Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.070305
ISSN 2371-8838 Vol. 7 Num. 3

Semi-supervised End-to-end Speech Recognition

Jiewen Ning, Yugang Dai*, Guanyu Li, Sirui Li, Senyan Li

Northwest Minzu University, Lanzhou, Gansu, 730000, China
**Corresponding author*

Keywords: End to end, Automatic Speech Recognition, semi-supervised, shared encoder, conformer

Abstract: The current popular end-to-end speech recognition technology requires a large amount of labeled data, which limits the development of speech recognition technology in low resource languages. In this paper, we propose speech recognition schemes based on semi-supervised learning methods. We try to influence unlabeled speech-to-text mapping by learning unlabeled text-to-text mapping using a shared encoder. The proposed scheme uses Conformer network as a shared encoder, which can extract both text features and speech features. Using CTC/Attention network as decoder, the model is iteratively self-trained using a small amount of labeled data and a large amount of unlabeled data. The WER of the rescoring results on the Aishell-1 dataset was 12.54, a 37% reduction in WER compared with the baseline system. Compared with the state of art supervised system, we use less labeled data and only improve the error rate by 56%, which shows that our method has great potential in semi-supervised speech recognition.

1. Introduction

With the development of transformer [1], conformer [2] and other frameworks to the field of automatic speech recognition (ASR), the accuracy of speech recognition has been greatly improved. End-to-end (e2e) ASR integrates the three modules of pronunciation dictionary, language model and acoustic model in traditional speech recognition, and can directly map the input speech into character sequences, which simplifies the traditional ASR recognition process.

At present, the popular e2e speech ideas are connectionist temporal classification (CTC) [3, 4], attention based encoder decoder (AED) [5], and transducers [6]. However end-to-end ASR usually needs a large amount of data for training, but there are many problems in the data that can be used for model training, such as the existing data is large but the amount of sample data that can be used for training is very small, or the sample data used for training requires material resources for manual tagging, and the level of manual labeling is uneven.

In order to solve these problems, many researchers hope to make effective use of massive data through data expansion, such as self-supervised learning, unsupervised learning, semi-supervised learning and so on. Chapelle [7] proposed semi-supervised learning (SSL). SSL divides the data set into two parts, one is the labeled dataset; the other is the unlabeled dataset. With the rapid development of artificial intelligence, many researchers use the SSL method to study the ASR task. Ouali [8] and others think that the SSL method can make the prediction result of the prediction function generated by the unlabeled dataset more accurate than the prediction function generated by the labeled dataset. Higuchi [9] uses the pseudo tags generated by unlabeled data and the recognition accuracy of the model is improved to a certain extent. Zhang [10] uses pre-training model, self-training and large-scale model parameters for semi-supervised training, and the performance of the model has been improved to a certain extent.

In this paper, the semi-supervised training method of end-to-end ASR is different from the traditional SSL method. The unlabeled dataset is composed of unlabeled text and unlabeled audio dataset. The purpose of this paper is to explore the influence of unlabeled text feature learning on unlabeled audio feature learning, and then improve the performance of ASR model. Inspired by Karita [11], the shared encoder in this paper adopts a conformer structure composed of self-attention and convolution network, which greatly improves the ability of local correlation and global information capture in model training. The conformer shared encoder can map voice and text respectively to get the corresponding advanced features.

The semi-supervised training method in this paper is as follows: first, a small amount of labeled data is used for encoder-decoder baseline model training; then a small amount of labeled data and a large amount of unlabeled data (voice and text) are used for iterative training; finally, the advanced features of the text are sent to the decoder to decode and get the corresponding character sequence.

2. Baseline Model

The baseline model is an encode-deocder architecture, which adopts multi-target training and joint decoding methods at the same time. As shown in Figure 1.

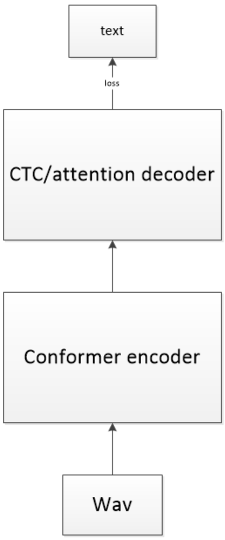

Figure 1: Speech recognition baseline model

The encoder is conformer and the decoder is CTC/transfomrer decoder. Among them, the conformer encoder maps the input sequence to the high-level-feature. The CTC/attention decoder decodes the high-level-feature h into a character sequence. The training of the baseline model is as follows:

$$h = ConformerEncoder(x) \tag{1}$$

Joint optimization goals, as follows:

$$P(y\%|h) = \partial L_{attention} + (1 - \partial) L_{ctc} \tag{2}$$

The conformer encoder has N blocks and the CTC decoder is composed of linear layer and softmax layer, and the transformer decoder is composed of self-attention layer and residual module.

3. Semi-supervised Speech Recognition Framework

3.1. Conformer Shared Encoder

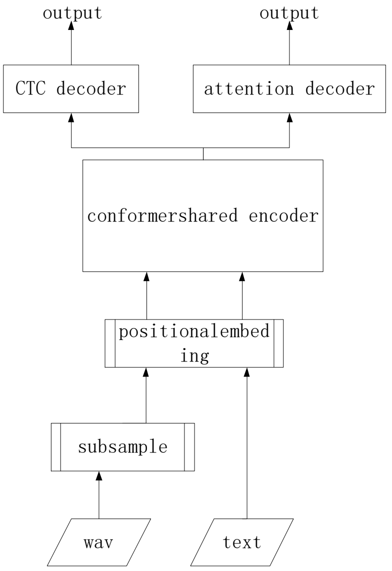

Figure 2: Semi-supervised speech recognition system.

Inspired by the SSL method of the text-speech automatic encoder in [12], the semi-supervised speech recognition system uses conformer as the shared encoder, making use of the local self-attention in the conformer structure and the global dependence of CNN, so that the shared encoder can map the features of speech and text at the same time. The semi-supervised speech recognition system is shown in Figure 2.

Because the input data types are different and conformer completely abandons the sequential LSTM network and depends on convolution and attention mechanism, the input data needs to be processed before it can be sent to the conformer shared encoder.

In order to obtain the position information between words in the text, position coding [13] is used. The text vector is first fed into the embedded layer for word vector random initialization, and the word vector with dictionary size and specified dimension is obtained after embedding. In order to reduce the redundant information of speech features, the speech is sent to the lower sampling layer and convoluted twice, and finally the time dimension and feature dimension of the speech are changed to 1/4 of the original.

3.2. Semi-supervised Training

The semi-supervised speech recognition model is shown in training algorithm 1.

The semi-supervised speech recognition system jointly trains the speech-text automatic encoder and the text-text automatic encoder, so the function loss is composed of three parts:

(1) Supervised loss

For labeled sets, the supervised loss is represented by the negative logarithmic likelihood between the shared encoder h and the real text, as shown below:

$$L_{\sup}(D_{\sup}) = -\log P_{\alpha}(Y_n \mid h_n) \tag{3}$$

(2) Text reconstruction loss

The process of text reconstruction is the process of decoding the intermediate representation of the text as the input of the decoder, which is a predictive character, and the loss is a negative logarithmic calculation formula as shown:

$$L_{text}(T_i) = -\log \Pr(\tilde{y} | h) \tag{4}$$

(3) Intermediate representation loss

It represents the intermediate representation loss of encoded speech and text. In this paper, KullbackLeibler divergence (KL) is used to measure the dissimilarity between two vector distributions.

Let P and Q be the z-dimensional Gaussian distribution of encoded speech and encoded text, respectively.

$$P = Normal(\mu_P, \Sigma_P) \tag{5}$$

$$Q = Normal(\mu_Q, \Sigma_Q) \tag{6}$$

$\mu_P = E_{x\sim S}[e(x)], \Sigma_P = Cov_{x\sim S}[e(x)],\ \mu_Q = E_{y\sim T}[e(y)], \Sigma_P = Cov_{y\sim T}[e(y)]$, the KL divergence of z-dimensional coded speech P and encoded text Q is as follows:

$$L_{dom} = 0.5[\log \frac{\det \Sigma_P}{\det \Sigma_Q} + tr\, \Sigma_Q^{-1} \Sigma_P + (\mu_P - \mu_Q)^T \Sigma_Q^{-1} (\mu_P - \mu_Q)] \tag{7}$$

In order to stabilize the training, we implement the determinant gradient (Gradient of determinant) in the logarithmic domain, so that the model can obtain better convergence.

Algorithm 1: Semi-supervised speech recognition model algorithm

```
The size of $D$ is T, which is divided into labeled datasets $D_{\sup}$ and unlabeled datasets $D_{un\sup}$ :
        $D_{\sup} = \{(X_n, Y_n) | n = 1,....., N\}, D_{un\sup} = \{(X_{N+1}, Y''_{N+1}) | n = N+1,....., T\}$
Where X stands for speech and Y for real label. $D_{un\sup}$ is made up of unpaired wav $S_i$ and text $T_i$.
$A$                                        ▷ encoder decoder ASR framework
$\alpha \in [0.5, 0.8], \beta \in [0.5, 0.8]$                ▷ hyperparameter
1: train a baseline model on A with labeled data sets.
2: initialize the super parameters of encoder and decoder:
3:for i in max( $len(D_{\sup}), len(D_{un\sup})$ ):
    Calculation Formula of Multi-objective function loss in semi-supervised speech recognition Model:
        $L = \alpha L_{\sup}(D_{\sup}) + (1-\alpha)\{\beta L_{text}(T_i) + (1-\beta) L_{dom}(S_i)\}$        (8)
    Update $\phi, \varphi$
end
```

4. Experimental Results and Analysis

4.1. Experimental Data and Platform

The open source Chinese dataset Aishell-1 [14] as the data set of this experimental model training, the sampling rate is 16khz, there are 400 speakers from China.

In the training of semi-supervised speech recognition model, the data set of the model is about 150 hours, the training data set of supervised speech recognition is about 30 hours, and the training data set of unsupervised speech recognition is about 120 hours. About 94156 unlabeled texts. The validation set is 18 hours and the test set is 10 hours.

The software environment used in this paper is a Pytorch deep learning environment built on 64-bit Ubuntu 20.04 operating system, and the verification of supervised and semi-supervised methods is carried out on the open source end-to-end voice toolkit wenet [15, 16].

4.2. Experimental Results

In order to verify the feasibility of semi-supervised speech recognition method and the influence of different decoding modes, this paper uses four decoding modes to carry out experimental analysis on semi-supervised speech recognition model, baseline model and the current best fully supervised model.

The four decoding modes are CTC greedy search [17], CTC prefix beam search [18], attention-based and rescoring [19]. The evaluation criterion of this paper is WER (word error rate), and all WER results are rounded to 2 decimal places. The recognition results of different decoding modes in Aishell-1 are shown in Table 1:

Table 1: Recognition results of different decoding modes in Aishell-1.

Mode model	Greedy search	Prefix beam search	attention	rescoring
state-of-the-art model	5.70	5.69	5.97	5.42
Baseline model	22.58	22.54	35.17	19.91
Semi supervised model	12.81	12.78	14.29	12.54

It can be seen from Table 1 that in the CTC greedy algorithm, CTC prefix tree search and attention decoding mode, the WER of the semi-supervised speech recognition model is about half of that of the baseline model, which clearly proves the feasibility of the semi-supervised speech recognition method. In other words, in the semi-supervised method, the conformer shared encoder and the calculation of the advanced feature loss after coding can improve the text-to-text mapping, which affects the speech-to-text mapping to some extent. However, there is still a certain gap between the WER of the semi-supervised model and the current fully supervised model.

As can be seen from Table 1, no matter which type of model it is, the recognition result of re-scoring is always the best. The better result of remarking recognition is determined by the idea of the algorithm, because the remarking decoding mode first uses the CTC prefix beam to search n best paths, which act as the attention of the text input and coding features of the transformer decoder and output the character sequence.

5. Conclusion

The end-to-end semi-supervised speech recognition model proposed in this paper uses a shared encoder to encode unlabeled speech and text at the same time. This semi-supervised learning method improves speech features by improving text feature extraction and then improves recognition accuracy. However, it is obvious that there is still a certain gap in recognition accuracy between this end-to-end semi-supervised speech recognition model and the fully supervised speech recognition model.

In the future research, in order to further improve the performance of the semi-supervised speech recognition model, we will explore in two aspects: 1) using bidirectional attention decoders (left-right decoder and right-to-left decoder) to solve the problem of reconstructed text 2) using contextual biasing [20] to improve the recognition accuracy of personalized words such as person name, street name, song name and so on.

Acknowledgments

This research was supported by Northwest Minzu University grant Research on multilingual intelligent processing (31920220010) of Basic research funds for central universities. I would like to thank my tutor for his guidance in my research direction and the laboratory of the Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province, Northwest Minzu University, Lanzhou, Gansu 730000, China.

References

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.

[2] Gulati A, Qin J, Chiu C C, et al. Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv: 2005. 08100, 2020.

[3] Graves A, Fern ández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks//Proceedings of the 23rd international conference on Machine learning. 2006: 369-376.

[4] Chorowski J, Bahdanau D, Cho K, et al. End-to-end continuous speech recognition using attention-based recurrent NN: First results. arXiv preprint arXiv:1412. 1602, 2014.

[5] Chorowski J, Bahdanau D, Serdyuk D, et al. Attention-Based Models for Speech Recognition. Computerence, 2015, 10(4):429-439

[6] Graves A. Sequence Transduction with Recurrent Neural Networks. Computer Science, 2012, 58(3):235-242.

[7] Chapelle O, Scholkopf B, Zien A. Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]. IEEE Transactions on Neural Networks, 2009, 20(3): 542-542.

[8] Ouali Y, Hudelot C, Tami M. An overview of deep semi-supervised learning. arXiv preprint arXiv:2006.05278, 2020.

[9] Higuchi Y, Moritz N, Roux J L, et al. Momentum Pseudo-Labeling for Semi-Supervised Speech Recognition. 2021.

[10] Zhang Y, Park D S, Han W, et al. BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Auto Lample G, Denoyer L, Ranzato M. Unsupervised Machine Translation Using Monolingual Corpora Only/ 2017.atic Speech Recognition. 2021.

[11] Karita S, Watanabe S, Iwata T, et al. Semi-Supervised End-to-End Speech Recognition//Interspeech. 2018: 2-6.

[12] Chan W, Jaitly N, Le Q, et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[13] Dai Z, Yang Z, Yang Y, et al. Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860, 2019.

[14] Bu H, Du J, Na X, et al. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline//2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA). IEEE, 2017: 1-5.

[15] Yao Z, Wu D, Wang X, et al. WeNet: Production oriented Streaming and Non-streaming End-to-End Speech Recognition Toolkit. 2021.

[16] Zhang B, Wu D, Peng Z, et al. WeNet 2.0: More Productive End-to-End Speech Recognition Toolkit. 2022.

[17] Chickering D M. Optimal structure identification with greedy search. Journal of machine learning research, 2002, 3(Nov): 507-554.

[18] Hannun A Y, Maas A L, Jurafsky D, et al. First-pass large vocabulary continuous speech recognition using bi-directional recurrent DNNs. arXiv preprint arXiv: 1408. 2873, 2014.

[19] Hori T, Hori C, Minami Y, et al. Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition. IEEE Transactions on audio, speech, and language processing, 2007, 15(4): 1352-1365.

[20] Aleksic P, Ghodsi M, Michaely A, et al. Bringing Contextual Information to Google Speech Recognition// International Conference on Concurrency Theory. Springer-Verlag, 2015.