Pre-training Transformer decoder for end-to-end ASR model with unpaired text data

Changfeng Gao, Gaofeng Cheng, Runyan Yang, Han Zhu, Pengyuan Zhang, Yonghong Yan

Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, China

University of Chinese Academy of Sciences, China

Abstract

- Pre-training method for encoder-decoder ASR models using text data only.
- Use the empty, worst or ideal artificial states to replace the real encoder states during pre-training
- Remain the network architecture unchanged and do not introduce extra component.

Background

Usage method for the text-only data

- Extra language model with shallow fusion. [Kar+19]
- Back-translation style methods with text-to-speech [Bas+19] or text-to-encoder [Hay+18] system.
- Pre-training methods like BERT[Dev+18] cannot be applied on the E2E ASR system.

Pre-training the decoder in E2E system using text-only data

Difficulty

- For a transformer decoder block, there are two multi-head attention machine:

$$\text{selfMHA}(X) = \text{MHA}(X, X, X) \quad (1)$$

$$\text{srcMHA}(X, Y) = \text{MHA}(X, Y, Y) \quad (2)$$

- .. the srcMHA needs the encoder states as input, which are unavailable during pre-training.

Solution

- LM pre-training: ignoring the srcMHA, pre-train a transformer LM and then initialize the parameters in the transformer decoder.
- AC pre-training: design an artificial condition (AC) states for the decoder during pre-training.

Language model pre-training

Method

- During pre-training, we feed the srcMHA with empty state and provide that when the state is empty, the srcMHA degenerates into an identity transformation

$$\text{srcMHA}(X, Y) = \begin{cases} \text{MHA}(X, Y, Y) & Y \neq empty \\ X & Y = empty \end{cases} \quad (3)$$

Problem

- The parameters of the srcMHA still remain randomly initialized.
- There is no rational explanation for the degeneration of the srcMHA

Artificial condition pre-training

Method

- We construct the artificial states as the input of the srcMHA to replace the encoder hidden states.
- The length of the artificial states is calculated according to the pronunciation duration and the value of the artificial states is designed by two assumptions.

Ideal condition pre-training

- The artificial states are generated by an ideal encoder
- The ideal encoder can transfer speech features into word vectors directly.
- The word vectors can be obtained by the LM pre-training.

Worst condition pre-training

- The artificial states are generated by a failed encoder
- The failed encoder loses all information in the speech features and converts them into random noise.

Results

Effectiveness of the pre-training

- Ac pre-training is better than the LM pre-training, WC pre-training is better than the IC pre-training.

Table 1:WER[%] for different models. Pre-trained with 960 hours and fine-tuned with 100 hours.

Pre-training method	Test clean	Test other	Dev clean	Dev other
Baseline	12.6	31.5	12.0	31.0
LM pre	11.6	30.5	11.2	30.4
IC pre	11.4	31.0	**10.8**	30.7
WC pre	**11.2**	**30.5**	10.9	**30.0**

Impact of the unpaired data

- Extra text can lead to better performance.
- The pre-train method is still useful when only use the transcription of the paired data.

Table 2:WER[%] and WERR[%] for the models pre-trained with different text.

Fine-tune data	Pre-train text	Test WER clean	Test WER other	Test WERR clean	Test WERR other
100hr	–	12.6	31.5	–	–
460hr	–	7.9	20.8	–	–
960hr	–	6.0	13.0	–	–
100hr	100hr	11.9	31.6	5.6	-0.3
100hr	960hr	11.2	30.5	11.1	3.1
460hr	960hr	6.7	20.0	16.4	3.8
960hr	960hr	4.9	12.2	18.3	7.7

What does the decoder learn ?

- pre-trained deocder can detect the correct pronunciation positions in the speech during the AC pre-training.
- We visualize the attention weights for the srcMHA to show what the decoder learn. See paper for details.

Conclusions

- We design a novel pre-training strategy for the decoder of the transformer-based E2E ASR model by using unpaired text data.
- Our pre-training method does not need extra component or change the neural network structure.
- Experiments on Librispeech corpus prove the effectiveness of our method and explain what the decoder can learn during pre-training.

More recent work

- We combined this proposed decoder pre-training method with some encoder pre-training methods, and further improved the E2E ASR performance.
- We proved this proposed method can be applied on other language like Chinese.
- We also further simplified the pre-training pipline for the WC pre-training.
- We are evaluating the proposed method on the RNN-based decoder.

References

[Bas+19] Murali Karthick Baskar et al. “Semi-supervised Sequence-to-sequence ASR using Unpaired Speech and Text”. 2019.

[Dev+18] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. 2018.

[Hay+18] T. Hayashi et al. “Back-Translation-Style Data Augmentation for end-to-end ASR”. 2018.

[Kar+19] S. Karita et al. “Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration”. 2019.