ISBN: 978-981-18-3959-7
2022 the 12th International Workshop on Computer Science and Engineering (WCSE 2022)
doi: 10.18178/wcse.2022.06.029

Detecting COVID-19 in Chest X-Ray Images Using Apache Spark and CNN

Yangjun Ou[1], Yihong Chen [2+], Yuhang Xie[1] and Ziyi Wang[2]

[1] School of Electronic Information Engineering,China West Normal University, China
[2] School of Computer Science,China West Normal University, China

Abstract.COVID-19 is highly contagious and highly pathogenic, It seriously threatens human life and health. Rapid detection of positive COVID-19 cases is very important in stopping the spread of the virus. At early diagnosis, It is the most simple and rapid indicator for judging changes in the illness. As the COVID-19 chest X-ray image dataset continues to expand, Researchers build a CNN-based COVID-19 detection model on Apache Spark. The model can effectively detect positive cases of COVID-19. This article first introduces the big data platform Apache Spark, Deep Learning Technology CNN, transfer learning techniques, etc. Then, it summarizes the characteristics and deficiencies of the research on chest X-ray image recognition of COVID-19 in recent years.Finally, Under the big data thinking, This paper proposes a technical direction for rapid detection of COVID-19 based on the big data analysis platform Apache Spark and the deep learning algorithm CNN for large-scale COVID-19 chest X-ray image datasets.

Keywords: Chest X-ray image of COVID-19, Apache spark, CNN .

1. Introduction

Since the global outbreak of the COVID-19 epidemic in the past three years, nearly 45 million people have been infected and nearly 6 million people have died around the world. And the numbers keep growing. Currently, Kit testing is prone to misdiagnosis and missed diagnosis, Nucleic acid testing has the disadvantage of waiting too long.Therefore, Rapid testing to detect positive cases of COVID-19 is critical. In the early stages of infection, Virus affects lungs of infected people, Positive cases can be effectively diagnosed by chest X-ray image detection. Timely detection of positive cases could stop the virus from spreading further.This can effectively reduce the fatality rate. In past research, CNN has achieved good results in medical image classification[1].Therefore, Detecting COVID-19 using COVID-19 chest X-ray images and CNN is now a hot research direction.

The World Health Organization recommends the use of test kits, But this approach could have serious consequences. Statistics show that cases of COVID-19-positive patients with multiple negative kit tests within a month, Also, the results of the kit test at different times of the day may be different. In short, the success rate of kit detection is relatively low, about 70%[2], and the sensitivity can only reach 60%-70%[3]. Therefore, identification through COVID-19 chest X-ray images is the most efficient and timely way to detect COVID-19.

This article introduces the related technologies currently used in the recognition of COVID-19 chest X-ray images. Summarized the two technical routes for the detection of COVID-19 chest X-ray images so far. At the same time, it summarizes the shortcomings of these two technical routes. Finally, it is proposed to study the detection of COVID-19 chest X-ray images under the thinking of big data, A new strategy for rapid identification of COVID-19 chest X-ray images based on the big data analysis platform Apache Spark and the deep learning algorithm CNN was proposed.

+ Corresponding author.
E-mail address: 405496216@qq.com

2. Relevant Knowledge

2.1. Big Data

As the COVID-19 outbreak spreads globally, the COVID-19 chest X-ray image dataset has also grown dramatically. For example, the COVID-19 chest X-ray image dataset of the China Chest CT Image Survey Consortium (CC-CCII) is a particularly large dataset.At the same time the dataset is growing rapidly. Therefore, in the study of COVID-19 chest X-ray image recognition, researchers must use the thinking of big data to conduct research. In previous studies on covid-9 chest X-ray image recognition, the concept of big data was not used in the research. The concept of big data is essentially defined as a large amount of data[4].

In the last year, the data of COVID-19 chest X-ray images have grown dramatically, showing the basic characteristics of big data. In the context of big data concepts, processing such a massive dataset of COVID-19 chest X-ray images requires specialized tools. The well-known big data platform Apache Spark has the characteristics of memory-based computing, and is especially suitable for the algorithm of multiple iterative computing such as machine learning [5].

2.2. CNN for COVID-19 Chest X-Ray Image Classification

COVID-19 chest X-ray image recognition is a very popular research direction since the outbreak, Researchers use CNN models VGG19, ResNet, etc. to identify COVID-19 chest X-ray images. On the one hand, compared with the traditional fully connected neural network, the neurons of CNN are partially connected, which reduces the training difficulty and training time. On the other hand, CNNs share weights among neurons in the same layer, in this way the number of weights is reduced, thereby reducing the complexity of the network model [6]. The basic structure of CNN is shown in Figure 1. The structure consists of an input layer, a convolutional layer, a pooling layer, a fully connected layer and an output layer. CNN extracts COVID-19 chest X-ray image features through convolutional layers, and then combines them into more abstract features through sampling layers to form a feature description of the image. Finally, activation functions are used to identify and classify COVID-19 chest X-ray images.

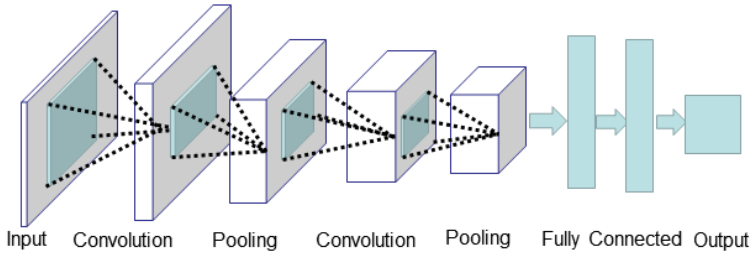

Fig. 1: The basic structure of convolutional neural network

Combining big data platform Apache Spark and CNN for COVID-19 chest X-ray image recognition, the combination of these two technologies makes COVID-19 chest X-ray image recognition efficient。It will become the main strategy for COVID-19 chest X-ray image recognition [7].

2.3. Apache Spark

Apache Spark is a framework structure based on memory computing, which makes it unique in deep learning algorithms. Existing studies have deployed deep learning algorithms on Apache Spark to recognize COVID-19 chest X-ray images, and have achieved certain results. Currently, we can summarize Apache Spark into five components: Spark SQL, MLLib for non-deep learning, GraphX, Spark Streaming,Deep learning. as shown in Figure 2.

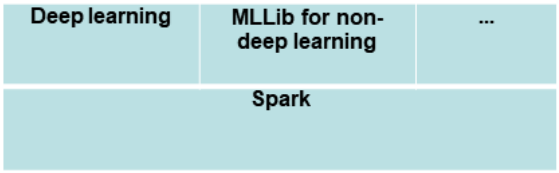

Fig. 2: Deep learning for Apache Spark

Spark MLLib only provides simple non-deep machine learning algorithms such as classification, regression, clustering, etc. Apache Spark is a distributed computing framework developed for large-scale data processing. Its in-memory computing framework is very beneficial to deep learning algorithms that require multiple iterations. So deploying deep learning algorithms on Apache Spark to recognize COVID-19 chest X-ray images has huge advantages.

2.4. Transfer Learning for COVID-19 Chest X-Ray Image Classification

The researchers applied the pre-trained CNN model on ImageNet to the COVID-19 chest X-ray image dataset to train a new CNN model to recognize the COVID-19 chest X-ray image. Transfer learning is a technique to transfer the experience gained on the 'source' dataset to a new 'target' dataset[8]. In the study of COVID-19 chest X-ray image recognition, transfer learning techniques solve the problem of insufficient COVID-19 chest X-ray image datasets. The concept of transfer learning can be represented in Figure 3.

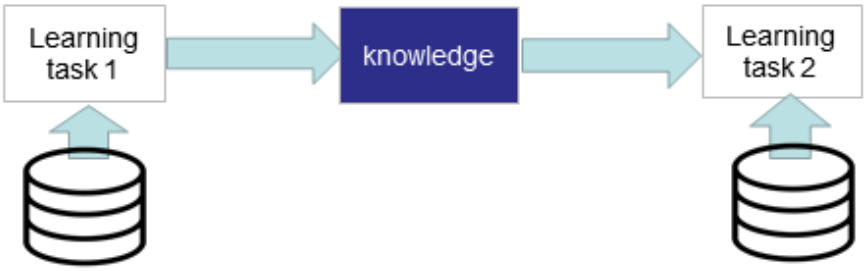

Fig. 3: Learning process of transfer learning

In the COVID-19 chest X-ray image recognition research, we use the pre-trained model on ImageNet to build a new model. New model applied to small COVID-19 chest X-ray dataset, This approach could address the shortage of chest X-ray image datasets for COVID-19.

2.5. CC-CCⅡ Dataset

The CC-CCⅡ dataset is a publicly available COVID-19 lung CT dataset, which is currently the largest dataset for COVID-19 lung CT images, with a total of approximately 617,775 CT slices. Including new coronary pneumonia (NCP), common pneumonia (CP), normal control group (CP) three categories [9]. The dataset is constantly being updated and expanded, and researchers should study the dataset under the concept of big data.

3. Application of COVID-19 Chest X-Ray Image Recognition

Early screening to detect positive infections has been a challenge since the start of the COVID-19 pandemic. However, the kit antigen detection is prone to misdiagnosis and missed diagnosis, and the nucleic acid detection consumes a long time. These two methods bring great hidden dangers to the prevention and control of the epidemic. Deep learning technology has been widely used in the recognition of medical images, and has achieved good recognition results in some aspects. Compared with traditional medical image recognition methods, deep learning can effectively mine the potential nonlinear relationship in medical images, thereby improving the feature extraction efficiency of medical images. Since 2019, many researchers have applied convolutional neural networks to the recognition of COVID-19 chest X-ray images, which has put forward new ideas and possibilities for timely and accurate screening of COVID-19 positive infections. In the research of COVID-19 chest X-ray image recognition by deep learning algorithms, the first problem to be solved is the problem of insufficient training dataset, CNN models must rely on large-scale training datasets to prevent overfitting, enhance generalization, and improve accuracy. Table 1 shows the application of deep learning CNN, Apache Spark, Transfer Learning in the field of COVID-19 chest X-ray image recognition in recent years.

Table 1:Research on the recognition of chest X-ray images of COVID-19

Reference	Architecture	Classes	Transfer Learning	data augmentation	Apache Spark	Data set COVID-19/Others	Accuracy
[10]	CNN	2	N	400 images increased to 800	N	489/980	99.17
		4	N	400 images increased to 800	N	498/2940	94.03
[11]	InceptionV3 ResNet50 VGG19	3	Y	N	Y	354/708	97.1 98.55 98.55
[12]	InceptionV3 ResNet50	2	Y	N	Y	160/160	99.01 98.03
[13]	MobileNetV2	2	N	13808images increased to 52000	N	3616/10192	97.00

Reference [10] proposed a new CNN model to achieve 2-classification and 4-classification of COVID-19 chest X-ray images. Using data augmentation to address insufficient datasets of chest X-ray images for COVID-19. Compared with the model proposed by the cited article, This CNN model uses the largest dataset of COVID-19 chest X-ray images. And the model achieves 99.01% and 94.03% accuracy in 2-class and 4-class, respectively, with the least number of parameters.

Reference [11] adopts the method of combining the big data platform Apache Spark and the transfer learning algorithm, and takes the lead in using the big data platform Apache Spark in the research of COVID-19 chest X-ray image recognition. A new research method is implemented by combining Apache Spark with pre-trained CNN models InceptionV3, ResNet50 and VGG19 on ImageNet. In the case of a small dataset of COVID-19 chest X-ray images, The 3 models achieved 97%, 98.55% and 98.55% accuracy in 3 classifications respectively. The training of this model takes a lot of time because of the transfer learning technique used.

Reference [12] also proposed a method of combining Apache Spark and transfer learning, and also used the pre-training models InceptionV3 and ResNet50 on ImageNet for transfer learning. In the 2 classification problem, it achieves 99.01% and 98.03% accuracy respectively.

Reference [13] proposed an improved MobileNetV2 model. The data augmentation technique was used in the experiment to increase the COVID-19 chest X-ray image dataset from 13,808 to 52,000. The improved MobileNetV2 model achieves 98% accuracy in 2 classification, and the pre-trained MobileNetV2 model on the same dataset achieves 97% accuracy. And the improved MobileNetV2 model requires less compilation time. Therefore, the improvement of the CNN model in the COVID-19 chest X-ray image recognition research is also a work that needs to be studied.

From the summary of the above literature, we can draw the following points:

- Among all the recognition experiments of COVID-19 chest X-ray images, only a few experiments are based on the big data analysis framework Apache Spark.
- The total size of the COVID-19 chest X-ray image dataset mentioned in all the above experiments varies from 200 to 4000 images.
- The above experiments using the transfer learning strategy used a pre-trained model on ImageNet.

In conclusion, the combination of the Apache Spark platform and transfer learning technology enables a breakthrough in the identification of COVID-19 chest X-ray images. This also provides a theoretical basis

for future research[14]. At the same time, it provides ideas for the realization of COVID-19 chest X-ray image recognition technology based on the big data analysis platform Spark and the deep learning algorithm CNN.

4. Conclutions

This paper introduces the research status of COVID-19 chest X-ray image recognition. In response to the shortage of COVID-19 datasets, the two major strategies used by researchers in the past: One idea is data augmentation; the other is Transfer Learning. This article summarizes the inadequacies of these two strategies, These two strategies are no longer applicable in future COVID-19 chest X-ray image recognition studies. Combination of Apache Spark, data augmentation and Transfer Learning technologies drives advances in COVID-19 chest X-ray image recognition with the efforts of researchers, At the same time, there are still many problems, and there is still a lot of work that we need to do:（1）Research on detection of COVID-19 in chest X-ray images based on Apache Spark and CNN technology requires more datasets，It is very important to introduce the CC-CCII dataset into future research。（2）For identification of COVID-19 chest X-ray images, it is difficult to determine which structured CNN model to use，Find a balance between recognition accuracy and model training time (the bigger the model, the more parameters it takes, the more training time it takes).

5. References

[1] Venkatesan N J, Shin D R, Nam C S. Nodule Detection with Convolutional Neural Network Using Apache Spark and GPU Frameworks[J]. Applied Sciences, 2021, 11(6): 2838.

[2] Mahase E. Coronavirus: covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate[J]. 2020.

[3] Xie X, Zhong Z, Zhao W, et al. Chest CT for typical coronavirus disease 2019 (COVID-19) pneumonia: relationship to negative RT-PCR testing[J]. Radiology, 2020, 296(2): E41-E45.

[4] Asri H, Mousannif H, Al Moatassime H, et al. Big data in healthcare: challenges and opportunities[C]//2015 International Conference on Cloud Technologies and Applications (CloudTech). IEEE, 2015: 1-7.

[5] Santosh T, Ramesh D, Reddy D. LSTM based prediction of malaria abundances using big data[J]. Computers in Biology and Medicine, 2020, 124: 103859.

[6] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[7] Xu J, Ma S. Image classification model based on spark and CNN[C]//MATEC Web of Conferences. EDP Sciences, 2018, 189: 03012.

[8] Tan C, Sun F, Kong T, et al. A survey on deep transfer learning[C]//International conference on artificial neural networks. Springer, Cham, 2018: 270-279.

[9] Zhang K, Liu X, Shen J, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography[J]. Cell, 2020, 181(6): 1423-1433. e11.

[10] Belman-López C E. Detection of COVID-19 and other pneumonia cases using convolutional neural networks and X-ray images[J]. Ingeniería e Investigación, 2022, 42(1).

[11] Awan M J, Bilal M H, Yasin A, et al. Detection of COVID-19 in chest X-ray images: A big data enabled deep learning approach[J]. International journal of environmental research and public health, 2021, 18(19): 10147.

[12] Benbrahim H, Hachimi H, Amine A. Deep transfer learning with apache spark to detect covid-19 in chest x-ray images[J]. Romanian Journal of Information Science and Technology, 2020, 23(S, SI): S117-S129.

[13] Akter S, Shamrat F M, Chakraborty S, et al. COVID-19 detection using deep learning algorithm on chest X-ray images[J]. Biology, 2021, 10(11): 1174.

[14] Z. M. Tun and M. Aye Khine, "Cardiac Diagnosis Classification Using Deep Learning Pipeline on Apache Spark," 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2020, pp. 743-746, doi: 10.1109/ECTI-CON49241.2020.9158314.