COVID-19 and Pneumonia Recognition based on Data Augmentation and Transfer Learning

Zhe Li*
Shanghai University of Electric Power, Shanghai, China

*Corresponding author e-mail: lizhe@shiep.edu.cn

Abstract. Up to now, COVID-19 has been diagnosed in the world as many as about sixty million, which has caused tremendous pressure and burden to hospitals and medical systems. The number of chest X-rays required to be reviewed by doctors of relevant specialties is more than ever before. Besides, in the process of recognition, there are many difficulties for human identification of chest X-ray images, such as the naked eye is difficult to find tiny abnormalities, the chest X-ray images researchers have is limited and doctors is unfamiliar with this new disease. In this case, it is very necessary to use deep learning to help doctors diagnose pneumonia and solve some urgent difficulties. In the field of deep learning, transfer learning focuses on saving solution model of previous problems, and takes advantage of it on other different but related problems. Therefore, we can use the existed Imagenet model to help us train the dataset of COVID-19 and pneumonia. Moreover, deep learning depends on a large amount of data, but when the data is few, data augmentation is able to increase data by some methods, so it is a effective way to overcome the lack of train data. In terms of the good function of these two ways, we use them to improve performance and help doctors improve diagnosis so as to relieve the current tense medical situation.

Keywords: COVID-19, Pneumonia, Deep learning, Data Augmentation, Transfer Learning, Convolutional Neural Networks

1. Introduction
COVID-19 (Corona Virus Disease 2019) is a novel corona virus infection caused by the 2019 new type of pneumonia. Its main symptoms are fever, dry cough and fatigue. A few patients are accompanied by nasal congestion, runny nose, diarrhea and other upper respiratory tract and digestive tract symptoms. Severe patients will rapidly develop into acute respiratory distress syndrome, septic shock, refractory metabolic acidosis, bleeding and coagulation dysfunction and multiple organ failure, thus endangering the lives of patients [1][2]. The transmission approaches of COVID-19 can be divided into direct transmission, contact transmission and aerosol transmission. One effective approach is isolation, which means everyone stays in a fixed place and no contact is allowed [3]. Nowadays, novel corona virus pneumonia has spread to almost all over the world, which has greatly damaged the lives and safety of people around the world. So far, the number of people infected
COVID-19 has surged to about 7.6 million in the world, with a cumulative death toll of about 160 thousand [4].

It is necessary to detect COVID-19 quickly and accurately, since it has highly contagious. As far as the current situation is concerned, in addition to the detection by the kit, doctors can also determine the degree of illness of patients through their chest X-ray images [5][6][7]. At present, the identification of pneumonia image type depends largely on how doctors interpret the chest X-ray images. But because pneumonia may overlap with other diseases and can be similar to many other benign abnormalities, even the best radiologists can easily misdiagnose it. In addition, due to the limitations of naked eye recognition, tiny image abnormalities may sometimes be ignored by doctors. What’s more, researchers do not have enough chest X-rays images of COVID-19 to use, and it is impossible and unreasonable for doctors to accumulate ample experience about COVID-19 quickly. To this point, a good fact is that even if they are not familiar with COVID-19, they know so much things of related diseases, and thus they have ability to refer to these related diseases to treat COVID-19 to some extent.

In recent years, artificial intelligence has been used in many medical fields, including the pneumonia field, and has been well developed. There are many powerful ways of AI to lend us a hand. In order to overcome and solve the problems about limitation of chest X-ray images of COVID-19, we can increase the number of chest X-ray images by data augmentation. To specific, we use the approaches, such as rotation, reflection, noise injection, mix, translation and so on, to produce more images. Neural network needs a large number of parameters, many of which are millions. To make these parameters work correctly requires a lot of data for training, but in many practical projects, it is difficult to find enough data to complete the task. Data enhancement makes limited data produce more data, increases the number of training samples and diversity, and improves the robustness of the model. In addition, changing the training samples randomly can reduce the dependence of the model on some attributes and improve the generalization ability of the model.

Plus, so as to solve the problem of not familiar with chest X-ray images about COVID-19, we can use existed model to deal with different but related image problems by transfer learning. One of the common uses of transfer learning is to use a pre-trained image classification model to transform it into a target detection model or a key point regression model. The reason for this is that the image classification datasets. We can easily obtain a large number of image classification datasets, such as the familiar Imagenet, while the number of target detection is small and the number of samples in the key point regression dataset is much less. If we do not use these few images for training directly through transfer learning, the effect can not achieve the desired accuracy, and it may cause overfitting phenomenon. Therefore, using Imagenet pre training model for transfer learning is of great significance to improve the generalization ability and accuracy of the model.

As a result, we are able to use deep learning to help doctors diagnose pneumonia, and use data augmentation and transfer learning to help improve diagnosis of COVID-19 and pneumonia, so as to speed up doctors' diagnosis and solve the severe problem in current situation.

The remaining framework of the paper is as follows. Section 2 talks about the materials and the methods we used in designed experiments. Section 3 describes the details of the experiments and lists the results obtained by running programs. Section 4 generalizes summary which we conclude in our experiments.

2. Methodology

2.1 Description of the Method
First of all, we input images and then train them by the basic model we had designed. Secondly, we input images and then use transfer learning to help us train. Thirdly, we input images, and process images by data augmentation and resize images, and then we use the basic model which we had designed to train these images which we had processed. Finally, we input images, and process images by data augmentation and resize images, and then these images we had processed are trained by utilizing transfer learning.
2.2 Several Methods in the Experiment
In this section, we give a description of several ways of preprocessing used in our experiments.

(a) Resize images
The size of images of COVID-19 dataset and pneumonia dataset is different. Under this situation, we must unify the size of these images from the same dataset. Thus, we set the image size to (224, 224).

(b) Data augmentation
The CNN model needs a large amount of data to train more effective and better [8]. Nevertheless, we just have 522 COVID-19 images which can be used to train. It is too little to train well. As a result, we must utilize data augmentation to help us preprocess data by translation, rotation, scaling, horizontal flipping and noise injection, which can enrich our images.

(c) Transfer learning
It is a way of machine learning. To specific, transfer learning is able to transfer the knowledge from one field(source filed) to other field(aim field), which can make aim field accomplish the better performance of study.

(d) Early stopping
When we train deep learning neural networks, we usually hope to get the best generalization performance. But all the standard deep learning neural network structures are easy to over fit: when the network performs better and better in the training set and the error rate is lower and lower, in fact, at a certain moment, its performance in the test set has begun to deteriorate.

The common method to prevent over fitting is to add regular terms to the model, such as dropout. However, deep neural network hopes to reduce the optimized parameters by deepening the network level, and get better optimization results. Early stopping can prevent over fitting by intercepting and saving the best parameter model in the whole process of model training.

(e) Reduce learning rate
To be specific, we use ReduceLROnPlateau to help us train better. In our experiments, if val_acc can not increase any more in a certain range, we will reduce learning rate to try to change this situation.

2.3 Basic CNN Model in the Experiment
In our experiment, except the experiment with transfer learning, we use the basic model (Fig.2) to train in other experiment. This model has nine layers, four convolution layers, two pooling layers and three full connection layers.

2.4 Model of Transfer Learning
We use the built-in VGG16 of keras, which includes five convolution groups and three full connection layers. The five convolution groups have 2, 2, 3, 3 convolution layers respectively, so there are 16 layers. VGG16 network has been pre-trained on the whole Imagenet, and pre-calculated weights downloaded from the Internet have been used. This can save a lot of computing power. The VGG16 network has been used for transfer learning again, and the network has completed the learning on Imagenet. This learning can be applied to new specific areas, and the experimental task can be completed by adjusting slightly.

3. Experiments and Results
In this section, we will describe two important datasets, including COVID-19 chest X-ray images dataset and pneumonia chest X-ray images dataset, and propose the results gained from several experiments. We used several methods, including Data Augmentation and Transfer Learning, to predict COVID-19 and pneumonia better and help us increase accuracy of prediction. Finally, we analyze and compare the results caused by the ways.

3.1 Dataset
In this experiment, we used two important datasets, including COVID-19 chest X-ray dataset and common pneumonia chest X-ray dataset. The COVID-19 dataset has 746 images, of which 349 images are diagnosed with COVID-19, and 397 images are normal pneumonia chest X-rays. There are 5856 pieces of common pneumonia dataset, including 4273 pieces of confirmed patients’ chest X-rays and 1583 pieces of normal chest X-rays. In order to do this task better, we divide dataset into train dataset and test dataset. Besides, we set the ratio of train dataset to test dataset is seven to three. So COVID-19 train dataset has 522 images (244 COVID-19 images and 278 normal images) and COVID-19 test dataset has 224 images (105 COVID-19 images and 119 normal images). Pneumonia train dataset has 4099 images (2991 pneumonia images and 1108 normal images) and pneumonia test dataset has 475 images (1282 pneumonia images and 475 normal images).
3.2 Experiments
First of all, we used the CNN network which we had designed to train COVID-19 dataset and pneumonia dataset, which could obtain a bunch of essential running results that we can use to analyze, including train loss, validation loss and validation accuracy.

Secondly, in order to increase validation accuracy, we made use of two methods, which is Transfer Learning and Data Augmentation. As far as Transfer Learning is concerned, we trained our dataset in the model which is from training ImageNet in VGG16. It could help us train related datasets easily and we need not to specially design a new network for our datasets. In terms of Data Augmentation, we used ImageDataGenerator of Keras to preprocess images by translation, flip, rotation, Gaussian noise and so on. Each time, we only processed images of definite batch size and then these images were trained in the network. This method can increase the amount of training data and enhance the generalization ability of the model and improve the robustness of the model.

3.3 Experimental Results and Analysis on COVID-19 Dataset

3.3.1 Experimental Results on COVID-19 Dataset. In table 1, we list the performance of COVID-19, including training loss, validation loss and validation accuracy, gained from different experiments we had designed.

| No.| dataset  | data | transfer | epoch | train loss | val loss | val_acc  |
|----|----------|------|----------|-------|------------|----------|---------|
| 1  | covid-19 | ×    | ×        | 1     | 133.5290   | 6.7481   | 67.86%  |
|    |          |      |          | 118   | 2.2148     | 3.0526   | 79.02%  |
| 2  | covid-19 | ×    | ○        | 1     | 30.4928    | 15.4640  | 56.70%  |
|    |          |      |          | 37    | 5.6100e-3 | 4.4680   | 81.25%  |
| 3  | covid-19 | ○    | ×        | 1     | 53.0440    | 3.8761   | 53.12%  |
|    |          |      |          | 300   | 1.5552     | 1.5848   | 86.61%  |
| 4  | covid-19 | ○    | ○        | 1     | 25.4846    | 17.7579  | 50.89%  |
|    |          |      |          | 95    | 0.7442     | 3.1134   | 81.28%  |
3.3.2 Experimental analysis on COVID-19 dataset. According to the results about COVID-19 which we obtained from table 1, we find that the experimental result without both data augmentation and transfer learning is worst because its validation accuracy is lowest. After we add data augmentation or transfer learning on the basis of No.1, No.2 and No.3 have improvement in different degree. To be specific, compared with experiment 1, the accuracy of No.2 is increased by about 2.23% and the accuracy of No.3 is increased by about 7.59%. From the experiment results, data augmentation has better improvement for accuracy on COVID-19 dataset than transfer learning.

3.3.3 Discussion on experimental results of COVID-19. It should be noted that even though we try to improve accuracy of No.1 by adjusting parameter, including learning rate, batch size and so on, using early stopping, and the final validation accuracy is 79.02%, from the fig.4.(a) we can see that the overfitting is still very serious in the process of training and. Through analyzing, we find that COVID-19 dataset do not have enough images and it just has 522 images used to train. Too few images lead to overfitting. However, we use data augmentation to increase accuracy, from fig.4.(b), overfitting is solved and accuracy has great improvement. Thus, data augmentation is beneficial to solve overfitting and enhance generalization ability.

![Fig. 4. (a) the figure about accuracy of No.1 in table 1 (b) the figure about accuracy of experiment 2](image)

In addition, we find that when we use both data augmentation and transfer learning together in the COVID-19 dataset, the accuracy is not as good as using data augmentation alone, but it is also better than not using them. This situation may because the model used to transfer learning is not very appropriate or COVID-19 does not have ample images.

3.4 Experimental Results and Analysis on Pneumonia Dataset

3.4.1 Experimental results on pneumonia dataset. We list the experimental results of pneumonia in table 2, containing training loss, validation loss and validation accuracy, obtained from different experiments we had designed.

| No. | dataset | data augmentation | transfer learning | epoch | train_loss | val_loss | val_acc |
|-----|---------|-------------------|-------------------|-------|------------|----------|---------|
| 5   | pneumonia | ×                 | ×                 | 1     | 32.2949    | 5.9061   | 82.07%  |
|     |          |                   |                   | 25    | 1.2970     | 1.7076   | 86.74%  |
| 6   | pneumonia | ×                 | o                 | 18    | 8.1690     | 4.0263   | 84.35%  |
|     |          |                   |                   | 164   | 0.0741     | 2.2771   | 89.47%  |
| 7   | pneumonia | o                 | ×                 | 1     | 38.9061    | 5.8769   | 77.58%  |
|     |          |                   |                   | 164   | 0.2219     | 0.3538   | 90.84%  |
| 8   | pneumonia | o                 | o                 | 1     | 13.2710    | 3.8959   | 80.59%  |
|     |          |                   |                   | 46    | 0.4329     | 1.0875   | 91.58%  |
3.4.2 Experimental analysis on pneumonia dataset. In addition, we find that when we use both data augmentation and transfer learning together in the COVID-19 dataset, the accuracy is not as good as using data augmentation alone, but it is also better than not using them. This situation may because the model used to transfer learning is not very appropriate or COVID-19 does not have ample images.

According to the results about pneumonia which we gained from table 2, we still find that the performance of experiment 5 without data augmentation or transfer learning is worst. When we utilize data augmentation on the pneumonia dataset to train, the validation accuracy is increased by about 4.1%. After we use transfer learning to train, the validation accuracy is increased by about 2.73%. The performance of using data augmentation is slightly better than of using transfer learning. All in all, if we use data augmentation and transfer learning together, the validation accuracy is highest and the validation accuracy is increased by about 4.84%. This suggests that data augmentation and transfer learning are good for enhancing accuracy, no matter how much the improvement is.

4. Conclusion
In this paper, in order to help doctors identify pneumonia images more efficiently and quickly, we use data augmentation to preprocess images and use VGG16 to accomplish transfer learning. From the results, these two methods have a good affect on improving the recognition of COVID-19 and pneumonia. Especially for the pneumonia dataset, its accuracy was increased to about 92%. If we can get more and more various chest X-ray images of COVID-19, and use more appropriate CNN model and take advantage of some effective approaches such as data augmentation, transfer learning and so on, COVID-19 dataset is likely to train better and has better performance on accuracy. Besides, in our experiments, the model which we used to transfer learning is not very appropriate, since the increase of recognition accuracy is not very well. Thus, we will design more effective and appropriate CNN model for training in the future work. Methods of deep learning can improve the level of medical diagnosis, which has a possible practical significance in current serious situation.

Reference
[1] Alaa S. Al-Waisy, Shumoos Al-Fahdawi, Mazin Abed Mohammed, et al. COVID-CheXNet: hybrid deep learning framework for identifying COVID-19 virus in chest X-rays images. 2020, :1-16.
[2] Mughahed A. Al-antari, Cam-Hao Hua, Jaehun Bang, et al. “Fast deep learning computer-aided diagnosis of COVID-19 based on digital chest x-ray images”. 2020, :1-18.
[3] Javor D., Kaplan H., Kaplan A., et al. Deep learning analysis provides accurate COVID-19 diagnosis on chest computed tomography. 2020, 133
[4] Chiranjibi Sitaula, Mohammad Belayet Hossain. Attention-based VGG-16 model for COVID-19 chest X-ray image classification. 2020, :1-14.
[5] Zhou Tao, Lu Huiling, Yang Zaoli, et al. The ensemble deep learning model for novel COVID-19 on CT images. 2021, 98
[6] Nayak Soumya Ranjan, Nayak Deepak Ranjan, Sinha Utkarsh, et al. Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: A comprehensive study. 2021, 64
[7] Desai Sudhen B., Pareek Anuj, Lungren Matthew P.. Deep learning and its role in COVID-19 medical imaging. 2020, 3-4
[8] Hoo-Chang S., Roth H.R., Gao M., Lu L., Xu Z., Nogues I., Yao J., Mollura D., Summers R.M. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging, 35 (5) (2016), p. 1285