The Study of the Impact of Data augmentation on the Performance Based on COVID-19 Dataset

Ruiqi ZHUANG
department of computer science, Hong Kong Baptist University, Kowloon, Hong Kong, 999077
* Corresponding author: 20251556@life.hkbu.edu.hk

ABSTRACT

Due to the emergent need of medical staffs to conduct certain medical operations, people term to use artificial intelligence to assist them like medical image analysis. However, currently medical image analysis is facing the dilemma of imbalanced data. This research used OpenCV to conduct offline image augmentation and then combined output of different augmentation methods to generate new training dataset that has balanced data. Therefore, it could deal with the issue of imbalanced data. Later, the research used the new datasets (i.e. the balanced dataset) to train the same Convolutional Neural Network (CNN) model as previously trained using imbalanced data, and compare different evaluation scores including F1 score, accuracy, precision and recall score for comparison. After looking into the histograms and table of the result, this research discovered a considerable increase in F1 score and precision, which indicated that, the offline image data augmentation generally has a good performance in COVID image dataset, except the situation where the method of sharpening is applied. However, one case that this research cannot include is that, simply using data from one augmentation method. (i.e. using output from increasing contrast only) This is because the size original dataset is small so that output from one augmentation method may not satisfyingly tackle the problem of imbalanced data.

Keywords: COVID-19, medical image analysis, image augmentation, image processing

1. INTRODUCTION

Coronavirus disease 2019 (COVID 19) is a kind of pneumonia outbroke in 2019. This pneumonia is highly contagious and has a fatality rate of 2.9%, up to 20 May 2020 this pneumonia has infected 4,806,299 persons and caused 318,599 deaths [1]. Conducting one nucleic acid detection requires around 24 hours to 48 hours, which is time-consuming. Additionally, considering the fact that COVID-19 has a high infectivity, conducting a large-scale nucleic acid detection may require a huge number of medical staff. Therefore, using artificial intelligence assisted diagnosis would be an effective solution to the issue of long waiting times as well as the lack of medical staff.

With the aid of deep learning, computers are able to develop machine learning models that receive a certain kind of images from patients, such as computed tomography (CT), and magnetic resonance (MR), and output the diagnosis result predicted by computer algorithms. That is the application of medical image analysis in artificial intelligence assisted diagnosis [2]. This method may strongly reduce the waiting time of the process of detecting COVID-19. Additionally, since the majority of the process is done by computers, it could solve the issue of a shortage of medical staff.

Nowadays due to the considerable variations in pathology together with the potential fatigue of humans, specialists have transferred the mission of image interpretation to the computer [2]. Unfortunately, the performance of medical image analysis models may be negatively affected by imbalanced training data. One research conducted by [3], it reveals that if the training data set is not able to fulfill a minimum balance, the performance of the model would decrease consistently. And considering the fact that the majority of the data including CT and MR are the privacy of patients, that means these data are not accessible by researchers. Therefore, a conclusion could be drowned that researchers and developers may come across the dilemma of imbalanced data resulting from insufficient data provided by hospitals. As a result, it is an emergent issue to tackle the problem of imbalanced training data.

One commonly used method is data augmentation, in which people perform some intensity or geometric transformation on pictures so that the training result may have a better performance, additionally, augmentation is able to generate new

images and therefore deal with the data imbalance issue [4]. Among all kinds of augmentation methods, generative adversarial network (GAN) is the most commonly used because it seems that GAN does have an excellent performance on image data generation [5]. It could base on the given image data and generate brand new pictures and fill in the training data set, and therefore deal with the problem of imbalanced data. However, one matter that people should take it into consideration is that those pictures generated by GAN are not the real pictures of patients. Thus, it exists the probability that the new pictures may not be able to capture the extremely small features of input images, especially on medical images. As a consequence, GAN may not be a satisfying solution to data imbalance in medical image analysis.

To explore a new method of solving data imbalance, this research used OpenCV to conduct offline data augmentation by rotation, intensity transferring, and reshaping pictures. And finally, it evaluates the result of augmentation by comparing the F2 score, recall precision and accuracy of two models trained by the augmented dataset and original dataset separately. This study is an exploration of efficient augmentation methods in, particularly, image analysis that requires high precision like medical image analysis. On the one hand, it prevents the hidden danger of learning unreal pictures; on the other hand, it provides researchers with new ideas of augmentation for extremely sophisticated images.

2. METHOD

2.1 Dataset description and preprocessing

The data used in this study is from a database of chest X-ray image for COVID-19 developed by a group of researchers from Qatar University, University of Dhaka Bangladesh collaborating with medical doctors from Pakistan and Malaysia. The complete data can be obtained from Kaggle community [6]. In total 4 categories are included in the dataset. The image data in the data involves COVID-19 positive cases, normal cases, lung opacity cases and viral pneumonia cases and their corresponding lung masks.

In the given dataset, there are 502, 187 images. 41, 720 in the training set and 42, 320 in the testing set. Each set has 4 categories which are COVID, Lung Opacity, Normal, and Viral Pneumonia. For each image in the dataset is of size 299 by 299 and with red, green, blue (RGB) values. Figure 1 provides sample images of the collected dataset.

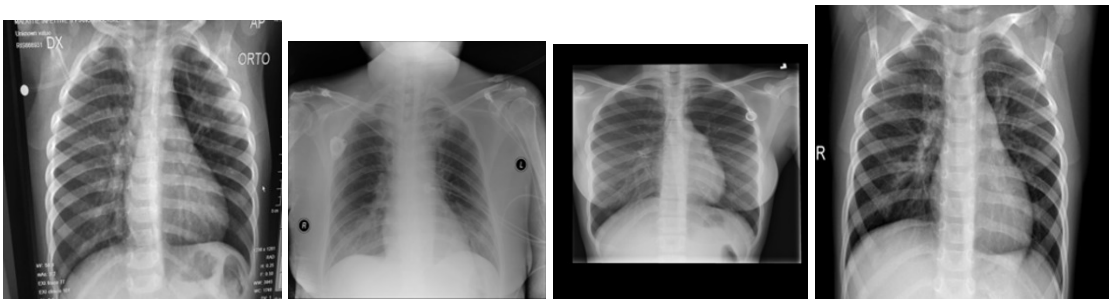

Figure 1. Sample data for COVID infected, lung opacity, normal, pneumonia cases.

2.2 Augmentation

In some cases, the deep learning models may come into the problem where models could not detect the features of image data excellently. Commonly, the problem is result from imbalanced data and blurred images. Therefore, image augmentation which includes rotating pictures, intensity variation, and reshaping images, could be applied to tackle these issues. In general, the augmentation can be divided into two categories, one is online augmentation which do not generate new image data, and offline augmentation which generates new image data.

To deal with the problem of imbalanced data, offline augmentation is the most preferable, since this method is able to generate new image data, so as to fulfill the insufficient data. This research used 4 ways to do augmentation, which are brightness increasing by 50, increasing contrast by 1.5, rotate for 90 degrees, sharpen the image, each generating 806 images. Later the research combinate output of each augmentation method and obtain 1612 images for each combination. In addition, normalization was also carried out in this study to speed up the training process and improve performance

2.3 CNN model

Convolution neural network (CNN) is a deep learning model that has a strong performance in dealing with images input, which is widely used in various fields [7-10]. The convolution layers included in the CNN that are able to detect and

extract features on the image by means of detecting rapid changes of color value like RGB value and gray level. Then as the layer goes deeper, CNN is able to detect bigger features. Finally, by combining different features it detected in convolution layers, the model would be able to have a clear image of the image that is being detected.

Additionally, CNN uses pooling layer to tackle the situation where the object being detected may be rotated or rescaled, and additionally, pooling layer may reduce the number of parameters, so that it may prevent overfitting. Two most commonly used pooling layers are max pooling layer and average pooling layer.

The architecture of CNN for this research consists of 3 blocks. The first block has two 2-dimension convolution layers followed by a max pooling layer and a dropout layer. The activation functions for convolution layers are all ReLu function. The second block is consisted of three 2-dimension convolutional layer followed by a max pooling layer and a dropout layer. The same as the previous block, ReLu activation functions are applied for each convolution layers. The final block is a fully connected neural network, there are 64 neurons for the first and the second layer, and softmax activation function is used in the output layer and output a vector of length of 4, each representing the probability of the corresponding label.

2.4 Implementation details

This research used TensorFlow to implement the CNN model. In order to access the image data stored in Google drive, the model is trained on Google Colab, and GPU is used to train the model, regarding the fact that GPU has a better performance in processing image data comparing to CPU. After several rehearsals on a tiny dataset, finally, the research made a decision that the learning rate for this model is set to be 0.001, chose RMSprop as the optimizer, and set the number of epochs to be 15. Loss function is Categorical cross entropy

3. RESULT AND DISCUSSION

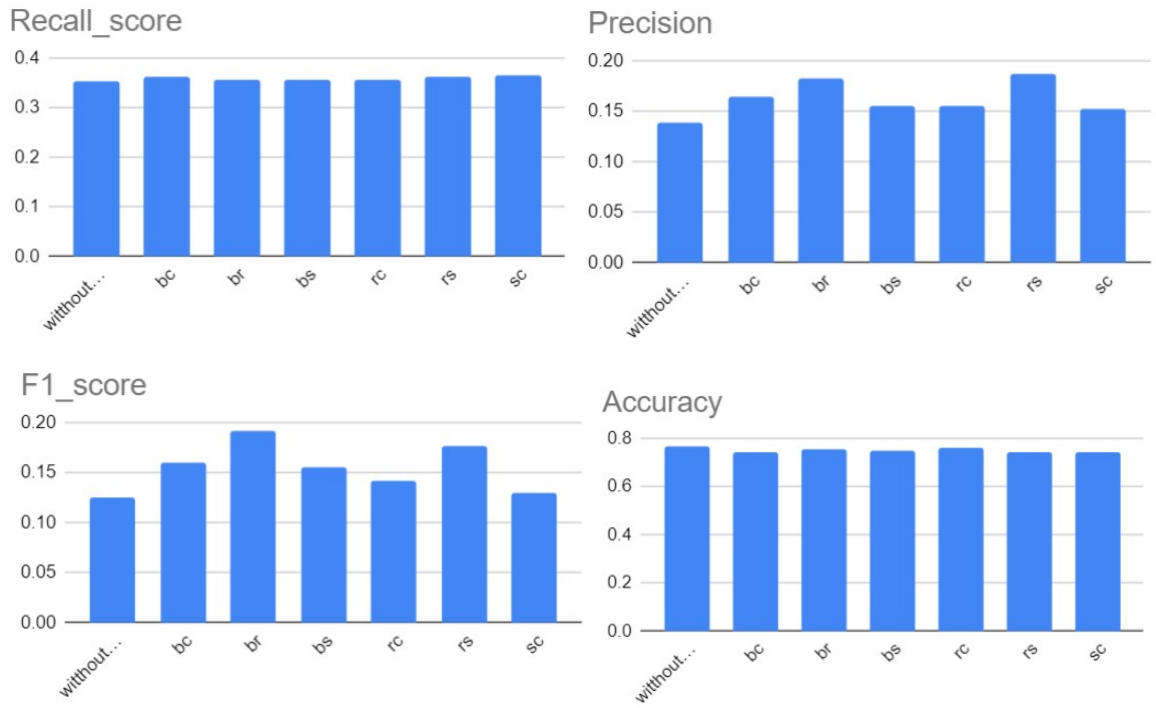

Figure 2. The performance of the model.

Figure 2 presents the result separately. In the x axis, different character stands for different method of augmentation method. The b means increasing brightness, r means rotation, c means adding contrast and s stands for sharpening the images. And the first column is the performance of original model without image data augmentation. Combinations of different character means combinations of augmentation method.

Table 1. Numerical result.

Item	Without augmentation	bc	br	bs	rc	rs	sc
Recall score	0.35159	0.362927	0.354899	0.354427	0.357024	0.362455	0.363636
Precision	0.13827993	0.16445402	0.18227216	0.15577191	0.1553719	0.1867284	0.1541636
F1 score	0.12452544	0.1594306	0.19147541	0.15523216	0.14145974	0.17638484	0.12995246
Accuracy	0.7667	0.7419	0.7554	0.7455	0.7587	0.74	0.74
SUM	1.38109537	1.42872162	1.48404657	1.41093107	1.41255564	1.46554824	1.38600482

Table 1 shows the numerical result of augmentation performance. In the first row, different character indicates different augmentation method. (e.g. bc means increasing brightness and contrast) The last row is the summation of the previous four rows. (i.e. Adding all of the evaluation score together.) The result of performance of different models is shown in Figure 2, it could be obviously observed that, there is an increase of recall score, precision and F1 score after applying augmentation and combining output of different augmentation methods.

Additionally, by observing Table 1, it could be discovered that, among all combination outcomes, increasing brightness in combine with rotating images has the highest enhancement in overall performance. Moreover, there exists considerable increasement for some evaluation scores. For example, it could be found that the F1 score could be increased from 0.12452544 to 0.19147541. However, some evaluation scores do not seem to indicate a significant enhancement. For example, there is not a remarkable increase on both accuracy and recall score.

This research discovered a good performance on recall score when augmentation method of adding contrast is involved. This may be due to that, increasing contrast may make the character of objects in pictures clearer and brighter, therefore CNN may be able to better extract or detect the objects in the pictures.

In general, it could be discovered that, the model has a satisfying performance on recall score precision and F1 score when the augmentation method of rotating images is involved. This may be due to the fact that rotation do keep the majority of features on original pictures. It therefore kept the features of images when training the CNN model. One more phenomenon is that sharpening images may not be a beneficial strategy for COVID image data augmentation. Since the majority of combinations that involved sharpening do not have a remarkable increase in most of the evaluation scores. This may be due to the fact CNN already has a strong ability to catch the boundaries. And thus, sharpening may not augment images well. However, one factor that should consider is that this research does not consider filling single augmentation result into new training dataset. This is because filling in single augmentation result may not satisfyingly tackle the issue of imbalanced data. As a result, the research combined different augmentation outcomes and fill in the dataset. This method may not consider the situation where different outcomes may, in contrast, confuse the features of original image.

4. CONCLUSION

This research helps deal with the issue of imbalanced data in artificial intelligence assisted diagnosis application. By feeding training dataset the combined results of image data after applying different type of image data augmentation methods. It generated multiple new training datasets, and then test each of them. The result indicates that most of the offline augmentation, except sharpening, increase the majority of evaluation scores. Thus, it can be concluded that offline augmentation does bring beneficial outcomes for medical image data in COVID dataset. However, one factor that should consider is that different augmentation outcomes may make the model confuse features on original images. Later study could feed insufficient images from one augmentation outcome and exam the performance.

REFERENCES

[1] Ciotti, M., et al., "The COVID-19 pandemic," Critical reviews in clinical laboratory sciences, Critical reviews in clinical laboratory sciences 57.6 (2020): 365-388.

[2] Shen, D., et al., "Deep learning in medical image analysis," Annual review of biomedical engineering, 19 (2017): 221.

[3] Larrazabal, A. J., et al., "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," Proceedings of the National Academy of Sciences, 117.23 (2020): 12592-12594

[4] Han, C., et al., "Combining noise-to-image and image-to-image GANs: Brain MR image augmentation for tumor detection," Ieee Access, 7, 156966-156977, 2019.

[5] Chen, Y., et al., "Generative adversarial networks in medical image augmentation: a review," Computers in Biology and Medicine, 105382,2022.

[6] Kaggle, "Convid19 radiography database", https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database, 2022.

[7] Qiu, Y., et al. "Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV." China Communications 17.3 (2020): 46-57.

[8] Chen, S., et al. "An end-to-end approach to segmentation in medical images with CNN and posterior-CRF." Medical Image Analysis 76 (2022): 102311.

[9] Li, Z., et al. "TFCNs: A CNN-Transformer Hybrid Network for Medical Image Segmentation." arXiv preprint arXiv:2207.03450 (2022).

[10] Nirmala, K., et al., "Investigations of CNN for Medical Image Analysis for Illness Prediction." Computational Intelligence and Neuroscience 2022 (2022).