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

Language model pre-training

Method

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

Problem

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.

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.

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

Difficulty

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

\[ \text{selfMHA}(X) = \text{MHA}(X, X, X) \]  
\[ \text{srcMHA}(X, Y) = \text{MHA}(X, Y, Y) \]  

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.

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.

Background

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

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.

\[ \text{srcMHA}(X, Y) = \begin{cases} \text{MHA}(X, Y, Y) & \text{if } Y \neq \text{empty} \\ X & \text{if } Y = \text{empty} \end{cases} \]  

Problem

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

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.

Impact of the unpaired data

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

What does the decoder learn?

The pre-trained decoder 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.

Results

Table 1: WER[\%] for different pre-trained method.

<table>
<thead>
<tr>
<th>Pre-training Method</th>
<th>Test WER</th>
<th>Dev WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>12.6</td>
<td>31.5</td>
</tr>
<tr>
<td>LM pre</td>
<td>11.6</td>
<td>30.5</td>
</tr>
<tr>
<td>IC pre</td>
<td>11.4</td>
<td>31.0</td>
</tr>
<tr>
<td>WC pre</td>
<td>11.2</td>
<td>30.5</td>
</tr>
</tbody>
</table>

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 pipeline for the WC pre-training.
We are evaluating the proposed method on the RNN-based decoder.

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.

References