Chest X-Ray Images for Pneumonia Classification

Yichen Isabel Zhou yz6126@nyu.edu
Center for Data Science
New York University
New York, NY 10003, USA

Yuyue Zhou yz6121@nyu.edu
Center for Data Science
New York University
New York, NY 10003, USA

Abstract

Pneumonia, such as COVID-19, is one of the most fatal diseases in the world, and chest X-ray is significant for pneumonia diagnosis. In this work,we applied CNN models that can detect and classify pneumonia from chest X-ray images. Our three models, ResNet152, DenseNet161 and GoogLeNet were trained on an open source dataset containing 5,856 chest X-rays. We found that GoogLeNet exceeds the performance stated in previous study on the binary classification: normal vs. pneumonia. Then we extended our task to normal, viral pneumonia and bacterial pneumonia classification and achieved compelling results. All codes and visualization results are at `https://github.com/izhou2015/DeepLearningForMedicine`

1. Introduction

Pneumonia, a disease usually caused by bacteria or virus, is one of the most common diseases that cause children to die (Rudan et al., 2008). Treatment guidelines for different types of pneumonia are quite different (Cao et al., 2018). Therefore, accurate and timely diagnosis is critical for further treatment. Physical signs and chest X-rays are two key criteria in pneumonia detection (Lynch et al., 2010). However, chest X-rays have limitations since it highly relies on interpretation from people with specific medical training. Finding a method to quickly and accurately detect and classify pneumonia using chest X-ray images has profound meaning.

Convolutional neural network(CNN) is a powerful deep learning network in image classification as it can extract features and perform classification accurately (krishna et al., 2018). (Kermany et al., 2018) applied transfer learning based on Inception V3 on 5,863 chest X-ray images from 1-5 years old children. They used pretrained weights and retrained the final classifier layer to classify chest X-rays. They performed two binary classifications: normal versus pneumonia classification and bacteria versus virus infection and achieved accuracy of 0.928 and AUROC of 0.968 for normal versus pneumonia classification. For bacterial and viral infection, the accuracy is 0.907 and AUROC is 0.940. Although they achieved good results on binary classification, their model cannot differentiate between normal, viral infection and bacterial infection in one shot.

Based on (Kermany et al., 2018), we proposed three goals of our study: (1) to improve different CNN model performances on chest X-ray pneumonia classification. (2) to expand Kermany's binary classification into 3-class classification: normal, viral pneumonia and bacterial pneumonia. (3) to visualize how CNN models classify images: using classification activation map (CAM) to see the regions that contain the most useful information for models.

2. Methods

2.1 Data

Chest X-ray images we used were collected from children of 1 to 5 years old from Guangzhou Women and Children's Medical Center, which are the same as what (Kermany et al., 2018) used. There are 5,856 images in total. The original dataset contains two folders: training folder with 5,232 images and test folder with 624 images. Each image has a label: normal, virus or bacteria. Details about the images are shown in Table 1. In the training set, normal images are all full-sized, while images labeled as virus or bacteria have been cropped by doctors, retaining infection area only. In the test set, all the images are full-sized. Images are biased, not only in the normal versus pneumonia proportion, but in the size as well.

	Training	Validation	Testing
Normal	1,349	120	114
Pneumonia(Virus)	1,345	80	68
Pneumonia(Bacteria)	2,538	114	128

Table 1: The number of Chest X-ray images in three sets

2.2 Evaluation Metrics

We considered F1 score, the area under ROC curve (AUROC) as our main metrics. Accuracy is also calculated for each epoch in the training and validation process.

2.3 Models

The first two models we considered are ResNet152 (He et al., 2016) and DenseNet161 (Huang et al., 2017) because they both have very low top-1 error from ImageNet. However, each of these two models has more than 400 parameters and would cost too much time to train. Therefore, GoogLeNet (Szegedy et al., 2015), which only has fewer than 200 parameters, were also examined. We also tried AlexNet (Krizhevsky et al., 2012) and VGG (Simonyan and Zisserman, 2014) but both caused gradient vanishing so they were abandoned. Then we trained all models and performed hyper-parameter tuning on both binary and 3-class classification.

2.4 Interpretability

We applied Class Activation Mapping (CAM) proposed by (Zhou et al., 2016) on our validation and test results. CAM can visualize regions that contain the most important information for models. By projecting the weights of the output layer to the previous layers and multiply them with activation maps from the last convolution layer, it enables CNN models to have localization ability.

3. Experiments

3.1 Training Process

We use Cross-Entropy loss for category and then the total loss function is the average mean over categories. The model was trained from scratch, and all layers are fine-tuned. Images are first resized to 364 × 364, then randomly cropped to 320 × 320 pixels, and horizontally flipped for data augmentation. Adam optimizer (Kingma and Ba, 2014), learning rate of 0.001, and batch size of 10 are used for all training process. The number of total epochs varied based on the speed of reaching convergence.

3.2 Validation Process

During validation, the same hyper-parameters are used. After receiving predictive probabilities, a softmax function is applied to translate class with the highest probabilities to predicted label. Then AUROC score and accuracy are recorded for each epoch. The best epoch for each model is selected based on AUROC score.

4. Results

4.1 Binary Classification Results

4.1.1 Validation Results

The validation results from all three models are shown in Table 2. GoogLeNet outperformed the others for all metrics. In addition, it took only half time to train compared to ResNet152 or DenseNet161. Therefore, if the task is to predict whether a patient is normal or catches pneumonia, GoogLeNet should be chosen. At the 74th epoch, the training loss is roughly converged and the validation accuracy and AUROC is the highest. ROC curve and confusion matrix of the best model on the validation set are shown in Figure 1.

Model	Accuary	AUROC	F1	Best Epoch
ResNet-152	0.962	0.956	0.969	178th epoch
DenseNet-161	0.965	0.957	0.972	198th epoch
GoogLeNet	0.968	0.965	0.974	74th epoch

Table 2: Validation Results for Binary Classification

4.1.2 Test Results

Since GoogLeNet performed the best on the validation set for normal versus pneumonia classification, we assessed the performance of it on the test set for pneumonia detection. Our accuracy and AUROC results are 0.932 and 0.980, both of them are higher than Kermany's

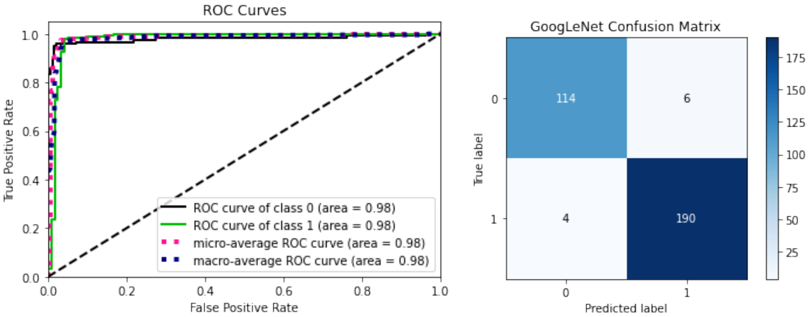


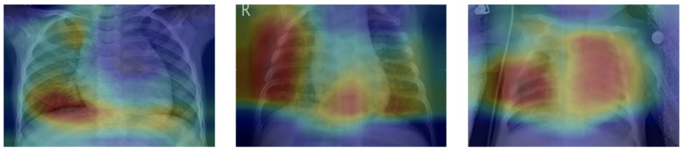
Figure 1: ROC curve and confusion matrix of GoogLeNet on binary classification problem.

results. GoogLeNet achieved F1 score of 0.947. Table 3 compares the performance of our model and Kermany's model.

Model	Accuary	AUROC	F1
Kermany's model	0.928	0.968	/
GoogLeNet	0.932	0.980	0.947

Table 3: Testing results for binary classification

For visualization, we randomly chose 3 corretly classified pictures, labeled normal, pneumonia (virus) and pneumonia (bacteria) from test set and performed CAM. Results are shown in Figure 2.



.

Figure 2: Visualization for binary classification on the test set using Class Activation Maps (left = normal, middle = pneu.(virus), right = pneu.(bacteria))

In the virus and bacteria image, we can see that CAM draws more importance even outside the lung. However, in the training set images, most of the areas lie on the right regions. The reason why this happened is mainly due to the data bias in training and test sets. As mentioned before, in the training set, only infection area is kept in pneumonia chest X-rays, while normal images and all images in the original test set are full-sized. The data bias is deceptive to the model, leading to inaccurate location in the CAM on the test set images.

4.2 3-Class Classification Results

4.2.1 Validation Results

We extended our model to predict not just pneumonia, but also which type of it, including viral and bacterial. The validation results from all three models are shown in Table 4. DenseNet161 outperformed the others for all metrics. However, GoogLeNet only required half the epochs to train than DenseNet161. Here comes the performance and efficiency tradeoff. If one cares more about performance, DenseNet161 is ideal but GoogLeNet could reach not as excellent but still good enough results within much fewer time. Figure 3 presents its ROC curve and confusion matrix of the best model on the validation set.

Model	Accuary	AUROC	F1	Best Epoch
ResNet-152	0.892	0.965	0.883	96th epoch
DenseNet-161	0.920	0.973	0.918	70th epoch
GoogLeNet	0.911	0.958	0.904	47th epoch

Table 4: Validation Results for 3-Class Classification

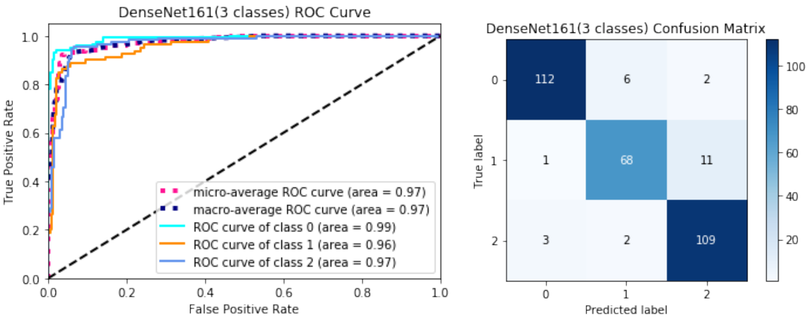


Figure 3: ROC curve and confusion matrix of DenseNet on 3-class classification problem.

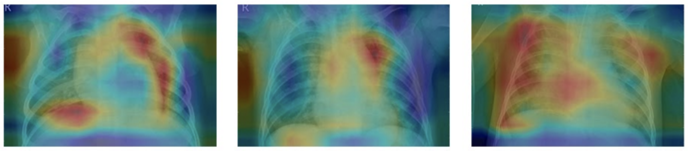
4.2.2 Test Results

DenseNet161 was deployed and assesses on the 3-class test data set. Results are shown in Table 5. We can see that DenseNet161 achieved impressive performance on 3-class classification task.

Model	Accuary	AUROC	F1
DenseNet-161	0.913	0.980	0.904

Table 5: Testing Results for Binary Classification

Results from CAM are shown in Figure 4. Part of CAM red regions lies outside the lung area. The reason is probably due to data bias as discussed in binary classification on test set.



.

Figure 4: Visualization for 3-class classification on the test set using Class Activation Maps (left = normal, middle = pneu.(virus), right = pneu.(bacteria))

4.3 Transfer Learning Results

Knowing that letting the model learn everything from scratch might not be ideal in terms of efficiency, we want to explore how transform learning could improve the performance. We decided to compare pre-trained GoogLeNet with the GoogLeNet trained from scratch that we initially used. Table 6 collected these validation results from these two models when doing binary classification. With a similar number of epochs required, pre-trained GoogLeNet performed better in all metrics and time efficiency. With the previous learning from massive ImageNet datasets (Deng et al., 2009), pre-trained GoogLeNet was able to outperform GoogLeNet from scratch.

Model	Accuary	AUROC	F1	Best Epoch (train time)
GoogLeNet	0.968	0.965	0.974	74th epoch($> 12h$)
Pretrained GoogLeNet	0.971	0.969	0.977	73th epoch($< 10h$)

Table 6: Transfer learning results for binary classification on validation set

A similar procedure is done in the 3-class classification world. This time, the pre-trained GoogLeNet does not win in metrics but it speeds up the learning process. Going back to the performance and efficiency tradeoff, one can choose the better model based on his or her needs. The results are as follows in Table 7.

Model	Accuary	AUROC	F1	Best Epoch (train time)
GoogLeNet	0.911	0.958	0.904	47th epoch($12h$)
Pretrained GoogLeNet	0.898	0.979	0.893	78th epoch($8h$)

Table 7: Transfer learning results for 3-Class classification on validation set

5. Conclusion

We applied different CNN models for chest X-ray pneumonia detection. Compared to previous study, GoogLeNet performed better on normal versus pneumonia classification, achieving higher accuracy and AUROC using the same dataset. In 3-class classification, DenseNet161 has the best results over ResNet152 and GoogLeNet. In all, CNN models show a promising future in the chest X-ray pneumonia detection. We hope that our study can assist in pneumonia chest X-ray detection and classification in the areas with poor healthcare service.

Appendix A. Contribution

Yuyue Zhou: literature review, data splitting, GoogLeNet experiments, paper writing.
Yichen Isabel Zhou: ResNet152 and DenseNet161 experiments, Classification Activation Map results visualization, paper writing

References

Bin Cao, Yi Huang, Dan-Yang She, Qi-Jian Cheng, Hong Fan, Xin-Lun Tian, Jin-Fu Xu, Jing Zhang, Yu Chen, Ning Shen, et al. Diagnosis and treatment of community-acquired pneumonia in adults: 2016 clinical practice guidelines by the chinese thoracic society, chinese medical association. *The clinical respiratory journal*, 12(4):1320–1360, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5): 1122–1131, 2018.

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

M krishna, M Neelima, Harshali Mane, and Venu Matcha. Image classification using deep learning. *International Journal of Engineering Technology*, 7:614, 03 2018. doi: 10.14419/ijet.v7i2.7.10892.

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.

Tim Lynch, Liza Bialy, James D Kellner, Martin H Osmond, Terry P Klassen, Tamara Durec, Robin Leicht, and David W Johnson. A systematic review on the diagnosis of pediatric bacterial pneumonia: when gold is bronze. *PloS one*, 5(8), 2010.

Igor Rudan, Cynthia Boschi-Pinto, Zrinka Biloglav, Kim Mulholland, and Harry Campbell. Epidemiology and etiology of childhood pneumonia. *Bulletin of the world health organization*, 86:408–416B, 2008.

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

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.