VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

ASR - VLSP 2021: Automatic Speech Recognition with Blank Label Re-weighting

Dang Dinh Son, Le Dang Linh, Dang Xuan Vuong, Duong Quang Tien, Ta Bao Thang*

Viettel Cyberspace Center, Ton That Thuyet, Cau Giay, Hanoi, Vietnam

Received 14 May 2021
Revised 27 August 2021; Accepted 1 November 2021

Abstract: End-to-end models have significant potential in most languages and recently proved their robustness in ASR tasks. Many robust architectures are proposed, and among many techniques, Recurrent Neural Network – Transducer (RNN-T) shows remarkable success. However, with background noise or reverb in spontaneous speech, this architecture generally suffers from high deletion error problems. For this reason, we propose the blank label re-weighting technique to improve the state-of-the-art Conformer transducer model. Our proposed system adopts the Stochastic Weight Averaging approach, stabilizing the training process. Our work achieved the first rank with a 4.17% of syllable error rate in Task 2 of the VLSP 2021 Competition.

Keywords: End-to-End Automatic Speech Recognition, Blank Label Re-weighting.

1. Introduction

Automatic Speech Recognition (ASR) has represented the growth of the speech processing field. The trend of developing ASR models is spreading worldwide in terms of theoretical studying and actual products. Starting from [1], sequence-to-sequence models have outperformed the hybrid HMM/DNN ASR systems. From this, many robust End-to-End (E2E) ASR architectures have been proposed. The complexity of original hybrid systems has been reduced by using a single network to directly transform an input voice into an output sequence. Among many modern deep-learning architectures, the study [2] shows that combining convolution and self-attention is better than using these methods individually. To our best knowledge, although this approach has proven its strength in many languages, there have not been many proposals or studies for Vietnamese speech. Therefore, we implemented the RNN-T with Conformer encoder architecture [3] for Vietnamese speech. While a backlog problem remains that E2E models generally suffer from a considerable deletion error rate, blank label

* Corresponding author.
E-mail address: tabaothang97@gmail.com

https://doi.org/10.25073/2588-1086/vnucsce.321

re-weighting is proposed. Our approach aims not only the blank-label suppression but also to keep the balance between blank and non-blank tokens. This technique directly improves our result, which reached 4.17% SyER. The result ranked first in the 8th International Workshop on Vietnamese Language and Speech Processing (VLSP 2021)[4].

The pipeline of this paper starts with corpus detail in section 2. Then in section 3, we present our methods, specifically with the Conformer transducer and blank label re-weighting. We finally show our experiments and evaluation in section 4, and conclude our work in section 5.

2. Corpus description

We collected and labeled about 2000 hours of audio from Youtube videos. Because our target is spontaneous speech in different real scenarios, we prioritized selecting conversational videos. All game-show or advertisement speech that does not contain diverse speakers, flexible context, or expressions is significantly removed. We use the WebRTC [5] as the tool for preprocessing stage.

We aim to ensure that the model learns the most accurate knowledge of dialogues' acoustic and language representation. Therefore, we clean the data by removing noisy audio, unclear speech, and samples with wrong transcripts. This preparation forces the model to be familiar with natural dialogue speech rather than pre-scripted ones. The duration distribution of the dataset is shown in Figure 1. The most frequent duration is centralized at the interval of 4 seconds. The most prolonged duration is about 15 seconds, and it only accounts for a small number of utterances

3. Method

In this section, we present the transducer model reasoning and implementation in Section 3.1. Then most importantly, we propose a blank label re-weighting algorithm in Section 3.2 to solve the deletion error problem.

3.1. Transducer model

Sequence to sequence learning with E2E approaches is divided into some main types such as Connectionist Temporal Classification (CTC) [6, 7], Recurrent Neural Network Transducer (RNN-T) [8], and Attention-based Encoder-Decoder (AED) [1, 9]. Comparisons have been raised and point out that the RNN-T architecture extends the superior advantages among these approaches [10]. Contrary to the old RNN-T that used LSTM-RNN for the encoder, the futuristic encoder is equipped with Conformer [3] architecture. This inherited and improved the strength of transformer [11] with the combination of convolution and self-attention.

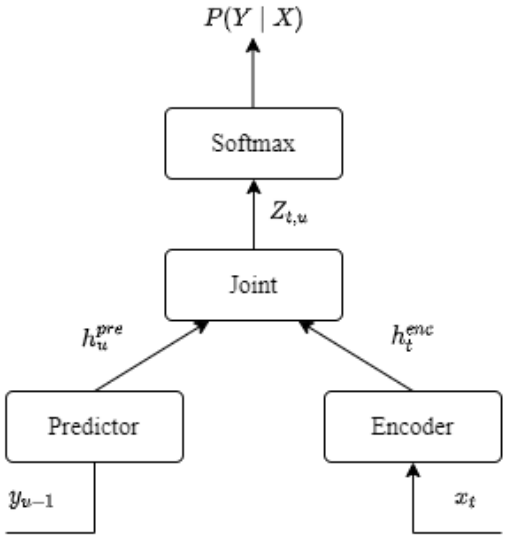

Figure 2. Transducer network [12].

With this tendency, we build a Conformer-Transducer E2E model. Particularly, the aim of the model is to predict the P(YjX) distribution of all output token sequences Y = [y_1, y_2,…., y_U] on condition that a sequence of acoustic feature
X = [x_1, x_2,….,; x_T] is taken as input. Our aim is getting a high-level representation from the encoded inputs. This transducer model in

Figure 2 contains a combination of a LSTM-RNN prediction network, a Conformer encoder and a joint network. The high-level representation from the encoder f^{enc} is setup as $h^{enc} = [h_1^{enc}, h_2^{enc}, \ldots, h_T^{enc}]$, with the output from the embedding vectors of labels $[y_1, y_2, \ldots, y_{u-1}]$, at step u is hupred and frame t-th, the joint network outputs the logits:

$$Z_{t,u}^{joint} = f^{joint}(h_t^{enc}, h_u^{pred}) \quad (1)$$

The distribution of probability over vocabulary at frame t-th and step u-th is calculated using the softmax layer.

3.2. Blank Label Re-weighting

With each frame fed into the network in transducer architecture, the model can generate an arbitrary number of tokens. This architecture uses an extra token called a blank token (). During model testing, the probability of a blank token is usually quite large [13]. This led to an increase in deletion error where output missed many tokens.

Algorithm 1 Blank Label Re-weighting (BLR)

Non-blank probabilities before BLR
$P_{t,u}(blank) \leftarrow 1 - P_{t,u}(non - blank)$

Scale factor calculation
$\gamma \leftarrow 1 + \frac{\beta \cdot P_{t,u}(blank)}{P_{t,u}(non-blank)}$

blank token probabilities after BLR
$P_{t,u}(blank) \leftarrow (1 - \beta) \cdot P_{t,u}(blank)$

Non-blank token probabilities after BLR
$P_{t,u}(non - blank) \leftarrow \gamma \cdot P_{t,u}(non - blank)$

To deal with this problem, we refer to the blank label re-weighting approach [13]. This original article proposed reducing the probability of a blank token. Generally, the model has the bias of generating blank tokens due to its huge likelihood. In this work, our method also suppresses the blank token probability. However, there has to be a balance between the blank and non-blank token scores. If the suppression factor is too high, deletion errors would be replaced by insertion errors, which is not what we expected.

Our improvement is to normalize the scores for non-blank symbols to ensure that the total sum of the probabilities over all RNN-T outputs (including both blank and non-blank) equals one. This helps to keep both the blank and non-blank token values balance. Blank label re-weighting process is described in the Algorithm 1. The β hyper-parameter is used as a scale factor of blank probability. This directly solves the problem of overvalued blank scores.

4. Experiments

4.1. Environmental Setup

To show improvement and evaluate the system performance on the datasets, we computed 80-channels of Mel-filterbank, set up the window width to 25ms and the stride to 10ms. We also apply speed perturbation [14, 15] and SpecAugment [16] with mask parameter F = 27, and ten-time masks with maximum time-mask ratio pS = 0:05, where the maximum size of the time mask is set to pS times the length of the utterance.

The filterbank features are first passed into two blocks of 2d-Conv layers, and then time reduction layers are added after each block to down-sample the frame rate to 4 before passing into the encoder. The encoder model consists of 16 layers of conformer block, where we set the model dimension to 640, with eight attention heads, 31 kernel size in convolution block, with the same setting as Conformer-L [3]. We use LSTM as our predictor, and the LSTM predictor contains one layer with 640 units and a projection layer with 640 units. The transducer's joint network is designed as a simple feed-forward layer.

The total number of parameters is about 166M. Our model is implemented in Pytorch optimized with Adam. We use the learning rate warmup for the first 10k updates up to 1e-4 peak for both transducer model training. The model is trained on 8 Nvidia A100 GPUs with a batch size of 128, for 30 hours of training

time . We observe that the model convergence can be improved by applying weight averaging and learning rate cycles in the training stage. Therefore, while the model is trained with 30 epochs, the last six epochs are used with stochastic weight averaging [17].

4.1. Evaluation

As shown in Table 1, we find out the best β coefficient in blank label re-weighting, with the value of 0.5. After employing testing experiments, we observed the relationship between the value of the coefficient and the error rate. The reason is that if the value of β is too low, deletion errors will happen more frequently. On the contrary, our model would suffer an insertion error problem when the coefficient value is too high.

Finally, we achieved 4.17% SyER on the VLSP 2021 test set. This result got first place in the VLSP 2021 evaluation campaign, surpassed all competitors, and proved the robustness of the proposed method.

Table 1. Influence of β on Syllable error rate (% SyER) over VLSP test datasets:

β	VLSP2019	VLSP2020	VLSP2021
0	13.109	5.393	4.463
0.1	12.251	5.358	4.389
0.2	12.231	5.319	4.277
0.3	12.192	5.256	4.191
0.4	12.18	5.210	4.175
0.5	**12.174**	**5.162**	**4.165**
0.6	12.543	5.35	4.165

5. Conclusion

This paper proposed blank label re-weighting on a transducer model to point out the best-chosen hyper-parameter. SWA was also implemented to improve the accuracy over the training stage. Concerning minimizing the syllable error rate, we evaluated the efficiency of our model on different β coefficient values while re-weighting the blank token. After many experiments, our best model achieved the result of 4.17% SyER. In future works, we will implement this method of blank token re-weighting in the training process. Instead of a fixed value of scale factor that might not be the best optimized, jointly training will be used to get the best re-weighting factor value.

References

[1] W. Chan, N. Jaitly, Q. V. Le, O. Vinyals, Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition, in: ICASSP, 2016.

[2] I. Bello, B. Zoph, A. Vaswani, J. Shlens, Q. V. Le, Attention Augmented Convolutional networks, 2020, arXiv:1904.09925.

[3] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Zhang, J. Yu, W. Han, S. Wang, Zhang, Y. Wu, R. Pang, Conformer: Convolution-Augmented Transformer for Speech Recognition, in: Interspeech 2020, ISCA, 2020.

[4] D. V. Hai, ASR Challenge: Vietnamese Automatic Speech Recognition, in : VLSP 2021.

[5] Google, http://web.archive.org/web/20080207010024/; http://www. 808multimedia.com /winnt/kernel.htm (2011).

[6] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural ’Networks, Vol. 26, 2006, pp. 369–376.

[7] A. Graves, N. Jaitly, Towards End-To-End Speech Recognition with Recurrent Neural Networks, in: E. P. Xing, T. Jebara (Eds.), Proceedings of the 31st International Conference on Machine Learning, Vol. 32 of Proceedings of Machine Learning Research, PMLR, Bejing, China, 2014, pp. 1764–1772.

[8] Graves, Sequence transduction with recurrent neural networks, 2012. arXiv:arXiv:1211.3711.

[9] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio, Attention-Based Models for Speech Recognition, arXiv preprint arXiv:1506.07503.

[10] J. Li, Y. Wu, Y. Gaur, C. Wang, R. Zhao, S. Liu, On The Comparison of Popular End-To-End Models for Large Scale Speech Recognition, 2020, arXiv: 2005.14327.

[11] Q. Zhang, H. Lu, H. Sak, A. Tripathi,

E. McDermott, S. Koo, S. Kumar, Transformer Transducer: A Streamable Speech Recognition Model With Transformer Encoders And Rnn-T Loss, 2020, arXiv:2002.02562.

[12] H. Shrivastava, A. Garg, Y. Cao, Y. Zhang, T. Sainath, Echo State Speech Recognition, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 5669–5673.

[13] Y. Zhang, S. Sun, L. Ma, Tiny transducer: A Highly-Efficient Speech Recognition Model On Edge Devices, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, https://doi.org/10.1109/icassp39728.2021.9413854

[14] H. Song, M. Kim, D. Park, Y. Shin, J.-G. Lee, Learning From Noisy Labels With Deep Neural Networks: A survey (2020). arXiv:arXiv: 2007.08199.

[15] M. A. Hasegawa-Johnson, P. Jyothi, D. McCloy, M. Mirbagheri, G. M. di Liberto, A. Das, B. Ekin, C. Liu, V. Manohar, H. Tang, E. C. Lalor, N. F. Chen, P. Hager, T. Kekona, R. Sloan, A. K. C. Lee, ASR For Under-Resourced Languages From Probabilistic Transcription, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 25, No. 1, 2017, pp. 50–63. https://doi.org/10.1109/taslp. 2016.2621659

[16] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, Specaugment: A Simple Data Augmentation Method for Automatic Speech Recognition, Interspeech 2019, http://dx.doi.org/10.21437/Interspeech.2019-2680

[17] S. Ling, C. Shen, M. Cai, Z. Ma, Improving Pseudo-Label Training For End-To-End Speech Recognition Using Gradient Mask, 2021, arXiv: arXiv:2110.04056.