Automated Radiology: Identifying Pulmonary Diseases with MIMIC CXR

Ben Fox & Nitish Bhardwaj
University of California, Santa Barbara

June 14, 2019

1 Introduction

Radiology is a common practice, with over one billion radiological reports every single year [1]. However, despite its common practice, over the past few years, the number of radiologists has decreased leading to poorer quality of care and reduced interpretation of radiological results [4]. This, in addition to the lack of radiologists in low resource countries (eleven total for Rwanda with a population of twelve million, and two total for Liberia with a population of four million) [4] and error rates of 3-5% radiologists make on a daily basis [1], elaborates necessity for an automated identification tool for x-ray images. With such a tool, this could improve care quality for patients by offering assistive interpretation of x-ray images easing the workload for radiologists. Thus, this project aims to train x-ray image classifier models to identify pulmonary diseases via classification of chest x-ray images utilizing four models: a vanilla convolutional neural network (CNN), VGG16 [7], Alexnet [5], and ResNet50 [2]. Given that the most common imaging performed in the world is chest x-ray imaging, which helps identify a number of heart and lung diseases, as well as fractures and support device locations, this could have substantial impact on the radiological and medical communities. Identification of pulmonary diseases via chest x-ray images was previously made popular by Stanford's CheXpert dataset [3]; however, this project utilizes a brand new dataset from PhysioNet - MIMIC CXR. All models were built using Python and Keras.

2 Data

The dataset used in this project was Physionet's MIMIC-CXR dataset [4], a new dataset as of January 2019 comprising 371,920 frontal, lateral, and other chest x-ray images (the largest chest x-ray imaging set to date) and corresponding diagnoses for fourteen pulmonary and chest related diseases. Diagnoses were acquired utilizing a free text analyzer (the same used in CheXpert [3]) to extract information from radiology notes associated with each image. For each of the fourteen disease labels, there were four classes: -1: uncertain, 0: negative diagnosis, 1: positive diagnosis, and NaN: no mention of disease. Labels that were uncertain (-1) and not mentioned (NaN) were changed to negative (0). Further, a subset of labels was chosen to predict, specifically "No Finding", "Edema", "Pneumonia", "Fracture", and "Support Device". All images not predicting of these labels were removed from the dataset. Each image in the dataset was acquired from an x-ray imaging device, converted to 8-bit pixel format, and was grayscale with an original size of approximately 1200 x 1200. In this project, a subset of the frontal images was used due to computational and storage constraints. The images were all transformed to 224 x 244, normalized by 255, and unit variance scaled. The train, test, and validation sets consisted of 15,535, 1,201, and 1,726. Table 1 summarizes the label class balances for the training and test sets. An example image can be seen in Figure 1.

	Train Set		**Test Set**	
Label	**Negative (0)**	**Positive (1)**	**Negative (0)**	**Positive (1)**
No Finding	9,204	8,057	549	652
Edema	14,402	2,859	1,030	171
Pneumonia	15,508	1,753	1,092	109
Fracture	16,723	538	1,173	28
Support Device	10,011	7,250	750	451

Table 1: Training and test sets label class balances

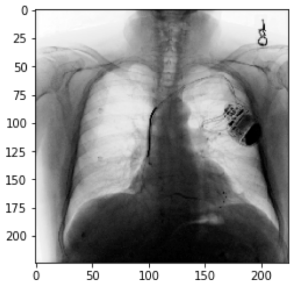


Figure 1: Example image from MIMIC CXR

Overall, this dataset was exciting to work with because there are currently no published studies for MIMIC-CXR and results have potential to be meaningful to the medical communities and offer solutions for better diagnoses and outcomes for patients.

3 Network

In this project, convolutional neural networks (CNNs) were used to train our imaging data. CNNs are known for high accuracy for classification and recognition problems and follow a hierarchical model which works by building a network of filters and max pooling layers to extract features. We utilized four different networks: a vanilla CNN, AlexNet (2012) [5]. VGG16 (2014) [7], and ResNet50 (2015) [2]. These models were chosen to see how they would perform on a new dataset and examine how the number of layers and filter size in a model architecture affects the classification results. Both depth and filter size make a huge difference in terms of information learned by the model. Deeper networks have the benefit of learning more parameters for classification; though, they run into the problem of the vanishing gradient. Further, models with smaller filter sizes use more parameters for learning, therefore learning finer differences about the images. For all models, Adam's optimizer with a learning rate of $1x10^{-5}$, a class weighted binary cross entropy loss function, and sigmoid activation function in the final dense layer were used.

3.1 Vanilla CNN

The vanilla CNN built was a four layer neural network with two convolutional and max pooling layers followed by two fully connected layers. The associated network graph is shown in Figure 2. Each convolution layer included a dropout rate of 0.75. The layers' activation, pooling, and filter sizes are detailed in Table 2.

Layer	Size	Filter/Pool Size	Activation
Input Image	[224,224,1]	NA	NA
CNN 1	[224,224,64]	[3,3]	Relu
Maxpool 1	[112,112,64]	[2,2]	NA
CNN 2	[112,112,64]	[3,3]	Relu
Maxpool 2	[56,56,64]	[2,2]	NA
Dense 1	[128,1]	NA	Relu
Dense 2	[5,1]	[3,3]	Sigmoid

Table 2: CNN model layer architecture

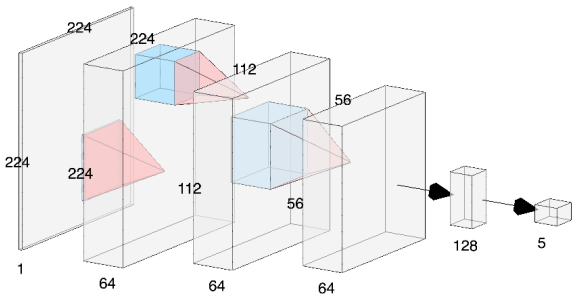


Figure 2: Vanilla CNN network graph

3.2 AlexNet

AlexNet was one of the first deep networks to utilize convolutions in image classification and showed promising results compared to traditional methodologies. It is composed of five convolutional layers followed by three fully connected layers for a total of eight layers. This model uses the relu activation function after each layer and dropout (with a rate 0f 0.6) after every fully connected layer. Its filter sizes are 11 x 11 for the convolutional layers with a pool size of 2 x 2. The full architecture is detailed here [5]. The associated network graph is shown in Figure 3.

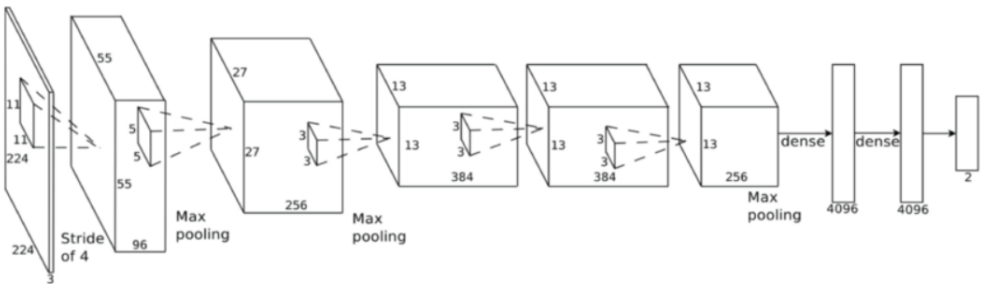


Figure 3: Alexnet network graph

3.3 VGG16

VGG16 is very similar to AlexNet but twice as deep for a total of sixteen layers. This helps identify more and more complex features from the images. It also makes the improvement over AlexNet by replacing large kernel-sized filters (11 x 11 in AlexNet) with multiple 3 x 3 filters, again helping to learn more complex featues. The VGG16 convolutional layers are followed by three fully connected, dense layers. The depth of the network channels start at a small value of 64 and increase by a factor of two after every sub-sampling/pooling layer. The associated network graph is shown in Figure 4. The full architecture is detailed here [7].

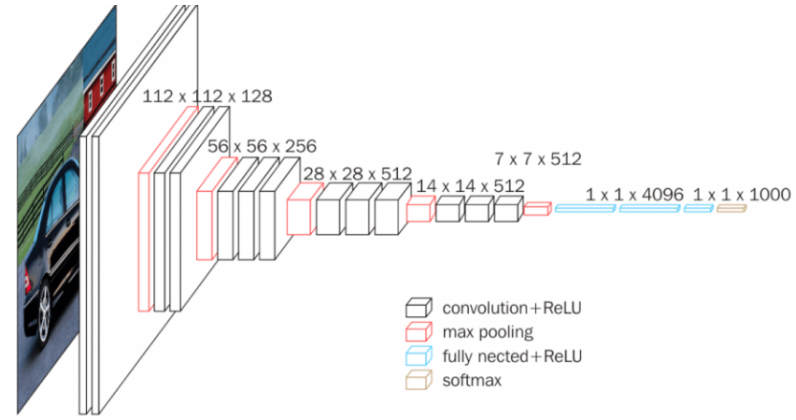


Figure 4: VGG16 network graph

3.4 ResNet50

The ResNet50 architecture is quite different from the other architectures detailed thus far as this network introduces skip connections. One of the common beliefs with CNNs is that accuracy should increase as the network gets deeper. However, the problem with increased depth is that the signal required to change the weights becomes very small at the earlier layers because of such increased depth. This is known commonly as the vanishing gradient problem. ResNet50 addresses this problem by introducing skip connections where the weights of layer x get copied to the next layer instead of creating new weights at each and every layer. No extra computation complexity and parameters are added to the network. ResNet50 is 50 layers deep and utilizes a 7 x 7 filter size. Figure 5 illustrates the ResNet50 model architecture. The full architecture is detailed here [2].

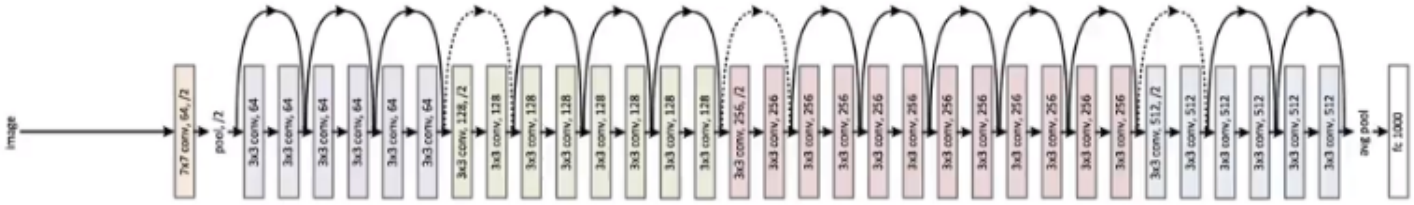


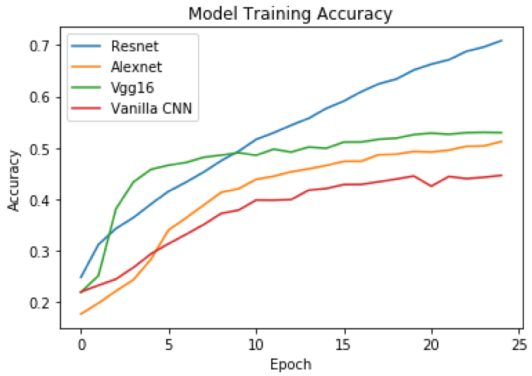
Figure 5: ResNet50 network graph

4 Training

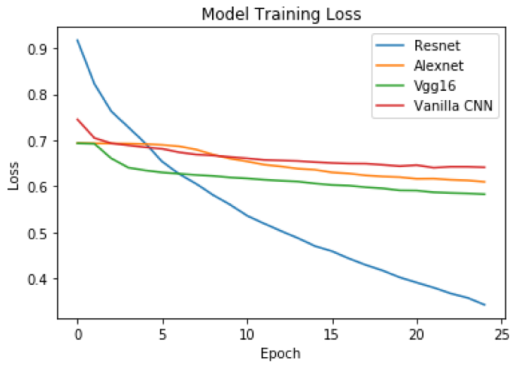
For all models, batch size was 64 images with 25 epochs of training. The corresponding categorical accuracy and weighted binary cross entropy loss per epoch are shown in Figure 6 below for all four models.

5 Validation

For all models, validation and test accuracies were low and improvement slowed after about 5-7 epochs of training. Figure 7 below shows the validation categorical accuracies and weighted binary cross entropy loss measures for all four models



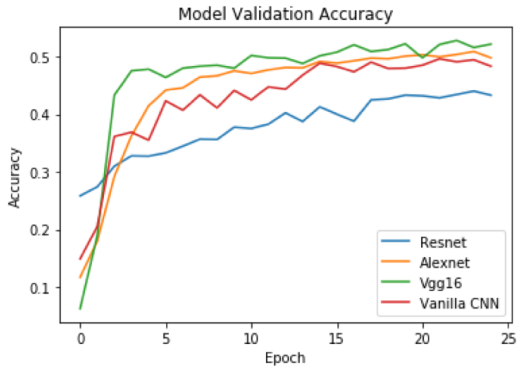
(a) Training categorical accuracy



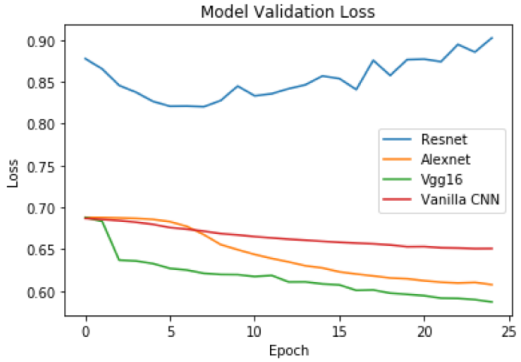
(b) Training weighted binary cross entropy loss

Figure 6: Model training accuracy and loss

per epoch. There was an indication of overfitting in the ResNet50 model as training accuracy continued to increase, while validation loss also increased after about 5 epochs of training. Additionally, accuracy, F1, precision, and recall measures are reported for the test set over all models. See Table 3.



(a) Validation categorical accuracy



(b) Validation weighted binary cross entropy loss

Figure 7: Model validation accuracy and loss

Metric	Vgg16	Alexnet	Vanilla CNN	Resnet
Accuracy	0.4346	0.3963	0.2939	0.3222
F1	0.6154	0.606	0.5216	0.4879
Precision	0.6975	0.6081	0.5974	0.5763
Recall	0.5507	0.6038	0.4628	0.4231

Table 3: Model test evaluations

6 Visualization

Images were visualized utilizing gradient weighted class activation maps (Grad-CAM) [6], which utilize the class-specific gradient information for each label flowing into the last CNN layer of a model to depict an activation map of the important regions in the image for the label specified. This has potential to assist radiologists in diagnosing pulmonary and chest diseases. It also helps discern if the model is identifying the correct portions of an image for identification of disease. Below, in Figures 8, 9, 10, 11, and 12 are examples of each label with an original image that was correctly predicted and the same image with the Grad-CAM heat map overlaid using the VGG16 model.

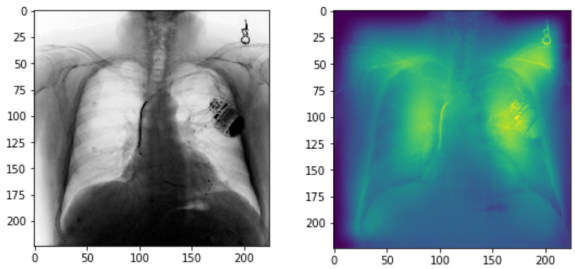


Figure 8: Support device

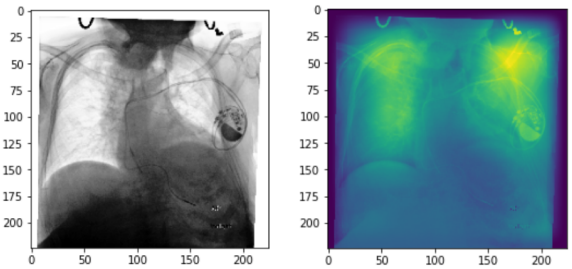


Figure 9: Fracture

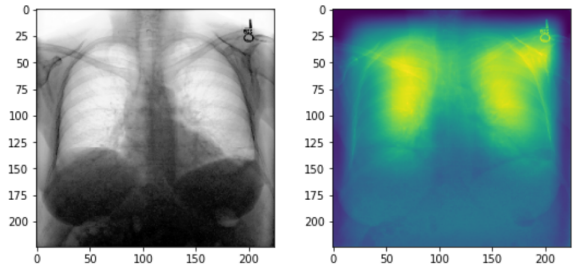


Figure 10: Edema

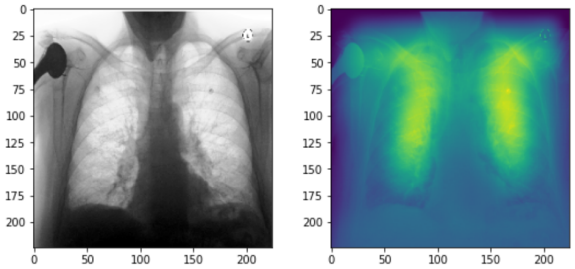


Figure 11: Pneumonia

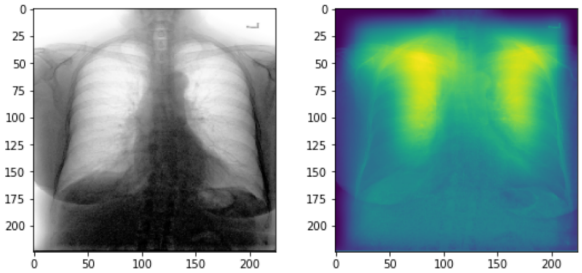


Figure 12: No finding

7 Discussion

Overall, this project provided a starting point for models built on the MIMIC-CXR dataset. Examining the four different models and their performance reveals the nuances of depth and filter size when training an image classifier. Namely, depth does not necessarily provide better classification results (as in ResNet50). This is likely because these deep networks require huge training sets and balanced classes to learn the differences between each class and in this case between each label. The small size of our dataset as well as the imbalanced classes within the labels could have contributed to the overfitting of the ResNet50 model.

Additionally we learned that in multi-label and medical datasets, the F1 score, precision, and recall are better metrics compared to accuracy. This is because these scores take into account false positives and false negatives, while accuracy only considers correctly predicted results. In medicine, false negative or false positive diagnoses could have drastic impacts on a patient and need to be minimized. Thus, considering these metrics, the two best models trained were VGG16 and AlexNet. Future work should be done to train the entirety of the MIMIC-CXR dataset on these models (or derivatives of them) along with all fourteen labels. Down the road, models built on datasets like this will ideally be able to assist doctors with diagnosis of pulmonary and chest diseases and improve patient care and outcomes.

References

[1] Adrian P Brady. Error and discrepancy in radiology: inevitable or avoidable? *Insights Imaging*, 8(1):171–182, Feb 2017.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019.

[4] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042, 2019.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[6] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

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