Deep_AUC_Maximization_A_New_Surrogate_Loss_and_ICCV_2021_paper.html. – Zugriffsdatum: 2021-11-04, 2021, S. 3040–3049

[52] Zagoruyko, Sergey ; Komodakis, Nikos: Wide Residual Networks. In: *Procedings of the British Machine Vision Conference 2016*. York, UK : British Machine Vision Association, 2016, S. 87.1–87.12. – URL http://www.bmva.org/bmvc/2016/papers/paper087/index.html. – Zugriffsdatum: 2021-07-05. – ISBN 978-1-901725-59-9

[53] Zoph, Barret ; Vasudevan, Vijay ; Shlens, Jonathon ; Le, Quoc V.: Learning Transferable Architectures for Scalable Image Recognition. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Juni 2018, S. 8697–8710. – ISSN: 2575-7075

Erklärung zur selbstständigen Bearbeitung einer Abschlussarbeit

Gemäß der Allgemeinen Prüfungs- und Studienordnung ist zusammen mit der Abschlussarbeit eine schriftliche Erklärung abzugeben, in der der Studierende bestätigt, dass die Abschlussarbeit — bei einer Gruppenarbeit die entsprechend gekennzeichneten Teile der Arbeit [(§ 18 Abs. 1 APSO-TI-BM bzw. § 21 Abs. 1 APSO-INGI)] — ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt wurden. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich zu machen.

Quelle: § 16 Abs. 5 APSO-TI-BM bzw. § 15 Abs. 6 APSO-INGI

Erklärung zur selbstständigen Bearbeitung der Arbeit

Hiermit versichere ich,

Name: _____

Vorname: _____

dass ich die vorliegende Masterarbeit – bzw. bei einer Gruppenarbeit die entsprechend gekennzeichneten Teile der Arbeit – mit dem Thema:

Die Auswirkungen der Bildauflösung und Modellskalierung auf Deep Learning basierte automatisierte Röntgen-Thorax Interpretation

ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

Ort Datum Unterschrift im Original