Improved features using convolution-augmented transformers for keyword spotting

Yi Wang[1,*], *Junan* Yang[2], *Jingtao* Liu[1], *Qiang* Chen[1], and *Song* Li[1]

[1]Unit 91977 of PLA, Beijing, China
[2]College of Electronic Countermeasures, National University of Defence Technology, Hefei, China

Abstract. Transformer can effectively model long rang dependency, but suffer from uncapable to extract local feature patterns. While CNNs exploit local features effectively. In this paper, we seek to combine convolution and Transformers improves over using them individually, and propose improved features using convolution-augmented transformers for keyword spotting. The convolution-augmented transformers are constructed with a ResNet front-end and a convolution-augmented transformers back-end in series. Using this improved feature for keyword spotting task. The results show that the improved features using convolution- augmented transformers can yield at least 3% improvement compared with other features.

Keywords: Keyword spotting, Attention, Convolutional neural networks, Transformers.

1 Introduction

Compared with the large vocabulary continuous speech recognition (LVCSR) tasks, keyword spotting has the advantages such as insensitive to circumstance change, low system resources and faster speed in detecting certain desired words in continue speech. Hence it has been widely used in speech data mining and audio indexing applications.

Recently, one of the research hotspots and difficulties in keyword spotting is focus on developing new features [1]; the reason is the state-of-the-art (SOTA) speech features are sensitive to noise on the one hand and have poor classifying capability on the other [2]. To combat this issue, Deep Neural Network (DNNs) [3], Convolution Neural Networks (CNNs) [4], Recurrent Neural Networks (RNNs) [5] based features have seen large improvements in recent years.

However, features with DNNs or CNNs each has its limitations. While DNNs based features are good at collecting and associating information from neighborhood, they are less capable to extract local feature patterns. On the other hand, CNNs based features can effectively capture local information, but they need many more layers or parameters to capture global information [6].

In recent years, the evolution of network architectures in LVCSR has have enabled

* Corresponding author: wygggg@126.com

significant progress, where the prevalent architecture today is the Transformer instead of the DNNs [7,8,9]. Designed for sequence modeling and transduction tasks, the Transformer is notable for its use of attention to model long-range dependencies in the data. Its tremendous success in the language domain has led researchers to investigate its adaptation to other domains. Now Transformer are widely considered as a strong contender for unified architecture of machine learning

In this paper, we seek to combine convolution and Transformers improves over using them individually, and propose improved features using convolution-augmented transformers for keyword spotting. The convolution-augmented transformers are constructed with a ResNet front-end and a convolution-augmented transformers back-end in series which are sequentially trained. The front-end using long-term raw feature (e.g., 3-frame concatenating features) as input, and output the posterior probabilities feature. In back-end, the posterior probability features which estimated by the front-end are transformed into a low-dimensional representation. Experiments are conducted with TIMIT database and using a Point Process Model which is a lightweight keyword spotting paradigm as the baseline system. The results show that the improved features using convolution- augmented transformers outperforms other features.

2 Related works

2.1 CNNs [10]

CNNs have attracted extensive attentions in many artificial fields, such as Computer Vision (CV), Natural Language Processing (NLP) and Automatic Speech Recognition (ASR). The unique network structure of CNNs can effectively reduce the complexity of the feedback neural network. Generally, the basic structure of CNNs includes a convolutional layer, an activation layer, a pooling layer and a full connection layer.

2.2 Transformer [7,11]

Transformer was originally proposed for machine translation in 2017. Later works show that transformer-based models can achieve SOTA performances on various tasks. Transformer consists of an encoder and a decoder, each of which is a stack of *N* identical blocks. Each encoder block is mainly composed of a multi-head self-attention module (MHSA) and a position-wise feed-forward network.

2.3 Convolution-augmented Transformers (Conformer) [12]

In order to organically combine convolutions with self-attention in ASR models. A novel model named Convolution- augmented Transformers (Conformer) has been proposed recently. As shown in Figure 1, It is comprised of a feed-forward module, a MHSA module, a convolutional module, and a feed-forward module. This combination has been shown to improve ASR performance compared to the transformer architecture as it better captures temporal information locally and globally [13].

3 Convolution-augmented transformers based features extraction model

Inspired by Conformer and other work, in this paper, we presented Convolution-augmented

Transformers based features for keyword spotting. The model is shown in Fig.2, It is comprised of two components, the front-end and the back-end.

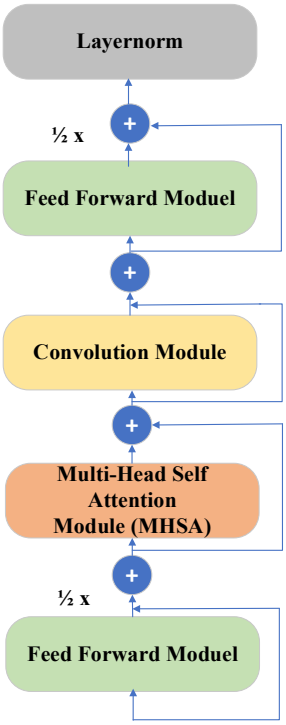

Fig. 1. Conformer model architecture.

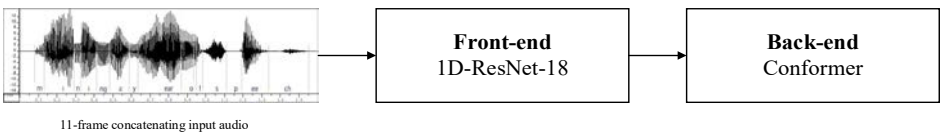

Fig. 2. Convolution-augmented Transformers based features extraction model.

The input audio we use in this paper is 11-frame concatenating, which are constructed by augmenting the current speech frame with its neighboring 3 frames within a context window (1+1+1). The reason is the gains of Transformers are mostly attributed to the feature vectors that are concatenated from a long temporal context

In the front-end block, we use a ResNet-18 based on 1D convolutional layers, where the filter size at the first convolutional layer is set to 80 (5ms). To down-sample the time-scale, the stride is set to 2 at every block.

In the back-end block, we use Conformer as the back-end for temporal modeling. It is comprised of a set of conformer blocks which is described in 2.3. In particular, the feed-forward module of conformer is composed of a d^{ff}-dimensional linear layer, followed by ReLU. The MHSA module receives queries Q, keys K, and values V as inputs. The matrix of outputs at i-th head self-attention is computed through Scaled Dot-Product Attention. The convolutional module contains a point-wise convolutional layer with an expansion factor of 2, followed by GLU, a temporal depth-wise convolutional layer, a batch normalization layer, a swish activation layer, a point-wise convolutional layer, and a layer normalization layer.

Compared with other speech features, Convolution-augmented Transformers based features has the following advantages: firstly, they do not require strong assumptions on data distributaries which make them can simply concatenate different distributaries features together; secondly, when trained on large amount of data, they are invariant to speaker characteristics and environment specific information such as noise; thirdly, they are able to learn both position-wise local features, and use content-based global interactions.; Last but not the least, the proposed features is low-dimensional and embedded with classification information.

4 Point process model

Point process model is a lightweight keyword spotting approach which operates within the sliding model. In this paper, we will use point process model to verify the practicality and reliability of convolution-augmented transformers-based features. Experimental results in [14, 15] have already proved that point process model has the capacity to generalize from a relatively small numbers of training examples and avoid the local optima of HMMs, the accuracy levels are comparable with other keyword spotting systems.

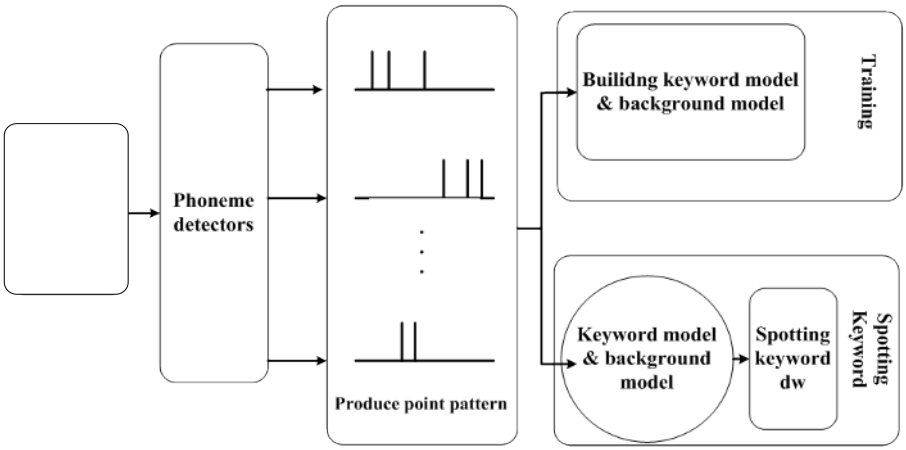

Fig. 3. Framework of point process keyword spotting system model.

The framework of point process model is shown in Fig.3.

5 Experiments and results

5.1 Dataset

The efficacy of convolution-augmented transformers-based features is evaluated by performing keyword spotting experiments on TIMIT speech corpora. TIMIT database consists of 4.3 hours of read speech. The training set has 4620 sentences collected from 462 speakers while testing set has 1620 sentences collected from 162 speakers, and there is no same speaker between the two sets. We analyse the input speech using a 25-ms Hamming window with 10-ms between the left edges of successive frames. And the features for each frame are 43-dimensional as we discussed in 2.2. The data were normalized to have zero mean and unit variance over the entire corpus.

5.2 Experimental Setup

Experiment compares Convolution-augmented Transformers based features with original MFCC, single MLP, single DNN, hierarchical MLPs based BN feature and Hierarchical DNNs based BN features to verify the improved feature is able to increase the keyword spotting accuracies.

5.3 Experimental Results

We adopt a ResNet-18 based on 1D convolutional layers for Front-end. The architecture of Front-end is shown in Table 1.

Table 1. The architecture of Front-end.

Conv1	Conv1d, 80, 64, stride 2
Res2	$\begin{bmatrix} \text{conv1d}, 3, 64 \\ \text{conv1d}, 3, 64 \end{bmatrix} \times 2$
Res3	$\begin{bmatrix} \text{conv1d}, 3, 64 \\ \text{conv1d}, 3, 64 \end{bmatrix} \times 2$
Res4	$\begin{bmatrix} \text{conv1d}, 3, 64 \\ \text{conv1d}, 3, 64 \end{bmatrix} \times 2$
Res5	$\begin{bmatrix} \text{conv1d}, 3, 64 \\ \text{conv1d}, 3, 64 \end{bmatrix} \times 2$
Pool6	Average pooling, stride 20

For the Back-end, we use 10M params, the model hyper-parameters are shown in Table 2.

Table 2. The model hyper-parameters of Back-end.

Num Params (M)	10.3
Encoder Layers	16
Encoder Dim	144
Attention Heads	4
Conv Kernel Size	32

For comparison, the structure of DNN is with 6 layer (with 4 hidden layer) DNN to build up the first DBN of the hierarchical DBNs, then the structure of the first DBN can be displayer as 129- [2048-2048-2048-2048]-129; the second DBN employs a 7 layer (with 5 hidden layer) DBN which the topological structure is 129-[2048-1048-43-1048- 2048]-129. the single MLP and single DBN is set to same as the second DBN of our hierarchical DBN. The hierarchical MLPs are set to just the same as the hierarchical DBNs. The number of Gaussian component in GMM based point process model is set to 8. All 4620 sentences in TIMIT training database are used in this experiment. Table 3 shows the Figure of Merits (FOM) of the 6 different features in keyword spotting.

Table 3. Performance Comparison between different features.

Features	**FOM (%)**
Original MFCC	92.17
Single MLP based BN features	92.66
Hierarchical MLPs based BN features	93.87
Single DNN based BN features	93.17
Hierarchical DNNs based BN features	95.55
Convolution-augmented Transformers based features	98.73

The results in Table3 demonstrate that the FOM of Convolution-augmented Transformers based features is at least 3% better than the other features.

6 Conclusion

In this paper, we propose convolution-augmented transformers based feature, and use this improved feature for keyword spotting task. The convolution-augmented transformers are constructed with a ResNet front-end and a convolution-augmented transformers back-end in series. Experiments are conducted with TIMIT database and using a Point Process Model as the baseline system. The results show that the improved features using convolution-augmented transformers outperforms other features. They can yield at least 3% improvement compared with other features.

References

1. M. Picheny, D. Nahamoo, V. Goel, B. Kingsbury, B. Ramabhadran and S. J. Rennie, Trends and Advances in Speech Recognition, IBM Journal of Research and Development, 55 ,2011, pp. 2: 1-2: 18.
2. H. Yang, S. Sharma, S. van Vuuren, and H. Hermansky, Relevance of time-frequency features for phonetic and speaker-channel classification, Speech Communication, 31 ,2000, pp.35-50.
3. Y. Wang, J Yang; H. Liu, Improved Bottleneck Feature using Hierarchical Deep Belief Networks for Keyword Spotting in Continues Speech, International Journal of Signal Processing, Image Processing and Pattern Recognition, 6 ,2013,pp. 375-386.
4. D. Horii; A. Ito; T. Nose, Analysis of Feature Extraction by Convolutional Neural Network for Speech Emotion Recognition, 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), 2021, pp. 425-426.
5. J. Yi, H. Ni, Z. Wen and J. Tao, Improving BLSTM RNN based Mandarin speech recognition using accent dependent bottleneck features, 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016, pp.1-5.
6. A. Srinivas, T. -Y. Lin, N. Parmar, J. Shlens, P. Abbeel and A. Vaswani, Bottleneck Transformers for Visual Recognition, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp.16514-16524.
7. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, 1(2017) 5998–6008.
8. L. Dong, S. Xu and B. Xu, Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp.5884-5888.
9. N. S. Mamatov, N. A. Niyozmatova, S. S. Abdullaev, A. N. Samijonov and K. K. Erejepov, Speech Recognition Based on Transformer Neural Networks, 2021 International Conference on Information Science and Communications Technologies (ICISCT), 2021, pp.1-5.
10. S. Kong, M. Kim, L. M. Hoang, and E. Kim, Automatic LPI radar waveform recognition using CNN, IEEE Access, 2018, pp.4207–4219.
11. T. Lin, Y. Wang, X. Liu, X. Qiu, A Survey of Transformers, (2021). arXiv:2106.04554

12. P. Ma, S. Petridis and M. Pantic, "End-To-End Audio-Visual Speech Recognition with Conformers," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 7613-7617.
13. A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, et al., "Conformer: Convolution-augmented transformer for speech recognition," in Interspeech, 2020, pp. 5036–5040.
14. Jansen A and Niyogi P, "Point Process Models for Spotting Keywords in Continuous Speech," IEEE Transaction on Audio, Speech, and Language Processing, vol. 17, no. 8, (2009), pp. 1457-1470.
15. Jansen A, "Point Process Models for Event-Based Speech Recognition," Speech Communication, vol. 51, no. 12, 2009, pp. 1155-1168.