Attention Model based Deep Learning Architecture for EEG Seizure Detection

Yukun Wang
keithwongszcn@163.com

Abstract. Because voice data is similar to EEG data, they are both temporal and concurrent. In order to explore the general appropriateness of speech recognition model in EEG database, this paper introduces the attention model which has been very good in speech field in recent years, and carries out experiments on TESC data set. The attention model in this paper is based on the traditional HMM model, and Z module is added to introduce the attention mechanism. At the input end of the whole framework, PCANet model is introduced to solve the problem of too high dimension of TESC raw data. The convolution size used in PCANet is the most important group of super parameters. It is optimized by convex optimization alone. Many groups of training have been carried out on the horizontal and vertical dimensions, and the maximum local accuracy has been obtained. In order to improve the recall rate, the penalty of incorrect data classification has been greatly increased. In order to further improve the accuracy, the depth of PCANet model is increased. The input and output dimensions of PCANet model are adjusted to the same size and cascaded in multiple layers. The final accuracy rate was raised to 60%.

1. Introduction
EEG has been in development for 60 years and has a wealth of resources in multiple application areas. In the past two years, deep learning has also begun to influence the EEG recognition algorithm, bringing new ideas to this field. The open source epilepsy database TUH EEG Seizure Corpus (TESC) [1] uses a large number of manual annotations with high accuracy. Predecessors have used a variety of deep learning models on TESC data [2]. They discussed the solutions of HMM, LTSM, CNN and RCNN prediction on the dataset respectively, and achieved good results, but the model has high universality and does not perfectly fit the characteristics of EEG data, and there is room for further improvement.

This paper tests for models that are more in line with EEG data characteristics. The EEG signal is time-series, and its analysis is mainly based on the frequency domain and the time domain[3], similar to the speech recognition task. The current speech recognition technology is very mature, so consider introducing a mature model in the field of speech to improve the predictive rate of epilepsy. The Attention Model (AM) [4] has performed well in the field of speech recognition in the past two years and is the focus of this paper. It assigns a focus to the "field of view" of the data, which is equivalent to the mechanism of attention distribution when people observe something. Experiments have shown that this bionic mechanism is very effective. Due to the low input dimension of the AM model, dimensionality reduction is required, and the epilepsy data features are sparse and easy to process. In order to preserve the time sensitivity of the dimensionality reduction data, the data collected each time is spliced in parallel, and the dimensionality-independent model is used for dimensionality reduction. In the region of data dimension reducing, ImageNet[5], ResNet[6] and CNN have great performance, and ResNet has a quick training method[7]. Furthermore, PCA[8] is also a good way to do the job.
This paper uses PCANet [9], which uses PCA to tune the convolution kernel in the CNN model. Epileptic data and the abstraction of the model lead to difficulty in model optimization. In this paper, convex optimization is used to optimize the hyperparameter of the topology of the PCANet model, and raise the level of accuracy.

The goal of this paper is to demonstrate the superior performance of the speech recognition model in the field of epilepsy detection by the accuracy of prediction. The overall structure of the model will be explained in detail in the second part, which uses PCANet and AM for dimensionality reduction and prediction. The third part shows the results and analysis of the experiment and gives the final conclusion.

2. METHODS FOR DETECTING EPILEPSY USING PCANET AND AM

2.1. Using PCANet to process raw data

The raw data needs to extract the spatial features of its spatial scale before passing through the AM model. This method is based on Iyad Obeid and Joseph Picone's research on EEG signals in 2017 [10]. This paper uses PCANet to accomplish this task.

Before starting, you need to preprocess the raw data. Under the comparatively coarse-grained estimation, this paper assumes that the field of view of the epilepsy data is 10s and the resolution is 0.1s. Based on this assumption, we extract data features every 0.1s. The raw data has a sampling rate of 250 Hz, a resolution of 16 bits, and 22 signal channels. Parallel splicing of each 0.1s data (25 frames) yields a 22-bit bitmap format of 22*25 (Fig 1).

The structure of the PCANet model is shown in Fig 1. It is divided into three layers, and the input size is 22 * 25 * 4 (16 bit) matrix, and the size of the convolution kernel is set to 4 * 5. The first two layers use PCA to extract features from the image matrix. The number of feature vector extractions for each layer is 5, and 5 * 5 packets are obtained at the end of the second layer (the size of each packet is). In the third layer, each packet is binary processed and binarized hash coded, and five output blocks are obtained. Finally, the histogram statistical cascade output is performed, and the decimal feature value is obtained for the next stage prediction. The size and number of convolution kernels in the first and second stages are hyperparameters, and the tuning strategy is discussed in the third subsection.

2.2. Using AM model to get the final prediction

The attention in the AM model is the importance of Y (indicated by the probability distribution) in the case of a specific input X, hidden layer H. In the model (Fig 2), the degree of concentration is
represented by $z$, and its value is updated after each input. The importance is the similarity between $z$ and input $x$. Under the similarity algorithm, the difference from the currently expected feature is larger. The importance of the data is reduced, thereby screening out the desired features.

The AM model in this section uses a hidden layer and a concentration module. The final output uses the softmax function, and the internal logic is simple. As described in the previous section, the raw data is processed as 100 frames (0.1 s per frame) as input to the AM model, initializing $z_0$ to 1, and matching using the cosine similarity function. As time passes, the sampling interval of the data is set to 5 s (that is, after the current data processing ends, the time window is shifted back by 5 s as a whole), thereby traversing the entire time axis.

LTSM is a variant of RNN, which divides the signal from the previous stage into two parts: the gate signal and the information. The gate signal is used to detect the depth of information transmission and decide whether it needs to continue to propagate. The way is similar to the degree of concentration. Due to the memory effect, feature points that are far from the current time point will also affect the result. Therefore, consider introducing LTSM and implanting a concentration module (Fig 2), which can effectively prevent gradient explosion or gradient. The disappearance occurred.

2.3. Hyperparameters tuning

Deep learning models are difficult to debug due to the abstraction of data and the complexity of the model.[11] This paper describes the debugging scheme for the two sets of hyperparameters for the size of the convolution kernel of the PCANet model and the number of features of the first two layers.

In this paper, 5*5 is used as the central dimension of the convolution kernel. The size of the convolution kernel is added and subtracted by 1 in the x and y directions, and tested on a small test set. The correct rate is compared. The results are shown in Table 1. However, the training cost of this method is relatively high. This paper assumes that the process of finding the optimal convolution kernel size is a convex optimization process, so it is verified by the interval sampling approximation. First, keep $X$ at 5, test three groups of $Y$, get $Y_5$ as the peak, then keep $Y$ at 5, and $X_4$ is the peak. According to the assumption of convex optimization, 4,5 is the best convolution kernel size.

| Coordinate | X3 | X4 | X5 |
|------------|----|----|----|
| Y4         |    |    | 51%|
| Y5         | 50%| 56%| 53%|
| Y6         |    |    | 47%|
The role of the first two layers in PCANet is to expand the data and obtain its main vector. In this paper, the number of vectors in each group of data features occupying 90% is counted by single-step debugging, and it is used as the feature of the first layer. The number, the same reason can obtain the number of features of the second layer, this method can ensure the maximum efficiency of subsequent training.

| Coordinate | X3 | X4 | X5 |
|------------|----|----|----|
| Y4         | 51%|
| Y5         | 50%| 56%| 53%|
| Y6         | 47%|

In view of the fact that this article is mainly used for epilepsy detection, it is necessary not only to improve the accuracy rate, but also to reduce the false negative rate (False Negative Rate/FNR). There are a large number of redundant data containing rich information in the training set. In the process of training, not only high returns should be given to correct predictions, but also high penalty should be given to the prediction of errors. In this experiment, the penalty multiple of FN is set to 3, and the reward coefficient of TP is 2. This set of hyperparameters can also be measured using the above-mentioned method of finding the size of the convolution kernel. For details, see (Table 2).

2.4. Use deeper model to increase accuracy
The number of hidden layers determines the upper limit of the complexity that the model can express. The model depth of CNN/DNN/ResNet [12] is thousands of layers, which makes it perform well in specific tasks. This paper attempts to increase the number of hidden layers in the PCANet and AM models, respectively.

In PCANet, since the increase of the number of layers will bring about exponential complexity, consider combining multiple PCANet models (Fig 3). This paper attempts to connect the three models end to end, and the results show that the correct rate is Promotion.

Fig 3. Composite PCANet model

The influence of the hidden layer of the AM model on the complexity is linear. In this paper, the hidden layer is set to ten layers, and the connection between the second layer and the eighth layer is randomly established.

The training result of our final model is shown in Fig 4.
3. CONCLUSIONS
In this paper, PCANet is introduced as a model for data dimensionality reduction and is identified using the AM model. After super-parameter tuning and model optimization, a composite of a three-layer typical PCANet overlay model and an AM model containing ten hidden layers are finally used. In the process of training, the Dropout rate is set to 0.5, the convolution kernel size of PCANet is 4*5, and training for 200 rounds.

The final prediction rate is around 60%. This recognition rate demonstrates the reliability of the AM model in the EEG epilepsy recognition task, and also confirms that the speech recognition model and EEG are similar.

References
[1] Shah, V., von Weltin, E., Lopez, S., McHugh, J. R., Veloso, L., Golmohammadi, M., ... & Picone, J. (2018). The Temple University Hospital Seizure Detection Corpus. arXiv preprint arXiv:1801.08085
[2] Golmohammadi, M., Ziyabari, S., Shah, V., de Diego, S. L., Obeid, I., & Picone, J. (2017). Deep Architectures for Automated Seizure Detection in Scalp EEGs. arXiv preprint arXiv:1712.09776
[3] Shah, V., Golmohammadi, M., Ziyabari, S., Von Weltin, E., Obeid, I., & Picone, J. (2017, December). Optimizing channel selection for seizure detection. In Signal Processing in Medicine and Biology Symposium (SPMB), 2017 IEEE (pp. 1-5). IEEE
[4] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473
[5] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105)
[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016, October). Identity mappings in deep residual networks. In European conference on computer vision (pp. 630-645). Springer, Cham
[7] Solomatov, K., & Stepanov, D. (2017). Speeding up ResNet training
[8] Shlens, J. (2014). A tutorial on principal comonent analysis. arXiv preprint arXiv:1404.1100
[9] Chan, T. H., Jia, K., Gao, S., Lu, J., Zeng, Z., & Ma, Y. (2015). PCANet: A simple deep learning baseline for image classification?. IEEE Transactions on Image Processing, 24(12), 5017-5032
[10] Obeid, I., & Picone, J. (2017). Machine Learning Approaches to Automatic Interpretation of EEGs. Biomedical Signal Processing in Big Data
[11] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778)
[12] Schirrmeister, R., Gemein, L., Eggensperger, K., Hutter, F., & Ball, T. (2017, December). Deep learning with convolutional neural networks for decoding and visualization of EEG pathology. In Signal Processing in Medicine and Biology Symposium (SPMB), 2017 IEEE (pp. 1-7). IEEE