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

Nima Chizari
Department für Informatik
Hochschule für Angewandte Wissenschaften Hamburg
Berliner Tor 5, 20099 Hamburg
Nima.Chizari@haw-hamburg.de

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 [7] dataset. [7] 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. The radiologists were only able to outperform the model on one pathology, namely Atelectasis. The authors used downscaled versions (320x320) of the chest screenings as input to their model. This paper empirically examines the impact of 3 different image resolutions on the detection rate Area Under Receiver Operating Characteristic Curve (ROCAUC) of the pathology Atelectasis using ImageNet pretrained models. The results hint at the potential of higher detection rates when both scaling model depth/parameters and image resolution of the input images.

1 Introduction

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. [7]

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 [7] and MIMIC-CXR [8] datasets. Both use the same automated labeling approach to extract observations from free text radiology reports.

[7] have shown that it is possible to train a deep learning model successfully on their dataset. They compared the detection rate of their model with the detection rate of 3 individual radiologists on a test set of 500 studies. In 3 of 5 pathologies, the model had a higher detection rate than all 3 radiologists. One radiologist outperformed the model on the pathology *Consolidation*, while all 3 radiologists outperformed the model on the pathology *Atelectasis*.

Although the chest screenings are provided in a resolution of up to 4000x4000 pixels, [7] used downscaled versions in a resolution of 320x320 as inputs to the model, without providing reasoning for their choice. This paper empirically examines the impact of 3 different image resolutions

(160x160, 320x320, 480x480) on the detection rate (receiver operating characteristic curve AUC) of the pathology Atelectasis using pretrained models.

This paper is structured as follows: in Section 2, related work on another chest radiograph dataset is summarized. In Section 3, the CheXpert dataset, which will be used for the experiments is presented and analyzed. Section 4 introduces transfer learning briefly and discusses the use of ImageNet pretrained models for medical imaging. In Section 5 lastly, the main experiments of this paper, including preprocessing, model architecture, training procedure and the results are presented.

2 Related Work

[13] also examined variations of convolutional neural network (CNN) performance for multiple chest radiograph diagnoses and image resolutions. They used the publicly available National Institutes of Health chest radiograph dataset (ChestX-ray8) [17] comprising 112.120 chest radiographic images from 30.805 patients. The network architectures examined included ResNet34 [5] and DenseNet121 [6]. Image resolutions ranging from 32x32 to 600x600 pixels were investigated.

For this dataset and the chosen architectures, maximum AUCs were achieved at image resolutions between 256x256 and 448x448 pixels for binary decision networks targeting emphysema, cardiomegaly, hernias, edema, effusions, atelectasis, masses, and nodules. Different diagnoses 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).

They further concluded that increasing image resolution for CNN training often has a trade-off with the maximum possible batch size, yet optimal selection of image resolution has the potential for further increasing neural network performance for various radiology-based machine learning tasks. Furthermore, identifying diagnosis-specific tasks that require relatively higher image resolution can potentially provide insight into the relative difficulty of identifying different radiology findings.

3 CheXpert Dataset Analysis

The CheXpert Dataset [7] 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, [7] decided on the 14 observations based on the prevalence in the reports and clinical relevance, conforming to the Fleischner Society's recommended glossary [3] 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.

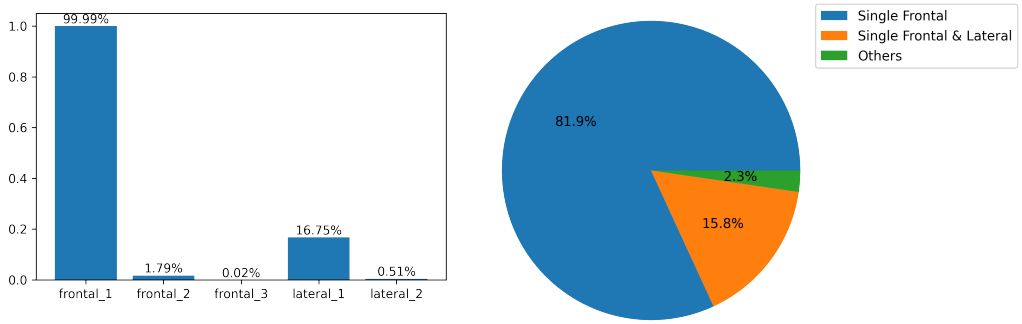
In the test and validation set however, [7] focuses on the evaluation of 5 observations which are called the competition tasks. These are selected based on clinical importance and prevalence: (a) Atelectasis, (b) Cardiomegaly, (c) Consolidation, (d) Edema, and (e) Pleural Effusion. The validation set consists of 200 studies on which the consensus of three radiologist annotations serves as ground truth.

The training and validation set 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.

As can be seen in 1a up to 3 frontal and 2 lateral views are available per study. Almost all studies include a frontal view x-ray image. The second highest occurring view is from a lateral perspective of the patient. In very few cases, additional frontal and lateral images are provided.

The constellation in which these views are provided can be seen in 1b. 81.9% of all studies consist of a single frontal x-ray image, followed by 15.8% of studies, which also include a single lateral image. The remaining 2.3% of studies are comprised of differing combinations. The highest amount of concurrent images provided in a study is 3.

In [7], the observations of the training set are reported in figure 1. [7] use the classification for each mention of observations to arrive at a final label for 14 observations that consist of 12 pathologies as



(a) Occurence of each view in studies

(b) studies grouped by view Constellations

Figure 1: Analysis of the available views per study. **Left:** Occurrences of the different radiographic views in each study. Up to 3 images per study are provided. These constellations can consist of 3 frontal and 2 lateral views. **Right:** Studies grouped by constellations. 81.9% of all studies provide only a single frontal radiographic image, followed by 15.8% of studies, where a single frontal and lateral radiographic image is provided. The remaining 2.3% of studies are made up of other constellations of views.

well as the "Support Devices" and "No Finding" observations. Observations with at least one mention that is positively classified in the report is assigned a positive (1) label. An observation is assigned an uncertain (u) label if it has no positively classified mentions and at least one uncertain mention, and a negative label if there is at least one negatively classified mention. They assign "blank" if there is no mention of an observation. The "No Finding" observation is assigned a positive label (1) if there is no pathology classified as positive or uncertain.

In the reported table 1, all unmentioned observations are assigned the negative classification. Assuming an unmentioned observation in the report conveys the absence of it could lead to bias and causes an imbalance in the distribution of classifications. The true distribution of classes can be seen in table 2. Most of the observations can be balanced by mapping the uncertain annotations to the less infrequent class or by mapping the"No Finding" class to the negative class.

Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16627 (8.86)	0 (0.0)	171014 (91.14)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)	168473 (89.78)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.3)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Pneumonia	4576 (2.44)	15658 (8.34)	167407 (89.22)
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.3)	1771 (0.94)	183429 (97.76)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Support Devices	105831 (56.4)	898 (0.48)	80912 (43.12)

Table 1: The reported class distribution of the CheXpert training set for the 14 labeled observations in [7]. All unmentioned observations are assigned the negative classifcation.

Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16627 (100.0)	0 (0.0)	0 (0.0)
Enlarged Cardiomediastinum	9020 (26.37)	10148 (29.67)	15032 (43.95)
Cardiomegaly	23002 (61.65)	6597 (17.68)	7712 (20.67)
Lung Opacity	92669 (90.89)	4341 (4.26)	4949 (4.85)
Lung Lesion	6856 (79.0)	1071 (12.34)	752 (8.66)
Edema	48905 (64.23)	11571 (15.2)	15667 (20.58)
Consolidation	12730 (22.77)	23976 (42.88)	19203 (34.35)
Pneumonia	4576 (20.74)	15658 (70.96)	1833 (8.31)
Atelectasis	29333 (49.12)	29377 (49.19)	1007 (1.69)
Pneumothorax	17313 (25.85)	2663 (3.98)	47006 (70.18)
Pleural Effusion	75696 (68.85)	9419 (8.57)	24832 (22.59)
Pleural Other	2441 (55.25)	1771 (40.09)	206 (4.66)
Fracture	7270 (74.93)	484 (4.99)	1949 (20.09)
Support Devices	105831 (94.5)	898 (0.8)	5257 (4.69)

Table 2: The actual class distribution of the CheXpert training set [7] without mapping unmentioned observations to the negative class.

4 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 activations used. These properties help model convergence. [4] Secondly there is *transfer learning*. In transfer learning, a base network is trained on a base dataset and task, and afterwards the learned features are repurposed or transfered to a second target network to be trained on a target dataset and task. [18]

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, then 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. [18]

Most base networks are trained on natural image datasets, usually *ImageNet* [12]. The human-annotated dataset comprises approximately 1 million images and 1.000 object classes. The classes range from animals (cats, dogs, etc.) to fruits (oranges, apples, etc.). The base task therefore consists of classifying these natural objects in the image. The transferability of parameters of such base network for use in a medical imaging interpretation model remains an open research question.

[18] 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.

[11] on the other hand, specifically investigated the use of ImageNet pretraining 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 [18], they also find that meaningful feature reuse is concentrated at the lowest layers.

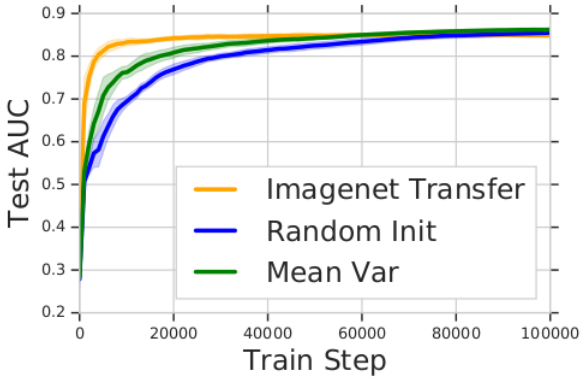


Figure 2: Convergence speed on the CheXpert consolidation pathology data and Resnet50 model architecture while using different parameter initialization methods. The figure compares ImageNet transfer learning and the Mean Var initialization scheme to random initialization. [18]

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, 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.

Contrary to their claims, using ImageNet pretraining offers faster convergence than the other schemes, while causing no overhead. These pretrained parameters are widely available and can be downloaded for most model architectures.

5 Experiments

This section focuses on empirical experiments regarding the impact of image resolution on the detection rate of one of the competition tasks in the CheXpert [7] dataset using ImageNet pretrained models. The pathology focused on is Atelectasis, because it is one of the few pathologies where the CheXpert model [7] did not achieve a better metric score than the radiologists. For each image resolution, multiple models will be trained using only the studies where the Atelectasis observation is present. The same performance metric (AUROC/ROC curve) and datasplit (Train/Valid) will be used to ensure comparability.

5.1 Image Resolution & Preprocessing

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 sizes 160x160, 320x320 and 480x480 using Lanczos algorithm implementation of the Python Image Library (PIL). The result can be seen in figure 3.

As can be seen in table 2, the Atelectasis observation is lacking negative studies. Therefore the studies of the *No Finding* observation are mapped to the negative Atelectasis class. [7] achieved the best result when mapping the uncertain annotations as positive for this pathology. Therefore this strategy

(a) 160x160 (b) 320x320 (c) 480x480

Figure 3: Frontal radiographic image from the validation set of CheXpert [7] in three 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.

was adopted for this approach as well. This results in a class distribution of 58,710 (77.28%) positive and 17,265 (22.72%) negative classifications.

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.

5.2 Model Architecture

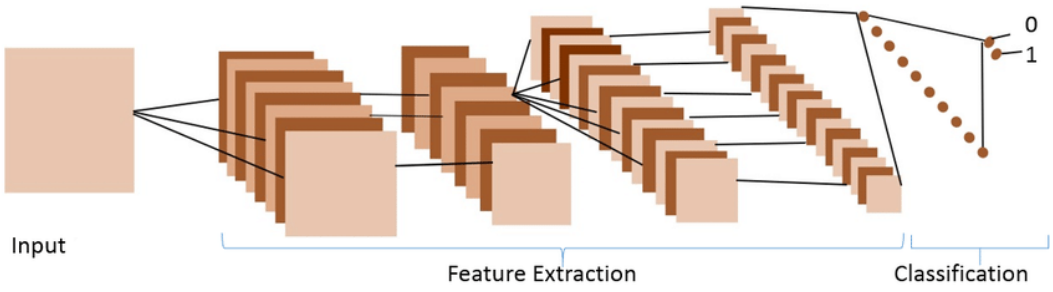


Figure 4: A CNN is composed of two basic parts of feature extraction and classification. Feature extraction includes several convolution layers followed by max-pooling and an activation function. The classifier usually consists of fully connected layers. [9]

The model can be broken down into 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

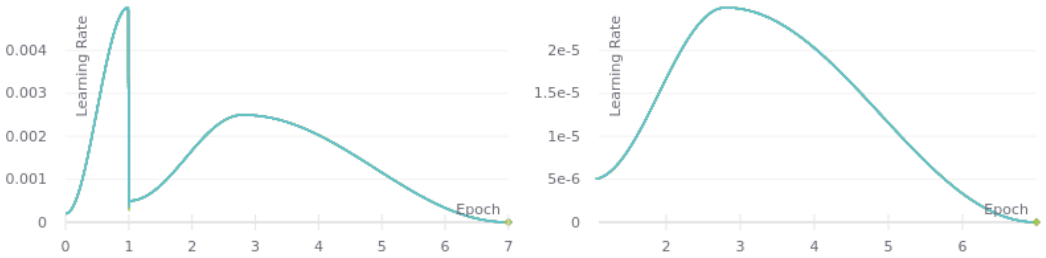
features of the model's body, this part of the neural network generates the final output. The general concept is visualized in figure 4. Because this is a binary classification problem, the output consists of a single value, which can be interpreted as the probability that the observation occurs in the study. When more than one view is available, the model outputs the maximum probability of the observation across the views.

The DenseNet model architecture [6] will be used as the body of the neural network, since it produced the best results for [7]. Two ImageNet pretrained variants of the architecture are leveraged, namely *DenseNet121* and *DenseNet161*, which only differ in the amount of layers present in the model. This also results in different amounts of features produced. This choice was made because we hypothesize higher image resolutions will require deeper models to scale in performance. The pretrained models were trained on the image resolution of 224x224.

The features produced by the body are then aggregated in a pooling layer and further processed in the head of the network. 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. 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 [4].

5.3 Training Procedure & Hyperparameters

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 6 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.



(a) Learning rate of model head

(b) Learning rate of model body

Figure 5: 1cycle learning rate schedules [15] of model head (left) and body (right). In the first epoch only the head of the model is trained with a high maximum learning rate. Afterwards the maximum learning rate is reduced and the body is unfrozen to be fine-tuned with a small learning rate.

Binary cross entropy loss and a batch size of 30 was used for all experiments. One element of a batch contains one study, which can consist of up to 3 x-ray images. Adam [10], in conjunction with the 1cycle learning rate [15] 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}. The learning rate was visualized over the epochs in figure 5. Further regularization was introduced with Dropout [16] 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.

To ensure the changes in the detection rate are caused by the differing image resolution, the same hyperparameters were used for each run. Because the parameter initialization of the head is dependent on random factors, each experiment is repeated 5 times with differing seeds. Additionally, each image resolution is also trained twice, once for the 121 and 161 layer DenseNet variant. This results

in 10 experiments per image resolution. The choice in the extensive number of 7 epochs per run and training with two variants of the same 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 5 NVIDIA Quadro P6000 24GB GPUs, kindly provisioned by [2]. 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 [1].

5.4 Results

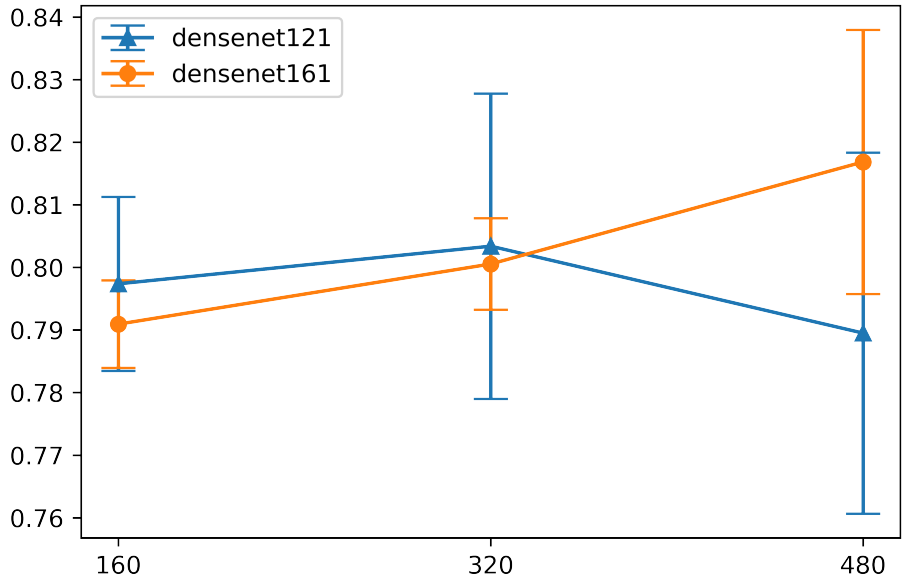


Figure 6: Results grouped by image size and model architecture. Each group consists of 5 runs with identical hyperparameters and differing seeds on 2 ImageNet pretrained DenseNet models, namely DenseNet121 and DenseNet161. Line plot shows the mean AUC per group for each model type. Error bars indicate mean standard deviation.

All results are reported using the scikit-learn's implementation [14] of calculating the Area Under Receiver Operating Characteristic Curve (ROCAUC) from the prediction scores of each model on the validation set. No test time augmentation was used.

Grouping the results by image resolution and reporting the mean and standard deviation AUC per image size and model depth results in figure 6. Of each run, the highest AUC on the validation set over all epochs was used. The line shows the AUC averaged over all runs per resolution and model, while the error bars reveal the standard deviation.

The highest mean AUC was achieved by using an image resolution of 480x480 and the DenseNet161 model architecture. This configuration also achieved the highest general AUC in any run with a score of 84.70. This is a good result for a single model, considering [7] achieved an AUC of 85.80 with ensembling methods.

Increasing the image resolution seems to scale well with model depth/parameters. The more shallow DenseNet121 architecture performed better at lower image resolutions, such as 160x160 and 320x320. Further increase of image resolution made the deeper DenseNet161 model surpass the performance of the more shallow model. The biggest gain in AUC was achieved by increasing the image resolution from 320x320 to 480x480 and using the DenseNet161 architecture. AUC improved by 1.63, which is a 2% improvement.

On the contrary, model convergence starts to become unstable when increasing the image resolution and thus using higher counts of pixels as input to the model. This can be seen by the increasing standard deviation when a higher image resolution is used as input. This indicates a stronger dependence on the randomly initialized starting parameters when initializing the neural network, because this is the only factor that differs in each run. Another reason could be the necessity of hyperparameter optimization, mainly the learning rate. This could have been set too high for runs with higher image resolutions, causing some runs to fail hitting a local minimum in the loss landscape.

6 Outlook

This paper empirically examined the impact of 3 different image resolutions (160x160, 320x320, 480x480) on the detection rate, namely Area Under Receiver Operating Characteristic Curve (ROCAUC) of the pathology Atelectasis from the CheXpert Dataset using ImageNet pretrained models. The results are visualized in Figure 6 and hint at the potential of higher detection rates when both scaling model depth/parameters and image resolution. To fully confirm the hypothesis, further experimentation with higher image resolutions, deeper models and other observations from the dataset is required.

References

[1] N. Chizari. Deep-Learning Pipeline zur Erkennung von Anomalien in Thorax-Röntgenbildern. page 13.

[2] CSTI. CSTI Creative Space for Technical Innovations. URL `https://csti.haw-hamburg.de/`.

[3] D. M. Hansell, A. A. Bankier, H. MacMahon, T. C. McLoud, N. L. Müller, and J. Remy. Fleischner Society: glossary of terms for thoracic imaging. *Radiology*, 246(3):697–722, Mar. 2008. ISSN 1527-1315. doi: 10.1148/radiol.2462070712.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv:1502.01852 [cs]*, Feb. 2015. URL `http://arxiv.org/abs/1502.01852`. arXiv: 1502.01852.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. doi: 10.1109/CVPR.2016.90. ISSN: 1063-6919.

[6] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. *arXiv:1608.06993 [cs]*, Jan. 2018. URL `http://arxiv.org/abs/1608.06993`. arXiv: 1608.06993.

[7] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *arXiv:1901.07031 [cs, eess]*, Jan. 2019. URL `http://arxiv.org/abs/1901.07031`. arXiv: 1901.07031.

[8] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv:1901.07042 [cs, eess]*, Nov. 2019. URL `http://arxiv.org/abs/1901.07042`. arXiv: 1901.07042.

[9] M. Khoshdeli, R. Cong, and B. Parvin. Detection of Nuclei in H&E Stained Sections Using Convolutional Neural Networks. ... *IEEE-EMBS International Conference on Biomedical and Health Informatics. IEEE-EMBS International Conference on Biomedical and Health Informatics*, 2017:105–108, Feb. 2017. doi: 10.1109/BHI.2017.7897216. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5455148/`.

[10] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, Dec. 2014.

[11] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding Transfer Learning for Medical Imaging. *arXiv:1902.07208 [cs, stat]*, Oct. 2019. URL http://arxiv.org/abs/1902.07208. arXiv: 1902.07208 version: 3.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec. 2015. ISSN 1573-1405. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

[13] C. Sabottke and B. Spieler. The Effect of Image Resolution on Deep Learning in Radiography. *Radiology: Artificial Intelligence*, 2:e190015, Jan. 2020. doi: 10.1148/ryai.2019190015.

[14] l. scikit. 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.

[15] L. N. Smith and N. Topin. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. *arXiv:1708.07120 [cs, stat]*, May 2018. URL http://arxiv.org/abs/1708.07120. arXiv: 1708.07120.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, Jan. 2014. ISSN 1532-4435.

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. 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)*, pages 3462–3471, July 2017. doi: 10.1109/CVPR.2017.369. ISSN: 1063-6919.

[18] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? *arXiv:1411.1792 [cs]*, Nov. 2014. URL http://arxiv.org/abs/1411.1792. arXiv: 1411.1792.