ISBN: 978-981-18-3959-7
2022 the 12th International Workshop on Computer Science and Engineering (WCSE 2022)
doi: 10.18178/wcse.2022.06.041

Alzheimer's Disease Detection through Spontaneous Speech Using Attention Augmented Convolutional Neural Network

Jiyun Li, Ya Hai, Chen Qian +

School of Computer Science and Technology Donghua University, Shanghai, China

Abstract. Alzheimer's disease (AD) is a neurodegenerative disease which affects patients' thinking, mood, and memory. Once diagnosed, it cannot be cured or reversed. Mild cognitive impairment (MCI) is the early stage of Alzheimer's disease, and medication at this stage can be used to slow down or even stop its development. A large number of studies have shown that AD can cause language barriers. There are significant symptoms in language, which can be used for early detection of AD. In this paper, convolutional neural network (CNN) is applied to the early diagnosis of AD. In addition, we introduced Convolutional Block Attention Module (CBAM) and incorporate CBAM into the CNN architecture to enhance the performance of the model. Experimental results show that the proposed model in this paper achieves 84.87% and 83.00% classification accuracy in long speech tracks and short speech tracks of the Alzheimer's Disease Recognition Competition, improved 5.07% and 9.00% compared to the baseline system.

Keywords: Alzheimer's disease, cognitive decline detection, convolutional neural network, convolutional block attention module.

1. Introduction

Alzheimer' s disease (AD) is a neurodegenerative disease and the most common cause of dementia, accounting for 50% to 70% of all types of dementia. With the globally aging population, conditions such as AD are likely to become more prevalent [1]. As the world's most populous country, China's elderly population has exceeded 200 million. In 2020, the number of people suffering from AD in China has reached 14.55 million [2]. There is a stage before the AD dementia stage called mild cognitive impairment (MCI). Early diagnosis of AD has become essential in disease management as it can offer interventions to slow or delay progression of symptoms [3]. Current diagnosis such as magnetic resonance imaging MRI, positron tomography PET, and cognitive assessment MMSE.

Research shown that one of the earliest symptoms of AD is speech impairment, including suffering anomia, reduced word comprehension, object naming problems, semantic paraphasia, and a reduction in vocabulary and verbal fluency [4], low speech rate and frequent hesitations at the phonetic and phonological level [5]. These early signs can be detected by having the patient perform a picture description task, such as the Cookie Theft task from the Boston Diagnostic Aphasia Examination [6]. Growing number of research has demonstrated that quantifiable indicators of cognitive decline associated with AD are detectable in spontaneous speech. A prime example of this is the ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) Challenge [7], aimed at generating systematic evidence for the use of such indicators in automated AD detection systems and towards their clinical implementation.

2. Related Work

In view of the excellent performance of convolutional neural networks in audio and music processing tasks such as speaker recognition, voice classification, and speech sentiment analysis [8]. Many researchers use the method of convolutional neural network to identify AD [9]. Among them, Karlekar et al. proposed a CNN-LSTM neural network language model and explored the effectiveness of part-of-speech tagging to improve the accuracy of AD patient classification. Their model achieved an accuracy of 91% [10]. Their

+ Corresponding author. Tel.: (86)021-6779-2382
E-mail address: chen.qian@dhu.edu.cn

method has a strong correlation between the transcription of the patient's conversation content and the language, and it is difficult to generalize to the research of other diseases or languages. Liu et al. [11] applied Automatic Speech Recognition (ASR) technology and hierarchical attention mechanism to neural networks to make them used for AD recognition tasks, and achieved an accuracy of 82.59% on the DementiaBank dataset. Chen et al. [12] proposed a network based on the attention mechanism. The network is composed of CNN and GRU modules. The special network structure enables the model to analyze local language patterns and overall macro language functions. The recognition accuracy rate of patients with disease is 97.42%.

3. Dataset

3.1. Overview of the Dataset

The dataset used in this paper is the official dataset provided by the Alzheimer's Disease Recognition Competition organized by the 16th National Conference on Man-Machine Speech Communication (NCMMSC2021). The dataset contains 280 Spontaneous Speech samples from 122 subjects. The duration of the voice samples is 15,073 seconds. Including 79 speech samples from 26 patients with Alzheimer's disease (AD), totalling 4,323 seconds (accounting for 28.7%); 93 speech samples from 52 patients with mild cognitive impairment (MCI), totalling 4901 seconds (accounting for 32.5%); 108 speech samples of 44 normal people (HC), with a total time of 5849 seconds (accounting for 38.9%).

3.2. Dataset Preprocessing

Since the voice samples in the dataset contain environmental noise, we used the open source audio processing tool Auditor to perform a unified noise reduction process on the voice material before conducting the experiment. The number of AD voice samples after deletion only have 32. We duplicate the AD voice samples to increase the proportion of AD samples in training. After preprocessing the dataset, the number of speech samples of AD, MCI, and HC are respectively 96, 93, and 96. The training set now contains 285 samples. We cut the speech samples of pre-processed training set into 6 seconds, in order to increase the number of samples. Taking into account the continuity between samples, we use 50% overlap to cut the samples. After cutting, the number of training set samples expanded to 4769, a total of 28614 seconds.

4. Methods

In this paper, we investigate the efficacy of convolutional neural network on AD detection. Inspired by the research of Chen et al. and the previous literature [13-18], the attention mechanism is introduced into the convolutional neural network, which is the channel attention mechanism and spatial attention mechanism. The introduction of the attention mechanism can not only tell us where to focus, but it can also increase the representativeness of interests.

4.1. CNN Structure

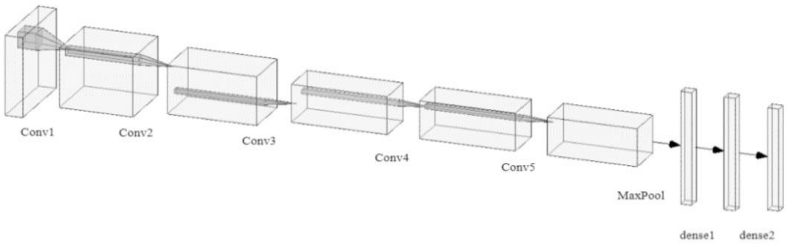

Fig. 1: Alexnet network structure diagram.

With the continuous development of convolutional neural network networks, many classic network architectures have emerged, such as LeNet, AlexNet [19], VGG-Net [20], ResNet [21]. The network model used in this paper is the 2DAlexNet model. Compared with the previous network structure, AlexNet uses CNN in a deeper and wider network, and its classification accuracy is higher. The basic structure diagram of

AlexNet's network is shown in Figure 1. AlexNet is composed of 5 convolutional layers and 3 fully connected layers, with a total depth of 8 layers.

4.2. CBAM

This network structure was proposed in 2018 [22]. Due to its excellent performance and strong adaptability, it has been widely used in neural network research in recent years. The structure diagram is shown in Figure 2. The CBAM structure includes an input module, a channel attention module, a spatial attention module, and an output module.

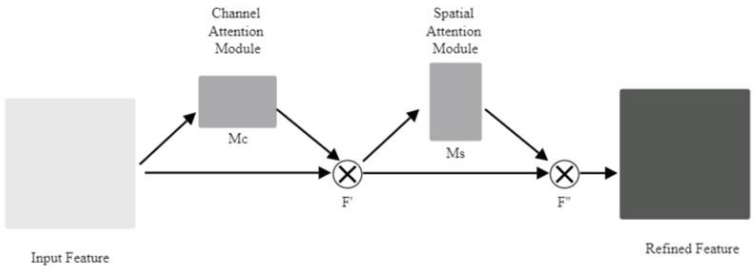

Fig. 2: Convolutional block attention module (CBAM).

The channel attention module integrates the information of each channel through average pooling and maximum pooling [23-24], and obtains two descriptors representing the average pooling feature and the maximum pooling feature. Pass the descriptors into the shared network to generate the channel attention feature map M_c. Finally, an element-wise multiplication is computed between F and M_c. The input feature F' of the spatial attention module is obtained.

For the spatial attention module, the average pooling and maximum pooling operations are performed along the channel axis, and then the pooled features are connected to generate a feature descriptor. The descriptor is convolved through the convolutional layer to generate a two-dimensional spatial attention map M_s, which emphasizes or suppresses features during the encoding process. The M_s obtained by the spatial attention module is multiplied by F' to obtain the final output F'' of CBAM.

4.3. Attention Augmented Convolutional Neural Network

This study uses a convolutional neural network for AD recognition based on speech signals. In order to allow CNN to focus on effective information points, we introduced the Convolutional Block Attention Module (CBAM). The overall architecture of the attention augmented convolutional neural network is shown in Figure 3. The backbone network includes 5 convolution modules, each of which contains a convolution layer, a ReLU function, and a maximum pooling layer. CBAM is embedded behind the first convolution module and the last convolution module to adaptively optimize features.

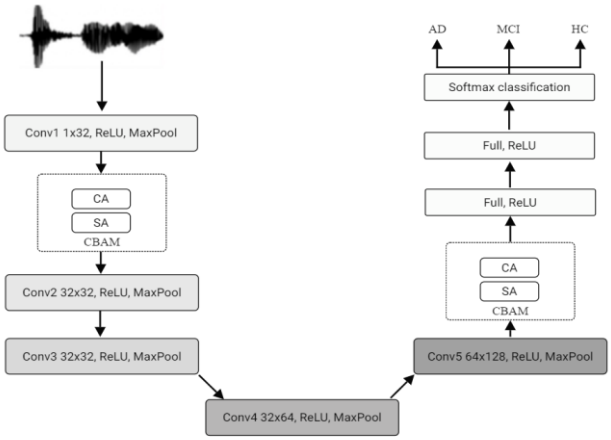

Fig. 3: Attention augmented convolutional neural network.

5. Experiments

5.1. Feature Extraction

This paper uses spectrogram (Spec), Melspectrogram (Melspec) and Mel-Frequency Cepstral Coefficients (MFCC) as the input features of the neural network. Spec [25] consists of the original audio through pre-emphasis, framing, windowing, and Short-Time Fourier Transform (STFT) and other operations are obtained. Mel spec [26] is also known as FBank. The spectrogram is composed of a series of Mel filters to form a mel spectrogram. Due to the overlap of adjacent filters, the characteristics of the Mel spectrogram are highly correlated. MFCC [27] is obtained through inverse discrete cosine transform (DCT) on the basis of Melspectrogram. It is a feature widely used in automatic speech recognition and speaker recognition.

5.2. Experimental Setting

In this paper, a 5-fold cross-validation method is used. Each model is trained and predicted 5 times, and the average accuracy of the output is used as the accuracy of the model. In the data preprocessing, in order to fully reflect the voice context information. We divide each 60 seconds audio into multiple 6 seconds audio, each with a 3 seconds overlap, and use the proposed model to recognize multiple 6 seconds audio. The result of the majority vote is used as the output of the speech recognition task. In training, the learning rate is set to 0.001, and the batch size is set to 32. The experimental results show that when the epoch is 20, the model achieves the best classification effect.

5.3. Experimental Results

(1) Experimental results of the training set

The experimental results of the cross-validation of the training set are shown in Table 1. Table 1 shows the classification effects of the three features of Spec, Melspec, and MFCC on the convolutional neural network.

Table 1: Experimental results of cross-validation of training set

Audio Length	Feature	Method	Accuracy
1 min (N=119)	Spec	AlexNet	0.801
		CBAM+AlexNet	0.842
	Melspec	AlexNet	0.883
		CBAM+AlexNet	0.931
	MFCC	AlexNet	0.793
		CBAM+AlexNet	0.814
6 sec (N = 1153)	Spec	AlexNet	0.794
		CBAM+AlexNet	0.840
	Melspec	AlexNet	0.879
		CBAM+AlexNet	0.926
	MFCC	AlexNet	0.782
		CBAM+AlexNet	0.811

(2) Experimental results of the test set

After using the best-performing settings in the training set, the model proposed in this article has achieved an accuracy of 84.8% on the long-speech track and 83.0% on the short-speech track on the official test set, which is significantly improved compared to the baseline. The experimental results of the model on the test set are shown in Table 2. Experimental results show that the best classification effect is achieved when using Melspectrogram as a feature for AD speech recognition. With the addition of the attention mechanism, the performance of the model has been further improved.

Table 2: Experimental results of the test set

Audio Length	Feature	Method	Accuracy	Recall	Precision	F1
1 min (N=119)	Baseline		0.799	0.785	0.786	0.798
	Spec	AlexNet	0.638	0.657	0.691	0.629
		CBAM+AlexNet	0.739	0.727	0.728	0.725
	Melspec	AlexNet	0.789	0.783	0.783	0.782
		CBAM+AlexNet	0.848	0.844	0.850	0.843
	MFCC	AlexNet	0.705	0.685	0.686	0.669
		CBAM+AlexNet	0.731	0.719	0.750	0.714
6 sec (N = 1153)	Baseline		0.740	0.737	0.723	0.718
	Spec	AlexNet	0.657	0.672	0.698	0.651
		CBAM+AlexNet	0.735	0.721	0.727	0.717
	Melspec	AlexNet	0.759	0.749	0.759	0.751
		CBAM+AlexNet	0.830	0.825	0.829	0.825
	MFCC	AlexNet	0.702	0.694	0.716	0.689
		CBAM+AlexNet	0.718	0.700	0.698	0.687

6. Conclusion

This paper proposes a convolutional neural network based on the attention mechanism. Apply it to the research of automatic recognition of Alzheimer's disease based on speech. This model has achieved an accuracy rate of 84.8% on the long-speech track and 83.0% on the short-speech track on the test set.

7. References

[1] R. Mayeux and Y. Stern, "Epidemiology of Alzheimer disease," *Cold Spring Harb. Perspect. Med.*, vol. 2, no. 8, Aug. 2012.

[2] Wang Y Q, Jia R X, Liang J H, et al. Dementia in China (2015–2050) estimated using the 1% population sampling survey in 2015 [J]. 2019, 19(11): 1096-100.

[3] J. Rasmussen and H. Langerman, "Alzheimer's disease - why we need early diagnosis," *Degener. Neurol. Neuromuscul. Dis*, vol. 9, pp. 123–130, Dec. 2019.

[4] K. A. Bayles and D. R. Boone, "The potential of language tasks for identifying senile dementia" *Journal of Speech and Hearing Disorders*, vol. 47, no. 2, pp. 210–217, 1982.

[5] Harold Goodglass and Edith Kaplan. *1983.Boston diagnostic examination for aphasia*, 2nd edition. Lea and Febiger, Philadelphia, Pennsylvania.

[6] S. de la Fuente Garcia, C. Ritchie, and S. Luz, "Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: A systematic review," *Journal of Alzheimer's Disease*, no. Preprint, pp. 1–27, 2020.

[7] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020.

[8] Abdoli, Sajjad et al. "End-to-end environmental sound classification using a 1D convolutional neural network" [J]. *Expert System with Application*, 2019, 136: 252-263.

[9] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Proc. of Neural Information Processing Systems (NIPS)*. (2012)

[10] Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018.Detecting linguistic characteristics of alzheimer'sdementia by interpreting neural models. In *Proceedings of the 2018 Conference of the North American Association for Computational Linguistics (NAACL2018)*

[11] Liu Zhaoci, Guo Zhiqiang, Ling Zhenhua, et al. Detecting Alzheimer's Disease from Speech Using Neural Networks with Bottleneck Features and Data Augmentation[C]// ICASSP 2021 - 2021 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*: IEEE, 2021: 7323-7327.

[12] J. Chen, J. Zhu, and J. Ye, "An attention-based hybrid network for automatic detection of Alzheimer's disease from narrative speech," *Proc. Interspace 2019,* pp. 4085–4089, 2019.

[13] Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention." advances in neural information processing systems. In: *Proc. of Neural Information Processing Systems (NIPS)*. (2014)

[14] Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recognition with visual attention. (2014)

[15] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. (2014)

[16] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. (2015)

[17] Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D.: Draw: A recurrent neural network for image generation. (2015)

[18] Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: *Proc. of Neural Information Processing Systems (NIPS)*. (2015)

[19] Kim, B. and S. Cho, Automated vision-based detection of cracks on concrete surfaces using a deep learning technique. *Sensors*, 2018. 18(10): p. 3452.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556, 2014.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "CBAM: Convolutional block attention module," in *Proc. of the European Conf. on Computer Vision*, Sep. 2018, pp. 3–19, Munich, Germany.

[23] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. *arXiv preprint arXiv*:1709.01507 (2017)

[24] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: *Computer Vision and Pattern Recognition (CVPR)*, 2016 IEEE Conference on, IEEE (2016) 2921–2929.

[25] D.H. Klatt and K.N. Stevens, "On the automatic recognition of continuous speech,Implications from spectrogram reading experiment," IEEE Trans. *Audio and Electroacoust.1973*, 21(3): 210 – 217.

[26] S. O. Arik, G. F. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," *CoRR*, vol. abs/1705.08947, 2017.

[27] MR.D.Mehendale. Speak identification [J]. *Signal&ImageProcessing: An International Journal*, 2011, 78(2): 62 – 69.