Visualizing Transfer Learning in COVID-19 X-rays

Matthew West MWEST@HSPH.HARVARD.EDU *Department of Biostatistics*
Harvard T.H. Chan School of Public Health
Boston, MA, USA

Anjali Jha *Department of Biostatistics*
Harvard T.H. Chan School of Public Health
Boston, MA, USA

Hillary Tsang *Department of Biostatistics*
Harvard T.H. Chan School of Public Health
Boston, MA, USA

Abstract

Classification of medical images is a major application at the frontiers of computer vision. While some studies have investigated the use of transfer learning in classification of COVID-19 via X-ray, there are substantial barriers to implementing such an approach in medical practice. Here we provide an implementation of transfer learning using ImageNet and CheXNet, and demonstrate the limitations of the former in medical transfer learning by visualizing learned features that contribute to network performance using Grad-CAM and activation maximization.

1. Introduction and Background

The Coronavirus disease 2019 (COVID-19) continues to have a devastating effect on global health, and accurately detecting the disease within the population is a crucial step towards addressing the pandemic. While RT-PCR and serological testing will play a substantial role in screening for the wider prevalence of disease, medical imaging can also play a critical role in diagnosis, prognosis, and disease staging in at-risk individuals (Salehi et al. (2020)).

Deep learning technologies, particularly those based on convolutional neural networks (CNN's), have been successfully applied to X-ray and CT data for a variety of purposes and pathologies to date (Pasa et al. (2019), Rajpurkar et al. (2017)). As well as perform classification and segmentation, such methods can aid in identifying and visualising salient features within an image that may have diagnostic relevance. A number of studies have been done on medical imaging data in the context of COVID-19 specifically. Researchers at a hospital in Wuhan validated a CNN-based model for CT data by correlating it with results from PCR testing (Ai et al. (2020)). Reproducing studies on a similar scale elsewhere has proved challenging, though open-source X-ray data scraped from publications has facilitated this (Cohen et al. (2020)). These studies typically combine this open source data of COVID-positive images with other publicly-available chest X-ray datasets (Wang and Wong (2020)).

While these results are promising, it remains to be seen whether a CNN-based classification method will reach a production-ready stage and be widely implemented in medical practice. These results should be interpreted with caution, taking into account implicit

assumptions around experimental design that may not reflect the distribution of data on which these methods are eventually implemented on. In recognition of this, the question of interest for this project is not necessarily optimizing validation performance, but quantifying how significantly different architectures contribute to performance on COVID X-rays by leveraging transfer learning and attempting to visualize these contributions. The project aims to circumvent existing data scarcity by also applying data augmentation to the data available. Having established benchmark classification performance and data augmentation practices in the first part, the latter part of this project focuses on using architectures and weights trained on large benchmark datasets, starting with general-purpose architectures then moving towards architectures trained specifically on X-ray data, with the expectation that this will improve performance and provide more useful features as the models get progressively more specialized. Furthermore, an analysis of visual features identified by classifiers is performed using Grad-CAM, saliency, and activation maximization.

2. Methods

Models based upon CNN's were built and evaluated in Python 3 using the deep learning framework Keras, with a Tensorflow backend (Chollet et al. (2015)). The repository of code is hosted on GitHub, accessible here, and individual models were trained on Google Colaboratory using a GPU backend, as well as on a Dell XPS 15 with an NVIDIA GeForce GTX 1650 GPU. Visualizations of filters and activations were produced using Keras-vis (Kotikalapudi and contributors (2017)).

2.1. Dataset

The dataset used was derived from a public open dataset of chest X-ray and CT images collected by a group at the University of Montreal (Cohen et al. (2020)). This dataset is scraped from publications and contributions from hospitals and physicians, and depicts a variety of disease states. A subset of X-rays was selected having posterior-anterior or anterior-posterior views, and with binary labels assigned based upon COVID or non-COVID findings.

Given the scarcity of data, bolstering the quantity of non-COVID examples with images from external datasets was considered, such as in the case of the COVIDx dataset constructed by Wang and Wong (2020). It was decided, however, that including a substantial quantity of images from another data source to constitute our negative class would confound classification due to origin, and this sort of experimental design may be a contributing factor to high validation performance in other studies using COVID X-ray data. While not ideal, this resulted in a dataset of 233 positive COVID-positive examples and 56 COVID-negative examples, which was split into representative training and validation sets of size 193 and 96, respectively.

2.2. Data Augmentation

Each data augmentation argument in Keras' `ImageDataGenerator` class was considered but only a subset was tested based on a literature search of CNN's applied to chest radiography images: horizontal flip, width shift, zoom, brightness range, channel shift, and

rotation range (Jun et al. (2018), Wang and Wong (2020)). Combinations of these augmentation parameters were tested and a final model utilized a subset of these. Unless otherwise specified, the data augmentation parameters used were `horizontal_flip=True`, `width_shift_range=1,` and `brightness_range=[0.5,1.2].`

2.3. Model Architecture

A variety of architectures were explored, initially for a benchmark classifier trained from scratch, and then using large pre-trained networks as initialization for transfer learning. In each case, X-rays were rescaled in size to 224 by 224 pixels, and an Adam optimizer with default parameters and binary crossentropy loss was used for training (Kingma and Ba (2014)).

For the benchmark CNN, the chosen architecture was a 2D convolutional layer with 32 filters of 3 by 3 pixels each and with same padding and ReLU activation function. Following a 2 by 2 max pooling layer, activations were passed to a second 2D convolutional and max pooling layer, but with 256 filters. This was then flattened and passed through a fully-connected dense layer with 128 units and ReLU activation, before being sent to the final output used in each of our networks, a single dense unit with sigmoid activation.

For the first network using transfer learning, a VGG16 architecture was used initialized with ImageNet weights (Simonyan and Zisserman (2014), Deng et al. (2009)). The final dense layer was removed and all layers prior to the dense layer were frozen during training. In addition to these layers, a 2D convolutional layer with 64 3 by 3 filters was added, followed by a global average pooling layer and single dense unit.

The third main class of networks trained was a DenseNet121 network (Huang et al. (2016)). Like VGG16, a 64 filter convolutional layer and global average pooling layer were added. One main difference was that this network was intialized using weights from CheXNet, the leading architecture in terms of performance on the ChestX-ray14 dataset (Rajpurkar et al. (2017), Wang et al. (2017)). These weights were provided by an implementation of CheXNet in Keras, which can be found at the following GitHub repository. In addition to a DenseNet trained with the base layers frozen and a 64 filter convolutional block, another DenseNet was trained without this block but with all its layers trainable.

2.4. Visualization: Grad-CAM, Saliency and Activation Maximization

Once models were trained, a variety of analyses were performed to compare performance and to visualize features learned by each of the models. The first of these was saliency, which involves backpropogating the gradient of the final class activation with respect to each individual input pixel (Simonyan et al. (2013)). This was followed by Grad-CAM, which instead computes gradients of output classes with respect to forward activation maps, and a weighted combination of these maps followed by a ReLU activation provides more coarse-grained visualization of attention in a CNN (Selvaraju et al. (2016)).

Finally, activation maximizations were generated in order to compare features that individual filters in the output layer and penultimate convolutional layers were visualizing, which is a method first introduced also in Simonyan et al. (2013). The penultimate layers typically hold the most complex features involved in classification, and so it was hypoth-

	Benchmark (no augmentation)	Benchmark CNN	VGG16 (ImageNet)	Dense121 (CheXNet)	CheXNet (all layers)
Batch Size	16	16	16	16	1
Epochs	12	5	12	24	24
Steps per Epoch	N/A	125	125	125	200
AUC	0.672	0.677	0.721	0.725	**0.816**

Table 1: Validation AUC and optimization hyperparameters on the set of models trained.

esized that they would allow visualization of features specific to architectures trained on ImageNet or CheXNet weights.

3. Results

3.1. Classifier Performance

The first experiment performed involved comparing the benchmark classifier with data augmentation to one without, using validation AUC to determine suitable augmentation parameters. Augmentation was found to offer minor improvements in performance, though it was noted that estimations of validation AUC would have a high variance due to the size of the validation set.

Following this, the chosen augmentation parameters were applied to training larger pre-trained classifiers. As can be seen in Table 1, these results broadly validated the hypothesis that performance would improve using transfer learning, and also the hypothesis that networks pre-trained on X-ray data would provide the most significant improvements in performance.

3.2. Visualizations: Grad-CAM and Saliency

Following evaluation of classification performance, the models were investigated using Grad-CAM and saliency to visualize areas in images that were being most activated in prediction. For these purposes, the models were used to make predictions on two images, one COVID-positive, and one COVID-negative. Clinical notes highlight that the COVID-positive patient also has *"progression of prominent bilateral perihilar infiltration and ill-defined patchy opacities at bilateral lungs"*.

From Fig. 1, it can be seen that for the benchmark CNN, though it makes accurate predictions, the entire image is activated evenly, with some general torso shape activations from the saliency map. This implies only general colour-filter type features are being utilized, and that the features being used to classify the image are not clinically relevant.

The VGGNet model predictions are similarly accurate, but with more targeted feature utilization from both the Grad-CAM and saliency plots. For the COVID-positive patient the top left of the image is activated, highlighting a text label which could be confounding classification. For the COVID-negative case, the bottom of the image is activated, though it is not clear that this is diagnostically relevant.

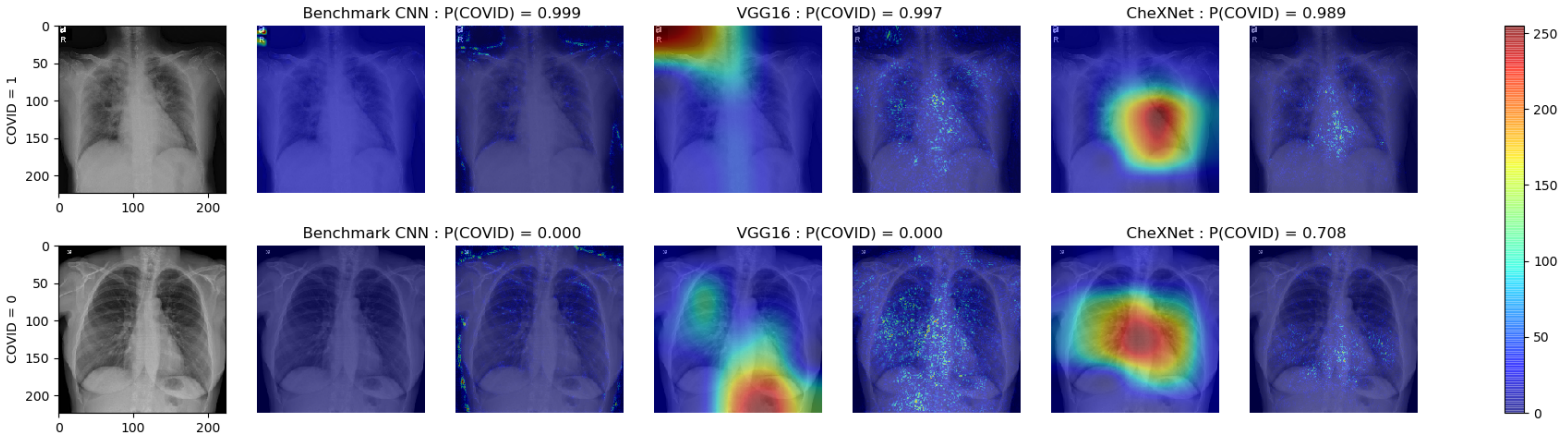

Figure 1: A plot of Grad-CAM and saliency for X-rays from a COVID-positive (top) patient and a COVID-negative patient (bottom), for each of the benchmark CNN, VGG16, and CheXNet models.

The CheXNet classifier has similarly localized Grad-CAM and saliency activation, though these appear to focus on the lungs and not any confounding information like colour or the presence of text. It is worth noting that for the COVID-negative patient, the CheXNet prediction using a threshold of 0.5 would be incorrect, though this classifier achieved a higher AUC across the validation set as a whole.

This visualization appears to validate the hypothesis that classifiers trained on X-ray data will utilize features more relevant to medical classification. Superfluous text within an X-ray image is clearly not diagnostically relevant, and this highlights that any medical imaging classifier trained on ImageNet data should be interpreted with caution.

3.3. Activation Maximization and Filter Visualization

Activation maximization images for each of the three classes of models were generated using the Keras-vis package. Fig. 2 shows images that maximally predict a COVID-positive example for the benchmark, VGG, and CheXNet classifiers. These images are generated by inputting an image of random noise and iteratively backpropagating updates to the image in order to maximize a given class activation.

Fig. 2a shows these images after 2048 such iterations, where a vague lung-shape can be seen in the benchmark classifier, and the patterns that activate the VGG classifier are similar to those typically seen in classifiers trained on ImageNet data (Olah et al. (2017)). The CheXNet max activation doesn't appear to show anything constructive, and so these experiments were repeated using 5% jitter, which moves pixels around while iterating, and removing the total variation regularization applied by default. Fig. 2b shows the results of this experiment, where the lung-like shape for the benchmark classifier is no longer visible, and the VGG activation shows similar albeit sharper ImageNet-style features. The

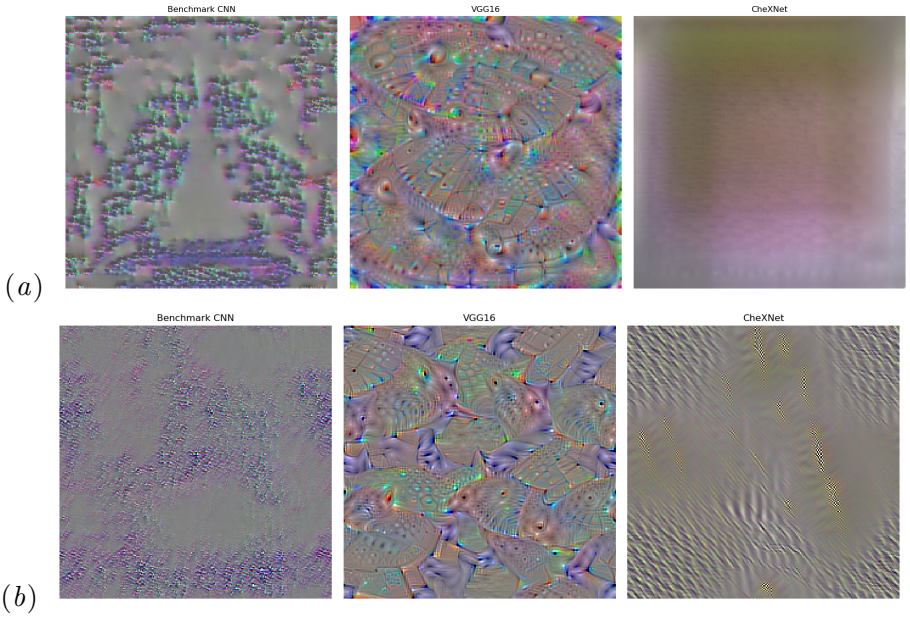

Figure 2: Activation maximization for each of the three models, using 2048 iterations of backpropagation both (a) without jitter, and (b) with 5% jitter.

CheXNet visualization shows that this classifier is utilizing textures distinct from those seen in ImageNet, which may map more closely to those used in classifying X-rays.

Fig. 3 shows the same method applied to the penultimate convolutional layer for each of the three classifiers. While the benchmark layer appears to be learning colour filters, and the VGG model again learning the familiar ImageNet patterns, CheXNet appears to utilize ripple-like textures similar to those seen in the class activation visualisations.

4. Discussion

The goal of this project was to explore how different network architectures impact classification performance on COVID X-rays, particularly by utilizing transfer learning. The results from initial experiments confirm the hypothesis that transfer learning can be useful for classification, and that using networks pre-trained on data of a similar type can improve classification performance.

Visualizations of class activation maps and saliency have shown that this improvement in classification performance correlates with networks utilizing more physically-reasonable image features in making predictions. This analysis has also shown ImageNet weights are likely to be a poor choice of initialization for transfer learning, and networks may learn features in images like text that would silently confound classification if not explicitly visualized. Similarly, activation maximization and visualizing specific convolutional layers has shown that convolutional filters utilize different kinds of textures in making predictions between those trained on ImageNet and CheXNet. This disparity highlights the necessity for similar benchmark classifiers in the domain of medical imaging for use in transfer learning

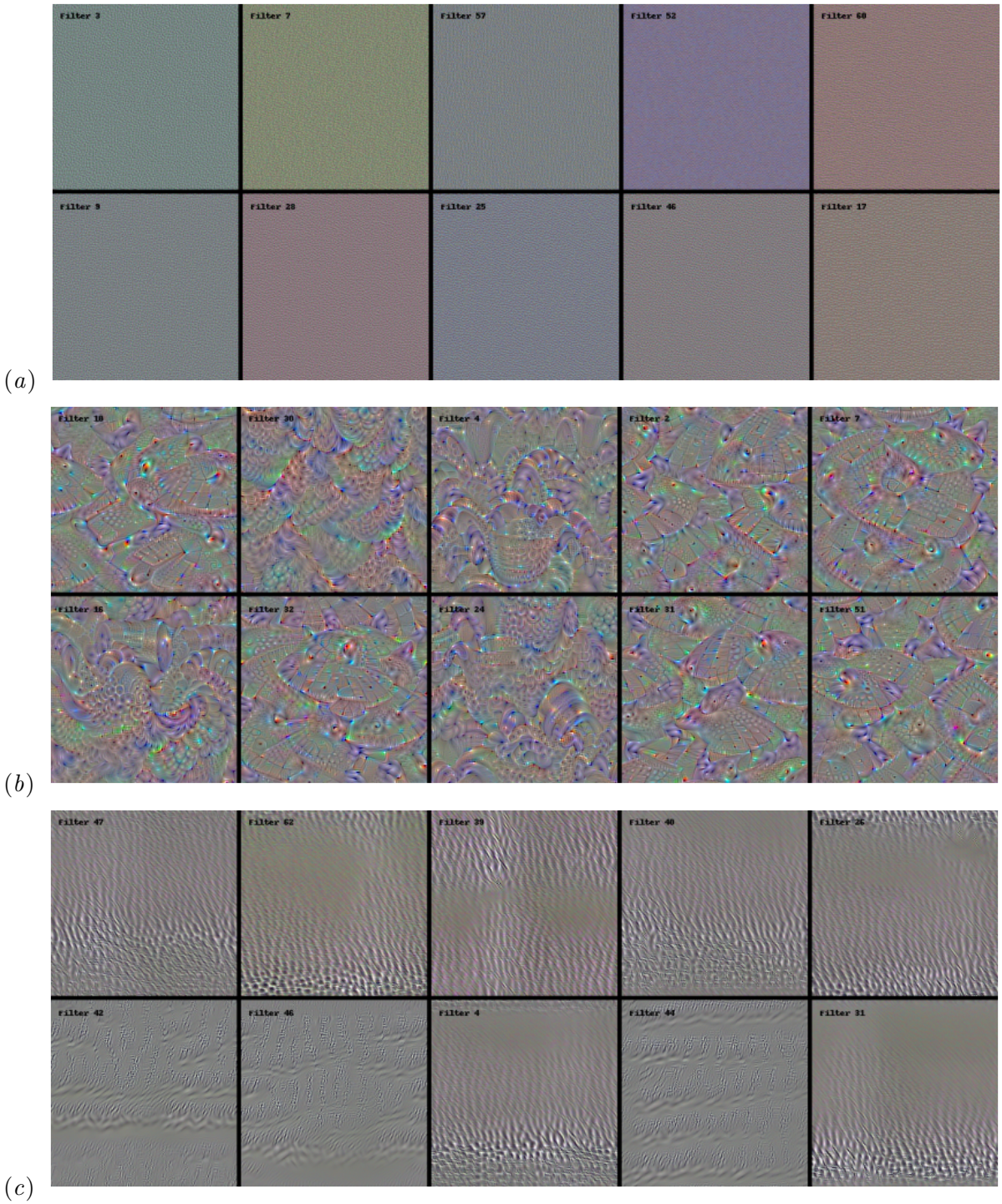

Figure 3: A sample of 10 filters from the penultimate convolutional layers of each of (a) benchmark CNN, (b) VGG16, and (c) CheXNet models, using 5% jitter and no total variation regularization.

applications. This sentiment is echoed and explored in more detail in the following paper from NeurIPS 2019 (Raghu et al. (2019)).

4.1. Limitations

The conclusions of this project could be made more robust if more high quality labelled data was made available, as well as expanding to networks trained on different X-ray datasets to see if similar features or textures are utilized across many different pathologies.

In addition to the numerous limitations in applying CNN's to medical imaging generally, there are limitations specific to the context of COVID-19 diagnosis. The data being used for COVID classification is subject to the problem of selection bias and generalizability, as it has been compiled from contributions by hospitals and physicians and represents a population that had a reason to be in the hospital and receive these tests. This cohort is qualitatively different from the general population that would benefit from a screening, and represents a substantial dataset shift that would have to be accounted for in any useful application of computer vision for COVID screening.

A futher limitation is that COVID patients can have a range of symptoms depending on severity of the disease, and it remains to be seen if clinically-relevant features are even discernible from X-rays. Since some milder cases may not manifest in the from of a observable effect on the lungs, the model may misclassify the milder cases as COVID-negative. While applying such methods to COVID may not be practical in traditional medical workflows and not replace RT-PCR testing, it remains of interest to perform visualization of classification in any application of computer vision to medical imaging, as doing so can provide invaluable interpretability regarding how these models make their decisions.

References

Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. *Radiology*, page 200642, 2020.

François Chollet et al. Keras. https://keras.io, 2015.

Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection. *arXiv preprint arXiv:2003.11597*, 2020.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. URL http://arxiv.org/abs/1608.06993.

Tae Joon Jun, Dohyeun Kim, and Daeyoung Kim. Automated diagnosis of pneumothorax using an ensemble of convolutional neural networks with multi-sized chest radiography images. *arXiv preprint arXiv:1804.06821*, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Raghavendra Kotikalapudi and contributors. keras-vis. https://github.com/raghakot/keras-vis, 2017.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

F Pasa, V Golkov, F Pfeiffer, D Cremers, and D Pfeiffer. Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization. *Scientific reports*, 9(1):1–9, 2019.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3347–3357. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8596-transfusion-understanding-transfer-learning-for-medical-imaging.pdf.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Sana Salehi, Aidin Abedi, Sudheer Balakrishnan, and Ali Gholamrezanezhad. Coronavirus disease 2019 (covid-19): a systematic review of imaging findings in 919 patients. *American Journal of Roentgenology*, pages 1–7, 2020.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. URL http://arxiv.org/abs/1610.02391.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Linda Wang and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images. *arXiv preprint arXiv:2003.09871*, 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.