ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746295

END-TO-END NETWORK BASED ON TRANSFORMER FOR AUTOMATIC DETECTION OF COVID-19

Cong Cai[1,2], *Bin Liu*[1], *Jianhua Tao*[1,2,3], *Zhengkun Tian*[1,2], *Jiahao Lu*[1,4], *Kexin Wang*[1,2]

[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
[4]College of Computer and Information Engineering, Tianjin Normal University, Tianjin, China

ABSTRACT

The novel coronavirus disease (COVID-19) was declared a pandemic by the World Health Organization. The cumulative number of deaths is more than 4.8 million. Epidemiology experts concur that mass testing is essential for isolating infected individuals, contact tracing, and slowing the progression of the virus. In recent months, some machine learning methods have been proposed utilizing audio cues for COVID-19 detection. However, many works are based on hand-crafted features and deep features to detect COVID-19. There is no evidence that these features are optimal for COVID-19 detection. Therefore, we proposed an end-to-end network based on transformer for automatic detection of COVID-19. It directly learns features from the raw waveform for end-to-end learning, rather than extracting features in advance. We propose a feature extraction module to automatically extract features. And we use the transformer architectures to model the dependencies between the extracted features. It is the first end-to-end learning based on raw waveform for COVID-19 detection. Experiments on COUGHVID dataset show that our method has achieved competitive results.

Index Terms— COVID-19, Digital Health, Speech Processing, End-to-End, Deep Learning

1. INTRODUCTION

The novel coronavirus disease (COVID-19), declared a pandemic by the World Health Organization on March 11, 2020. The cumulative number of confirmed cases reported globally is now over 234 million and the cumulative number of deaths is more than 4.8 million as of 5 October 2021 Epidemiology experts concur that mass testing is essential for isolating infected individuals, contact tracing, and slowing the progression of the virus [1, 2]. Reverse transcription-polymerase chain reaction (RT-PCR) tests are the gold standard for detecting COVID-19 in clinical practice, due to their high sensitivity and specificity. However, they present several disadvantages: uncertain diagnosis time, expensive and collecting samples requires professionals. Therefore, an inexpensive, rapid, and extensible screening test for COVID-19 is necessary to help limit its spread.

Deep learning and machine learning technology provide solutions to this problem, which could be used to analyze cough sounds of infected patients and infer predictions [3]. In recent months, a variety of cough recording datasets have been collected by various groups and used to train machine learning models for COVID-19 detection. For example, Coughvid [4], Coswara [5] and Virufy [6]. Moreover, the INTERSPEECH 2021 holds two COVID-19 detection competitions: the INTERSPEECH 2021 Computational Paralinguists Challenge (ComParE) [7] COVID-19 Cough Sub-Challenges (CCS), and Diagnosing COVID-19 using acoustics (DiCOVA) [8]. Ruben et al. [9] leverage transfer learning to develop a set of three expert classifiers, and win the championship in the ComParE CCS. And Mahanta et al. [10] use the CNN network to classify COVID-19 positive and negative, and on top of the leaderboard of DiCOVA.

However, previous work is based on spectrogram or handcrafted features, there is no evidence that these features and transformations is optimal for COVID-19 detection. An alternative method is to directly model the raw waveform signals, which can automatic extract features and do end-to-end learning. This method has achieved outstanding performance in depression detection [11].

To solve the high-dimensional sparse problem, Bai et al. [12] proposed the temporal convolutional network (TCN). Its dilated convolution architecture can increase the receptive field exponentially to capture long-term dependence.

The most used architecture from the range of image processing is the convolutional neural network (CNN). Akman et al. [13] employ the ResNet to classify the COVID-19 positive and negative based on mel-spectrogram or Mel Frequency Cepstral Coefficients (MFCCs). The Transformer [14] has achieved impressive results in natural language processing. Recently, it has been shown that transformer can be a competitive approach to CNNs in the field of image processing [15]. In this paper, we evaluate the performance of the trans-

former architecture in the field of speech signal processing. Different from the mel-spectrogram and Vision Transformer (ViT) used in [16], we utilize the transformer architecture to model features extracted from the raw waveform.

Our main contributions can be summarized as follows: 1) We propose an end-to-end model to extract features from the raw waveform for COVID-19 classification. Compared with other hand-crafted features and deep features, the extracted features can play the role of transformer architectures to a greater extent. 2) We employ the transformer architectures to model the dependencies between feature sequences. We validate several different transformer classification methods in speech signals: class token and global average pooling. 3) The experimental results on the COUGHVID dataset demonstrate the effectiveness of our method.

2. DATASET AND PROCESS

To the best of our knowledge, as at the time of conducting this research, the largest publicly available cough dataset of COVID-19 patients was the COUGHVID dataset [4]. It includes a total of more than 20000 cough records, and the label includes participants' self-reports and expert annotation. In this paper, we utilise the COUGHVID dataset to validate our method. It contains three status: COVID, symptomatic, and healthy. We assign them into two groups, positive: COVID, negative: symptomatic and healthy.

First, We filter all audio recordings that have a degree of certainty below 0.9, and use the XGB model provided by official COUGHVID organizer to extract cough from audio recordings, namely voice activity detection. Due to the highly imbalanced dataset, we have applied data augmentation techniques to create a more balanced training dataset. By adding gaussian noise, shifting the time signal and stretching the time signal, we double the number of positive samples (only training set). And part of negative samples are randomly selected. To reduce the input dimension, we down sample the raw waveform to 16 kHZ for each audio recordings. We fix the audio length at 6 seconds, and the sample is padded with repeated versions of itself if length is less than 6s. Last, we divide dataset into three parts: training set, development set and test setthe amount of data is shown in Table 1. And oversampling technology is adopted to keep the number of positive and negative samples is the same during training.

Table 1. The number of the training set, development set and test set.

Partition	Positive	Negative	Total
Training	1000	3000	4000
Dev	100	300	400
Test	100	300	400
Total	1200	3600	4800

3. METHODOLOGY

As shown in Figure 1, our method consists of a one-dimension convolution, the feature extraction module, transformer, global average pooling and fully connected layers. First a one-dimension convolution is used to reduce the input dimension. The feature extraction module is used to extract features from raw features. And we employ the transformer to model the dependencies of extracted feature sequences. Finally, a global average pooling and two fully connected layers are used to predict final result.

3.1. Feature Extraction Module

The feature extraction module consists of N_1 dilated convolution blocks.

3.1.1. Dilated Convolution

The traditional convolution is difficult to process temporal sequence and capture long-term contextual information due to the limitation of convolution kernel size. So Yu et al. [17] proposed the dilated convolution, which can aggregate multiscale contextual information because of its exponentially growing receptive field. The filter of dilated convolution is applied over an area larger than its length by skipping input values with a certain step. It is equivalent to a convolution with a larger filter derived from the original filter by dilating it with zeros, but the dilated convolution is more efficient.

In this paper, we use the dilated convolution to make the receptive field expand exponentially and the number of parameters only increase linearly. The cough in patients with COVID-19 may last long, so dilated convolution is capable of aggregating contextual information related to cough at multiple scales.

3.1.2. Dilated Convolution Block

Motivated by the temporal convolution network (TCN) [12], we propose the dilated convolution block to extract features. Figure 1 shows the structure of dilated convolution block. First the dilated convolution is used, and the batch normalization and the nonlinear activation function ReLU are followed. Then a 1×1 convolution (pointwise convolution) changes the number of output channels to make it consistent with the input channels. Finally we use the residual path to speed up convergence and enable training of much deeper models. The output of each block is the input of the next block.

The dilation factor is doubled for every block up to a limit and then repeated: e.g. 1, 2, 4,..., 2^n,1, 2, 4,..., 2^n. The dilation factor increases exponentially to ensure to capture sufficiently large temporal contextual information related to COVID-19. It expands the network's receptive field and captures the COVID-19 cough features of the entire speech with fewer layers of stacking.

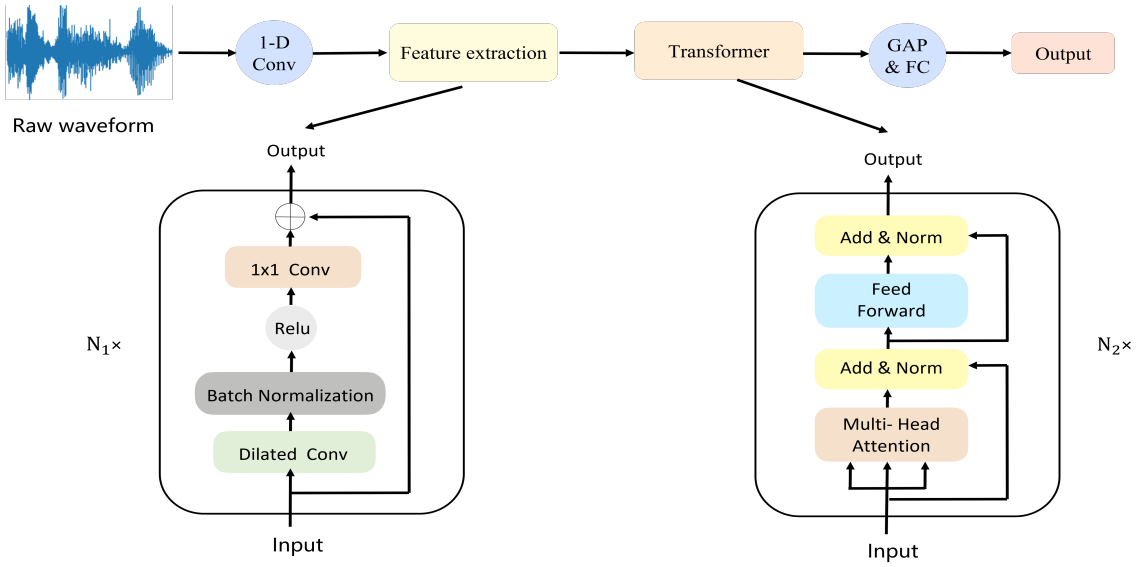

Fig. 1. Overall framework of proposed method. It contains a one-dimension convolution, the feature extraction module, transformer, global average pooling and fully connected layers.

3.2. Transformer

We use transformer [14] to model the context relationship between feature sequences. Specifically, we only use the encoder part of the transformer. It is based on multi head attention mechanism. Multi head attention operates multiple self attention operations in parallel. The formula is as follows:

$$\text{Attention}\,(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

where d_k is the key dimensionality. In self-attention, queries, keys and values come from the output of the previous layer. The multi-head attention mechanism obtains h different representations of (Q, K, V), computes scaled dot-product attention for each representation, concatenates the results, and projects the concatenation through a feed-forward layer. It can be defined as:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \qquad (2)$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \qquad (3)$$

Multi-head attention runs multiple self-attention operations in parallel, then subsequently fusing the individual embeddings into a single embedding. This style of architecture was first seen in form of the Transformer neural network [14], which relates a single sentence (sequences of word embedding) to it self.

Recently in the image processing, a new transformer architecture for classification has emerged [15]. Specifically, they use a learnable class token and the input feature sequences for multi-head attention calculation, and then use the output class token for subsequent classification. And Chu et al. [18] demonstrate that class token can be replaced by global average pooling (GAP). Therefore, we use the GAP following by the transformer.

And two fully connected layers are used to predict final results.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

We down sample the raw waveform to 16 kHZ and fix the audio length at 6 seconds for each audio recordings. Therefore, the input dimension of the model is 96000. To reduce the input dimension, there are 32 kernels in the first one-dimension convolution layer, whose size is 100 and stride is 50. For the first 1-dimension convolution, the size of the convolution kernel is 100 and the stride is 50. For feature extraction module, there are 24 dilated convolution blocks in total and the highest dilation factor is 32. The dilation factor is doubled for every block up to a limit and then repeated:1, 2, 4,..., 32,1, 2, 4,..., 32. And there are 64 convolution kernels with size of 3 and stride of 1 in the dilated convolution block. The last 1×1 convolutional layer has 32 kernels, whose size and stride are the same as the previous one. For the transformer network, we stack 6 encoded blocks, and the number of heads in the multi-head attention layer is 8, the number of nodes in the fully connected layer is 256.

To train the model, we use Adam optimizer [19] with learning rate of 0.0001 and batch size of 300, which is also dependent on the size of the input features. And our loss function is focal loss [20], which is proposed for the problem of unbalanced data. The training set is used to train model. The development set is used to adjust the experimental parameters and verify the effectiveness of each module in the model. The test set is used to compare our method with expert.

Table 2. Performance comparison between the features extracted by the proposed method and other features on the development set.

Features	AUC
Raw Waveform	72.1%
Wav2vec	78.6%
MFCC	81.5%
Proposed Method	**83.2%**

Table 3. Performance comparison of different transformer classification methods on the development set.

Metheds		AUC
ResNet-50		82.5%
Transformer	None	81.6%
	Class Token	82.8%
	GAP	**83.2%**

4.2. Comparison with Other Features

We compared the hand-crafted feature MFCC and the deep feature Wav2vec [21] to compare our methods. Specifically, these features are input into the transformer model for classification. The performance is shown in Table 2. It can be found that the effect of MFCC is better than wav2vec. It may be because the MFCC is designed from perceptual evidence. And Our method surpasses MFCC and achieves the best performance, which shows that it is feasible to classify COVID-19 by end-to-end learning features. This may be because the hand-crafted features may lose some useful information related to COVID-19 when transforming. And the feature of unsupervised representation learning, like Wav2vec, may not be suitable for COVID-19 classification.

In addition, we directly input the raw waveform to the transformer (reshape into two-dimensions), that is, without the feature extraction module. The experimental results show that our feature extraction module does learn useful knowledge from the raw waveform.

4.3. Transformer Classification Strategy

Recently, in the field of computer vision, transformer has different uses in classification. Dosovitskiy et al. [15] use a learnable class token and the input feature sequences for multi-head attention calculation, and then use the output class token for subsequent classification. And Chu et al. [18] believe that class token can be replaced by global average pooling (GAP). Therefore, We verified the class token, GAP, and direct use of two full connection layer classification. The results are shown in Table 3. The "None" means to directly use two full connected layers to classification. It can be found that there is little difference between using GAP and class token, but GAP is slightly better. Both GAP and class token significantly better than direct classification. This result is consistent with [18]. We infer that class token may need a larger dataset to perform well.

Moreover, we compared the performance of CNN. Specifically, we use ResNet-50 instead of transformer for classification. It can be found that the performance of CNN is not bad, but it is slightly inferior to transformer.

Table 4. Performance comparison between the proposed model and expert diagnosis on the test set.

Metheds	Specificity	Sensitivity	AUC
Expert	79%	25%	/
Proposed Method	**87%**	**63%**	**78.4%**

4.4. Comparison with Expert

Because COUGHVID dataset does not divide the training set and development set, some researches test performance in their own division set. Therefore, our method is not comparable with other work. Therefore, we compare it with the expert diagnosis in the COUGHVID dataset, and the results are shown in Table 4. Our method outperforms expert diagnosis in detecting COVID-19 from cough audio recordings.

5. CONCLUSIONS

Hand-crafted features and deep features may not be optimal for COVID-19 detection. Motivated by this speculation, we propose an end-to-end network based on transformer for automatic detection of COVID-19, which uses the feature extraction module based on temporal convolutional network to directly extract features from the raw waveform. The transformer is used to model the dependencies of extracted feature sequences. In addition, We verified the different methods used by transformer for classification. It is found that the GAP achieves the best performance. The experimental results on COUGHVID dataset show that our method has achieved promising performance in COVID-19 detection. In the future, we will evaluate our method on other COVID-19 datasets to improve the robustness and universality of our method.

6. ACKNOWLEDGEMENTS

This work is supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (No. 202200012), the National Key Research Development Plan of China (No.2017YFC0820602) and the National Natural Science Foundation of China (NSFC) (No.61831022, No.U21B2010, No.61901473, No.62101553).

7. REFERENCES

[1] Philip J Rosenthal, "The importance of diagnostic testing during a viral pandemic: early lessons from novel coronavirus disease (covid-19)," *Am J Trop Med Hyg*, vol. 102, no. 5, pp. 915, 2020.

[2] Marcel Salathé, Christian L Althaus, Richard Neher, Silvia Stringhini, Emma Hodcroft, Jacques Fellay, Marcel Zwahlen, et al., "Covid-19 epidemic in switzerland: on the importance of testing, contact tracing and isolation," *Swiss medical weekly*, vol. 150, no. 1112, 2020.

[3] Bless Lord Y Agbley, Jianping Li, Aminul Haq, Bernard Cobbinah, Delanyo Kulevome, Priscilla A Agbefu, and Bright Eleeza, "Wavelet-based cough signal decomposition for multimodal classification," in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, 2020, pp. 5–9.

[4] Lara Orlandic, Tomas Teijeiro, and David Atienza, "The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, no. 1, pp. 1–10, 2021.

[5] Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, P Kumar Ghosh, Sriram Ganapathy, et al., "Coswara–a database of breathing, cough, and voice sounds for covid-19 diagnosis," *arXiv preprint arXiv:2005.10548*, 2020.

[6] Gunvant Chaudhari, Xinyi Jiang, Ahmed Fakhry, Asriel Han, Jaclyn Xiao, Sabrina Shen, and Amil Khanzada, "Virufy: Global applicability of crowdsourced and clinical datasets for ai detection of covid-19 from cough," *arXiv preprint arXiv:2011.13320*, 2020.

[7] Björn W Schuller, Anton Batliner, Christian Bergler, Cecilia Mascolo, Jing Han, Iulia Lefter, Heysem Kaya, Shahin Amiriparian, Alice Baird, Lukas Stappen, et al., "The interspeech 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates," *arXiv preprint arXiv:2102.13468*, 2021.

[8] Ananya Muguli, Lancelot Pinto, Neeraj Sharma, Prashant Krishnan, Prasanta Kumar Ghosh, Rohit Kumar, Shrirama Bhat, S Raj Chetupalli, Sriram Ganapathy, Shreyas Ramoji, et al., "Dicova challenge: Dataset, task, and baseline system for covid-19 diagnosis using acoustics," *arXiv preprint arXiv:2103.09148*, 2021.

[9] Rubén Solera-Ureña, Catarina Botelho, Francisco Teixeira, Thomas Rolland, Alberto Abad, and Isabel Trancoso, "Transfer Learning-Based Cough Representations for Automatic Detection of COVID-19," in *Proc. Interspeech 2021*, 2021, pp. 436–440.

[10] Saranga Kingkor Mahanta, Shubham Jain, and Darsh Kaushik, "The brogrammers dicova 2021 challenge system report," Tech. Rep., Tech. rep., DiCOVA Challenge (March 2021). URL https://dicova2021. github ....

[11] Cong Cai, Mingyue Niu, Bin Liu, Jianhua Tao, and Xuefei Liu, "Tdca-net: Time-domain channel attention network for depression detection," *Proc. Interspeech 2021*, pp. 2511–2515, 2021.

[12] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[13] Alican Akman, Harry Coppock, Alexander Gaskell, Panagiotis Tzirakis, Lyn Jones, and Björn W Schuller, "Evaluating the covid-19 identification resnet (cider) on the interspeech covid-19 from audio challenges," *arXiv preprint arXiv:2107.14549*, 2021.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] Steffen Illium, Robert Müller, Andreas Sedlmeier, and Claudia-Linnhoff Popien, "Visual Transformers for Primates Classification and Covid Detection," in *Proc. Interspeech 2021*, 2021, pp. 451–455.

[17] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[18] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen, "Conditional positional encodings for vision transformers," *arXiv preprint arXiv:2102.10882*, 2021.

[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988.

[21] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.