Classification of pulmonary lesions based on CNN and chest X-ray images

Wen Xue
School of Mathematics & Statistics, Southwest University, Chongqing 400715, China
pzx@swu.edu.cn

Abstract. This article mainly introduces the convolutional neural network (CNN) and uses CNN to realize the processing and classification prediction of chest X-ray images (CXR), to determine whether the lung has lesions, and finally the final AUC score of 0.85556 through CNN. In order to further improve the accuracy, after referring to many documents and considering the actual situation, I chose to perform principal component analysis (PCA) in the image preprocessing part, replace the random initial sample with the principal component initial sample, and replace the random initial kernel with the principal component initial kernel. To avoid staying in the local optimum when stochastic gradient descent finds the optimal kernel. The PCA + CNN model predicted an AUC score of 0.89333, an increase of 0.03777.

Key words: Pulmonary lesions; Chest imaging; Predict; AUC scores.

1. Introduction

1.1. Research background
Since 2020, coronavirus disease (Coronavirus) has spread throughout the country. The difference between coronavirus and other lung diseases was discussed. Compared with severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (mers), coronavirus has the characteristics of low mortality, strong concealment and strong transmission ability, so timely detection, diagnosis and treatment are particularly important. At the beginning of the epidemic, sars-cov-2 nucleic acid detection reagents were insufficient, so Chinese researchers strongly recommend CT imaging as the main basis for the diagnosis of coronavirus in the current situation, and chest imaging can show the progress of the disease. At present, 90%—95% of suspected coronavirus patients have chest CT, which has a high detection rate of viral pneumonia. In contrast, the quality of the existing sars-cov-2 nucleic acid detection kits varies greatly, the detection rate is low, and the number of people who need to be detected also exceeds the capacity of medical institutions [1]. Although various types of pneumonia have certain imaging features, coronavirus and other viral pneumonia, bacterial pneumonia and some lesions have some common imaging features [2].

At present, chest imaging is widely used, and in order to improve the accuracy of image analysis for more complex chest processing. In order to avoid doctors' subjective judgment errors caused by different chest X-ray images affected by patients' age, gender, smoking and other factors, now scientists can use computer to detect chest lesions through mature technology. This paper takes the chest X-ray (CXR)
image of Shenzhen hospital as an example to try to use computer classification technology to predict whether the lung lesions. If the final result is good, it can be applied to accurate diagnosis and treatment of chest X-ray image, brain image diagnosis and treatment and other digital diagnosis and treatment.

1.2. Research issues
This paper tries to solve the following problems
1. Feature extraction of CXR image in Shenzhen hospital;
2. How to use CNN training images in R to predict chest lesions;
3. Optimization: the kernel vectors of the convolutional neural network with PCA are no longer randomly initialized;
4. Other improved methods of CNN model. A total of 662 CXR images (< 1000) are collected in this model, which belongs to small data. Therefore, more attention should be paid to the number of cycles when using this model. Although the higher the number of cycles, the more accurate it is, it is inevitable that over fitting will occur.

1.3. Literature Review
Through this epidemic, some researchers think that we should focus on the results of sars-cov-2 nucleic acid detection reagent [2], and some researchers think that the results of chest imaging are more important [1]. Therefore, we should not only pay attention to the results of reagent detection or chest X-ray imaging, if we want to accurately judge whether a patient has a new type of coronary pneumonia, we should comprehensively consider the results of both. It is easy for medical institutions to make objective judgment when testing reagents. However, unlike testing reagents, the diagnosis after chest imaging mostly depends on the subjective judgment of doctors, so there are misdiagnosis and missed diagnosis. With the development of the times, the information technology and artificial intelligence have been greatly promoted, so there is a digital diagnosis and treatment technology. Researchers have developed support vector machine (SVM), recurrent neural network (RNN) and convolution algorithm through constant thinking. In the case of complex image classification, the CNN algorithm can be widely used in the pursuit of accuracy. After the first attempt, we found that the training time of CNN is too long [3-6]. In order to improve the traditional CNN model, we found two methods to improve the accuracy and reduce the training time: feature extraction based on principal component analysis (PCA) of initial kernel [7,8] and pooled feature learning based on principal component analysis (PCA) [9]. Because all principal components are orthogonal to each other, there is no redundant information. Each principal component is a linear combination of the original variables, which maximizes the information retained in the low dimensional space. This paper mainly uses the kernel PCA feature extraction, which has no requirement for the data in the original image, and can extract nonlinear features more effectively.

2. Theoretical framework

2.1. Basic model
In the CXR image feature extraction, the patient's image size, height, age, gender, nationality, occupation and other free variables cause the high variability of this group of images.

![Figure 1](image.png)

*Figure 1.* The results of dimensionality reduction techniques such as lung segmentation and bone shadow elimination from reference [10] (the middle is the final rendering)
(1) Convolutional neural network (CNN)
In the usual research, we often encounter the problem of studying multiple groups of images. Each block in the graph represents different layers of CNN, which are input layer, convolution layer, pooling layer, dropout layer and fully connected layer. After the last layer is processed, the model will automatically output classification results.

(2) Principal component analysis (PCA)
PCA has a large number of variables and complex relationship, which brings us the complexity of the analysis problem. Principal component analysis can synthesize multiple variables into a small number of variables (dimension reduction), and make these components represent the vast majority of the information of the original variables and can be unrelated to each other. From the literature review, it can be seen that the classical CNN model generates the initial weights randomly. Finally, although the gradient descent method can better find the appropriate weights, the gradient descent method can only search the local optimal, not the global optimal, and it takes a long time. Therefore, this paper uses the method of reference [7] to extract features based on PCA, so as to realize the initialization of multi-layer convolution kernel parameters, which avoids the high error rate and too many iterations when the convolution kernel parameters are randomly initialized.

(3) In the deep learning of stochastic gradient descent (SGD)
It is sometimes necessary to adjust the weights of the neural network to the original model to reduce the loss of accuracy, and then optimize the loss function through the optimization algorithm in order to find the optimal parameters and minimize the value of the loss function. In the optimization algorithm of solving machine learning parameters, the gradient descent algorithm is widely used. The random gradient descent is mostly used in classifier.

2.2. Basic assumptions of the model
1. All the images are natural chest X-ray images under natural conditions;
2. The effect of dimensionality reduction techniques such as lung segmentation and bone shadow elimination on CXR image is negligible.

2.3. Model test: ROC curve and AUC score
ROC curve: receiver operating characteristic curve, is a coordinate diagram analysis tool. The ROC curve is a threshold. If the threshold of the two classification problem is small, the real number of patients may be higher than the number of detected diseases when the threshold is large.

AUC score: the area of curved edge formed by ROC curve and cross axis. Obviously, the value of this area will not be greater than 1. Any adjacent two points and horizontal axis on ROC curve can form trapezoid, and AUC can be obtained by adding all such trapezoid areas. Generally speaking, the more training samples, the more different scores will be after the scores are determined as positive ones. In this way, the more points on ROC curve will be, and the estimated AUC will be more accurate. This idea is very similar to the differential method commonly used in calculus. The probability meaning that AUC can express can be calculated by probability.

Method to calculate AUC. Do n random tests. In each experiment, one positive sample and one negative sample are randomly sampled. When the score of positive sample predicted by the model is larger than that of the negative sample predicted by the model, the count will be added 1. The final count is n (n must be less than or equal to n), then the AUC is obtained by n/n. ROC and AUC are often used to compare models in machine learning.
3. Modeling

3.1. Model 1: Classic CNN architecture for chest X-ray image analysis

CXR image of Shenzhen Hospital (source: https://www.kaggle.com/yoctoman/shcxr-ung-mask/download), in the program, any image is only a set of RGB numbers in vector format.

CNN is a series of neural network layers, each layer converts one three-dimensional activation function to another through differentiable function.

![Figure 2. CNN structure chart in CXR image](image)

3.1.1. Filter. When convolution layer performs complex mathematical operations, in computer vision, the typical method of image processing is to use filter to convolute it to extract salient features. This is the first step in CNN. The input image applies filter logic to activate mapping or feature mapping, that is, after the input image, in the computer program, the function will be adjusted according to the threshold value shown by the function, which looks white, gray value. In this case, we extract part of the window (which can be regarded as the block moment dimension of the m matrix, m < n), and then multiply the gray value of each image by the gray value of the corresponding feature image, and add the results, and finally divide the results according to the total number of gray values.

![Figure 3. The schematic diagram of CNN in filtering](image)

3.1.2. Pooling and standardization. In this step, a window (2 or 3) will be selected and a step range of pixels will be selected (2 selected in this article). The window is crossed through the filtered image and the maximum value is taken for each window. The results of pooling and each pixel are selected and the linear rectification function relu is used to standardize. If any value is negative, set it to 0.

3.1.3. Voting and classification. The last layer is the full connection layer, which votes the group to determine the output classification results. The full join layer is only the combination matrix of all previous results, and the output is determined according to the highest voting category because it is the last layer.
3.2. Model 2: Classic CNN architecture for chest X-ray image analysis

According to the possible problems of traditional CNN for small data sets, PCA is applied to optimize the existing convolutional neural network, which mainly optimizes the initial random kernel vector to improve the limitation of gradient descent method, so as to get better kernel vector.

Relevant literature shows that there is a high degree of similarity between the learning results of PCA process and the learning results of self-coding neural network when the number of hidden layer neurons is limited. Therefore, according to this characteristic, we can use PCA to calculate many times to obtain all the required convolution kernel groups, and then realize the initialization of CNN convolution kernel parameters. The specific method of parameter initialization of CNN convolution kernel based on PCA algorithm is as follows:

After some special preprocessing of data, we can get the pixel value of a black and white image, of which 0 represents black and 1 represents white. After successfully converting to data, 75% of the data is divided into training set and 25% of the remaining is used as test set. At the same time, because CNN model in mxnet package can only deal with 4-dimensional x matrix. So in data preprocessing, we need to increase the 2-D image pixel matrix to 4-D. The difference between the classical convolutional neural network model and the convolutional neural network after PCA is that the act type = tanh "hyperbolic tangent function in the first layer and the second layer of convolutional layer of the classical convolutional neural network code generates the convolution layer. The PCA + CNN algorithm input the sample image data after the principal component analysis at the initial time, so the latter takes a little longer.

4. Relevant conclusions

Because of the importance of learning rate in neural network, we can not simply set the learning rate randomly. We can optimize the learning rate by parameter training. In this code, 0.02, 0.04, 0.06, 0.08, substituting mx.model.FeedFirward In the. Create function, assume that the momentum parameter is 0.9 (it will be trained later). Finally, we can draw the AUC score when we take different learning rates. Combined with the results in Figure 5, we take the learning rate as 0.06. Similarly, for the training of momentum parameters, the values of 0.5, 1 and 1.5 can be selected first, and then the training process is completed mx.model.FeedFirward The. Create function obtains the result and draws figure. Combined with figure, we can see that the momentum parameter is 1. Finally, the training set and test set are trained and evaluated. Train accuracy represents the number of iterations.
Figure 5. Schematic diagram of CNN preprocessing based on PCA(A); AUC of different learning rates (1); AUC of different momentum (2)(B); The ROC curve of CNN on the test set (top), where the horizontal axis is the specificity and the vertical axis is the sensitivity. The ROC curve of CNN + PCA on the test set (bottom), where the horizontal axis is the specificity and the vertical axis is the sensitivity (C).

When the number of iterations is 156, the training accuracy of the terminal output is stable at 0.79422. In comparison, when the number of iterations is 134, the training accuracy of the terminal output is stable at 0.85556. Similarly, when the number of iterations is 128, the training accuracy of the final output of CNN terminal combined with PCA is stable at 0.89333. For more intuitive comparison, this paper draws the ROC curve. By observing the ROC curve, we can know that the accuracy of PCA + CNN algorithm is higher than that of CNN algorithm.

4.1. Advantages of PCA + CNN model
(1) Considering the influence of various factors, the classical convolutional neural network model framework is established to extract the nonlinear features of the image, and the principal component analysis is added to extract the nonlinear features of the image more efficiently;
(2) Based on the results of the model, in order to better predict whether the lung lesions, the classical CNN model combined with principal component analysis (PCA) is used to change the original randomly initialized kernel vector into several initial kernel vectors with the largest variance contribution rate selected by PCA, so that the random gradient decreases and the weight of the neural network is adjusted. It has less precision loss and higher accuracy;
(3) In the preprocessing, the image function in ebimage package is used for gray processing, and the size function is used for image scaling (the original image pixel is 3000 * 3000, and the scaled pixel is 28 * 28), so as to reduce the number of iterations and improve efficiency between.

4.2. Disadvantages of PCA + CNN model
(1) It is known that there are missing variables in the existing CXR images. For example, if there is shadow in the lung, it may be ignored because of gray processing. When there is bone crack in the rib, it will also be ignored because of the elimination of bone shadow in the dimensionality reduction operation.
(2) Although it is not troublesome to analyze the lung lesions through the model, the accuracy rate is 86.5%, but in real life, the wrong classification will bring a lot of unnecessary trouble to doctors and patients, and even bring economic losses to the hospital.
4.3. Improvement of PCA + CNN model

In order to solve the above problems, the most reasonable method is to add more samples into the training set of the model, because the sample size used in this paper is 662, in fact, for the neural network, it is a small amount of data, and it is easy to over fit. However, if non hospital personnel have no other way to obtain the latest medical data except in other open source medical databases such as kaggle or GitHub. Therefore, in order to solve the shortcomings of this model, if the time is enough, we can even build a deep learning network, but it still requires too much equipment. Another solution is to build several convolutional neural networks based on PCA, and then sample new feature atlas from these neural networks and vectorize them into a new data set, in which each image feature comes from the output of a CNN, and then take the output data set as the input of the next PCA + CNN algorithm. Of course, it also takes a lot of time.

References

[1] Dai W, Zhang H, Yu J, et al. CT Imaging and Differential Diagnosis of COVID-19[J]. Canadian Association of Radiologists Journal, Vol.71, No.2(2020), 195–200.

[2] Zeng Y T, Song L, Zhang X C. The role of HRCT in the diagnosis of COVID-19[J]. ChineseJournal of Medical ence Research Management, Vol.33(2020): E015.

[3] Zhang, Q. H., and C. X. Wan. "Review of Convolutional Neural Networks." Journal of Zhongyuan University of Technology (2017).

[4] He Pengcheng. Improved convolutional neural network model and its application [D]. Dalian: Dalian University of Technology, 2015

[5] Li Yandong, Hao Zongbo, Lei hang. Review of convolutional neural networks [J]. Computer applications, 2016, 36 (9): 2508 – 2515 + 2565

[6] Wang Yuanyuan, Zhou Tao, Lu Huiling, et al. Computer aided diagnosis model of lung tumor based on integrated convolutional neural network [J]. Journal of Biomedical Engineering, 2017, 34 (4): 543-551

[7] Ma Yichao, Zhao Yunji, Zhang Xinliang. CNN handwritten digit recognition algorithm based on PCA initialization convolution kernel [J]. Computer engineering and applications, 2019,55 (13): 134 – 13

[8] Gao Xuwei. Research on feature extraction method of nuclear PCA and its application [D]. Jiangsu: Nanjing University of Aeronautics and Astronautics, 2009

[9] Liu Yanping. Feature extraction and camera estimation based on CNN convolutional neural network [J]. Electronic quality, 2018, (2): 19 – 23

[10] Stirenko S, Kochura Y, Alienin O, et al. Chest X-Ray Analysis of Tuberculosis by Deep Learning with Segmentation and Augmentation[C]. IEEE International Conference on Electronics and Nanotechnology, 2018.

[11] Xu Zhonghui, LV Weishuai. Image coloring based on convolutional neural network [J]. Application of electronic technology, 2018, 44 (10): 25 – 28

[12] Wang J, Luo C, Huang H, et al. Transferring Pre-Trained Deep CNNs for Remote Scene Classification with General Features Learned from Linear PCA Network[J].

[13] Pu Yaru. Polarimetric SAR image classification based on CNN feature learning and SVM [D]. Xi'an: Xi'an University of Electronic Science and technology, 2015

[14] Marcos L, Bichinho G L, Panizzi E A, et al. Classification of chronic obstructive pulmonary disease based on chest radiography[J]. Radiologia Brasileira, Vol.46, No.26(2013), 327–332.

[15] Emerman C L, Cydulka R K. Evaluation of high-yield criteria for chest radiography in acute exacerbation of chronic obstructive pulmonary disease[J]. Annals of Emergency Medicine, Vol.22, No.4(1993): 680–684