UDC 621.391

DESIGN OF SPEECH RECOGNITION SYSTEM BASED ON ATTENTION MECHANISM

YALU GAO

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus

Received March 20, 2023

Abstract. Attention mechanism is to let the machine pay attention to more key information and ignore the secondary information in complex tasks, that is, to assign different weights to different information to represent different degrees of attention so as to reduce the amount of calculation. The common end-to-end speech recognition model structure is to directly model speech and text, which not only simplifies the speech recognition model but also improves the recognition performance. In order to study the function of attention mechanism in the end-to-end speech recognition system, this paper will take wenet as the basic network framework, and study the conformer part of the encoder and the transformer structure of the decoder in the end-to-end basic framework. These two structures make full use of the attention mechanism.

Keywords: speech recognition, attention mechanism, wenet network, transformer structure, conformer structure.

Introduction

This design will use part of the data in aishell as the data set to train the speech data. By constantly modifying the network framework and changing the attention model, we can get a better recognition rate, so as to get the most appropriate network parameters and achieve a good speech recognition function. In addition, we compare the attention mechanism with other networks, such as CNN, CTC, etc. Through this comparative experiment, we get the shortcomings of attention mechanism itself and its compensation method, such as the combination of attention mechanism and DNN, which makes up for its inability to focus on local features, and the combination of CTC can be used for streaming speech recognition [1].

The main algorithm

Attention mechanism: The attention mechanism is divided into three categories: the first category is the soft attention mechanism, which assigns weight to each input item. The weight is between 0 and 1, because it will be considered, so the amount of calculation is relatively large. The second category is the hard attention mechanism. The weight distribution of each input item is either 0 or 1, so some of them will not pay attention. The advantage is the amount of calculation small, with the disadvantage of possible loss of information. The third type is the self-attention mechanism, and its weight distribution is mainly the input item. We mainly use the self-attention mechanism.

Conformers and Transformers: In the Wenet used in this design, we mainly use two kinds of networks, Transformer and Conformer network [2]. These two networks are typical applications of the attention mechanism, in the part of the encoder network, both can be used. You can choose according to the specific situation, and in the decoder network. For the network part, only the Transformer network can be used. Transformer is composed of multiple Transformer Block groups. As a result, self-attention, res, relu, Feed-Forward layers will be used in each block, as shown in Figure 1. Conformer is composed of multiple Conformer Block composition, each block will use convolutional layer, self-attention layer, residual res, activation relu, Feed-Forward layers [3].

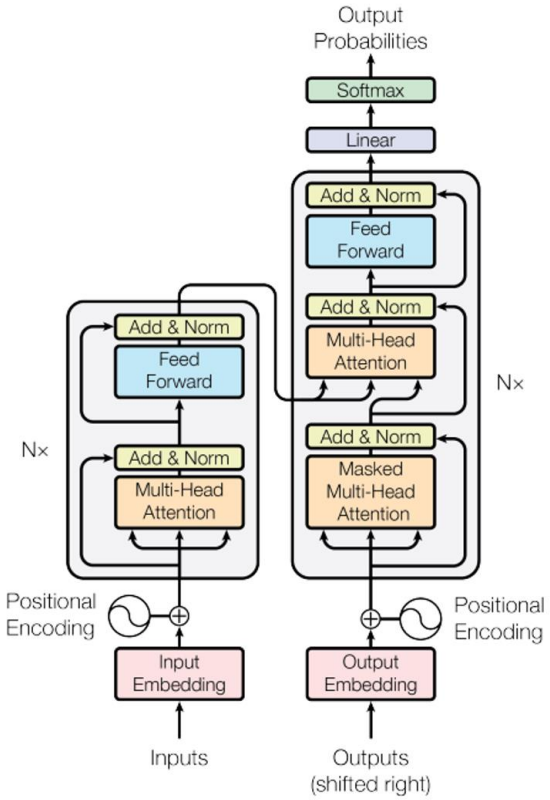

Figure 1. Transformer

Wenet network structure

In order to perform both non-streaming scene recognition and streaming scene speech recognition, Wenet using the U2 model, which is a joint model of CTC and AED. It is composed of Shared Encoder, CTC decoder and Attention decoder [4]. A Shared Encoder can select Transformer or Conformer, the attention decoder is composed of transformer layers, CTC decoder is composed of a Fully Connected Layer and a Softmax Layer. While training at this time, by using the dynamic chunk training method, the Shared Encoder can handle different lengths voice clips. When recognizing, the first step is to go through the CTC decoder to get multiple backups with the highest score. Use the results, and then use the Attention decoder to re-score the candidate results, and choose the highest score final result. When the selected voice segment is infinitely long, it is suitable for non-streaming scenarios that is to get decoding after a complete speech segment, it is also suitable for streaming languages when the selected speech segment is of finite size sound recognition scene [5]. The shared Encoder implements incremental forward operation, and the result of the CTC decoder is also displayed as intermediate results. During decoding, the CTC decoder operates in steaming mode in the first pass, while the attention decoder is used in the second pass to give more accurate results as shown in Figure 2.

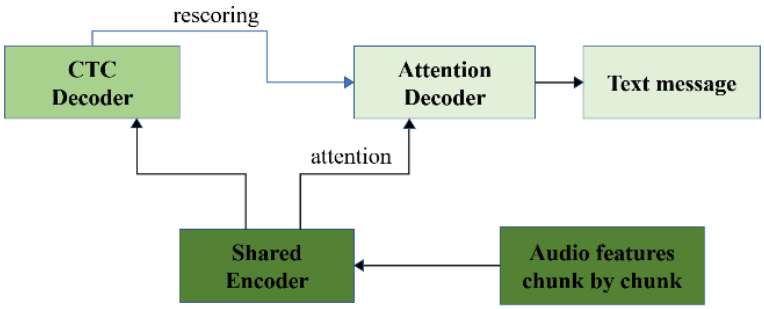

Figure 2. Structure of U2

Comparison of the effects of different decoding methods

This experiment explores two different decoder methods in the non-streaming model, focusing on comparing the real-time rate (RTF) between attention decoder and attention resoring, and the results of the four decoding methods. The full context is used in the model, and the conformer is used for training with a standard convolution kernel size of 15. In AED decoder, $\text{beam} = 10$ is used for decoding; CTC uses prefix beam search to form top n_best hypotheses as reference information for final re-scoring. "/" is used to represent that means it's not important in the process.

Table 1. **Comparison results of different decoding methods**

Decoding method	CTC weight	RTF	CER
Attention decoder	/	0,297	6,02
CTC prefix beam search	/	0,0	6,05
Attention rescoring	0,0	/	5,73
Attention rescoring	0,5	0,182	5,52

Conclusion

The first set of experiments: attention decoder, as shown in Table1. Although the Attention model works well, it still has its own problems. Questions are as follows: 1. Suitable for phrase recognition, poor for long sentence recognition; 2. Training is unstable when noisy data. The second group of experiments: CTC prefix beam search, where the real-time rate of CTC is not the focus of this experiment, so "/" is used to represent it. The third and fourth sets of experiments: attention rescoring, after CTC prefix beam search and attention rescoring, it is found that CTC prefix beam search produces many errors, which can be corrected by the attention rescoring strategy. However, some CTC decoding correct results will also be corrected into errors after attention rescoring, which means that CTC plays an important role in some cases. Therefore, a CTC weight score needs to be added. In addition, tested different CTC weight scores from 0,1 to 0,9, and found that when $\lambda = 0,5,$ it is the most stable.

References

1. Yan Z.J., Huo Q, Xu J. // A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR[C]. Interspeech 2013. P. 104–108.
2. Yao Z., Wu D., Wang X., [et al]. // Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit[J]. arXiv preprint arXiv. 2021.
3. Vaswani A., Shazeer N., Parmar N., [et al]. // Attention is all you need[J]. Advances in neural information processing systems. 2017. P. 30.
4. Tian Z., Yi J., Tao J., [et al]. // Self-attention transducers for end-to-end speech recognition[J]. arXiv preprint arXiv. 2019.
5. Hannun A. // Sequence modeling with ctc[J]. Distill. 2017. P. 8.