Improving Transformer-based Speech Recognition with Unsupervised Pre-training and Multi-task Semantic Knowledge Learning

Song Li\textsuperscript{1}, Lin Li\textsuperscript{1}, Qingyang Hong\textsuperscript{2}, Lingling Liu\textsuperscript{1}

\textsuperscript{1} School of Electronic Science and Engineering, Xiamen University, China
\textsuperscript{2} School of Informatics, Xiamen University, China

lilin@xmu.edu.cn, qyhong@xmu.edu.cn

Abstract

Recently, the Transformer-based end-to-end speech recognition system has become a state-of-the-art technology. However, one prominent problem with current end-to-end speech recognition systems is that an extensive amount of paired data are required to achieve better recognition performance. In order to grapple with such an issue, we propose two unsupervised pre-training strategies for the encoder and the decoder of Transformer respectively, which make full use of unpaired data for training. In addition, we propose a new semi-supervised fine-tuning method named multi-task semantic knowledge learning to strengthen the Transformer’s ability to learn about semantic knowledge, thereby improving the system performance. We achieve the best CER with our proposed methods on AISHELL-1 test set: 5.9\%, which exceeds the best end-to-end model by 10.6\% relative CER. Moreover, relative CER reduction of 20.3\% and 17.8\% are obtained for low-resource Mandarin and English data sets, respectively.

Index Terms: unsupervised pre-training, speech recognition, Transformer, multi-task learning, semi-supervised learning

1. Introduction

Sequence-to-sequence attention-based models have recently shown very promising results for automatic speech recognition (ASR) tasks \cite{1–5}. Compared with the traditional speech recognition systems based on hidden Markov model (HMM), they directly learn speech-to-text mapping with the pure neural networks, without a special phoneme dictionary to convert words into phonemes. Transformer \cite{6} is one of the state-of-the-art sequence-to-sequence architectures, which has performed promisingly in building end-to-end speech recognition systems \cite{7–10}. Compared with RNN, Transformer introduces multi-head attention to study features from multiple subspaces and directions, so that the model can extract more representative features. In addition, it calculates in parallel, which is faster than RNN. However, end-to-end speech recognition usually requires a great deal of paired data to train, in order to achieve better recognition results. This is unfriendly to some low-resource applications. Relative to supervised data (including both speech data and the corresponding text data), unpaired data are much easier to collect. Therefore, how to use a large amount of unpaired speech and text data from real life scenario to strengthen the training performance of speech recognition systems has become one of the hot topics for researchers.

In order to make full use of unpaired data, researchers had proposed two main strategies: unsupervised pre-training and semi-supervised learning. Unsupervised pre-training strategies \cite{11–13}, like bidirectional encoder representations from Transformers (BERT) \cite{11} and generative pre-training (GPT) \cite{13} in the natural language processing field, aim to learn a gener-
are described. In Section 3, we introduce specific experimental details. Experimental results are presented in Section 4. Finally, the paper is concluded in Section 5.

2. Our proposed methods

2.1. System overview

In this paper, we investigate our proposed unsupervised pre-training strategies and fine-tuning method on the Transformer architecture, which includes three significant components: the encoder, attention, and the decoder. We replace the position-wise fully-connected feed-forward network layers in the standard Transformer architecture with one-dimensional convolution (Conv1D) layers, which introduce more non-linear characteristics to speech recognition systems. To explore the efficient contributions of sufficient amounts of unpaired data, we pretrain the encoder and the decoder models with our proposed unsupervised pre-training strategies, respectively. The SPC pre-training aims to integrate useful acoustic semantic information contained in speech into Transformer’s encoder by predicting some masked features in the speech feature sequence. The TPC pre-training provides Transformer’s decoder with rich linguistic semantic information by an autoregressive language model objective. Fig.1 illustrates the details of our Transformer block and unsupervised pre-training strategies.

![Figure 1: A schematic representation of our unsupervised pre-training strategies and Transformer block.](image)

2.2. SPC model

In order to obtain acoustic semantic knowledge, SPC takes an MLM-like objective to obtain general feature representations of speech. To pre-train the SPC model, time masking is first applied to the acoustic features, so that a series of consecutive time steps \{t_0, t_0+t\} are masked. The parameter \(t\) is chosen randomly from a uniform distribution from 0 to the time mask parameter \(W\), and \(t_0\) is chosen randomly from \([0, T - t]\), where \(T\) is the time step of acoustic features and \(W\) is set to 30. Similar to time mask, frequency mask is used so that the frequency range of \([h_0, h_0+h]\) are masked. In addition, in order to make SPC model training easier, we subsample and normalize the FBank coefficients with the normalization layers and convolution layers. The down-sampled FBank coefficients are further extracted by an encoder, which is composed of our Transformer block, to obtain high-level acoustic feature representations. Finally, the clean FBank coefficients are reconstructed through a linear layer and transposed convolution layers. The detailed structure for the SPC is shown in Fig.1b. Instead of calculating the loss of all feature frames, we only calculate the loss of the corresponding masked FBank coefficients. The loss function of SPC model is defined as follows:

\[
L_{\text{SPC}} = \frac{1}{B \cdot T} \sum_{b=1}^{B} \sum_{i=1}^{T} L_{\text{Huber}}(\{x'_{b}, x_{b+i}\})
\]

2.3. TPC model

The linguistic semantic information contained in the text is very rich. In order to obtain this information, the TPC model uses an autoregression language model objective to obtain general feature representations of the text. As shown in Fig.1c, TPC consists of a word embedding layer, a linear layer, and a decoder. The decoder is composed of our Transformer block with a masked multi-head self-attention. We train a TPC model using a large amount of unpaired text. The loss function of the TPC model is defined as follows:

\[
L_{\text{TPC}} = \frac{1}{B \cdot N} \sum_{b=1}^{B} \sum_{i=1}^{N} y_{bi} \log p(y_{bi} | x_{b1}, x_{b2}, ..., x_{b(i-1)})
\]

where \(B\) is the training batch size, and \(N\) is the index for text tokens. \(\{x_{bi}, i = 1, 2, ..., n\}\) are the input text sequences for TPC, which consist of a sentence with a starting symbol \(<\text{sos}>\). \(\{y_{bi}, i = 1, 2, ..., n\}\) are the autoregressive output sequences of TPC. \(\{y_{bi}, i = 1, 2, ..., n\}\) are the target sequences of TPC, which consist of a sentence with a terminator symbol \(<\text{eos}>\).

TPC model calculates the conditional probability of the next word based on previous words in an autoregressive manner, which enables the TPC model to learn about the relationships between the former and the latter words in a sentence. These relationships usually represent rich linguistic semantic knowledge from the text. The Transformer-based speech recognition system is also an autoregressive model. If we incorporate linguistic information in each decoding step, the performance of speech recognition systems improves, which is exactly the benefit of the TPC model. After pre-training, we use the TPC model’s weights to initialize Transformer’s decoder, except for the encoder-decoder attention part.

2.4. Multi-task semantic knowledge learning

In order to prevent Transformer from forgetting acoustic and linguistic semantic knowledge during the fine-tuning process, we propose a multi-task semantic knowledge learning (MTSL) method, which is shown in Fig.2. Specifically, an auxiliary task is introduced at the encoder output of Transformer, which reconstructs the clean acoustic features with the same loss func-

\[
L_{\text{Huber}}(y - y') = \begin{cases} \frac{1}{2} (y - y')^2, & |y - y'| \leq \delta \\ \delta |y - y'| - \frac{1}{2} \delta^2, & \text{otherwise} \end{cases}
\]

where \(\{x_{bi}, i = 1, 2, ..., n\}\) are the output of SPC, \(\{x_{bi}, i = 1, 2, ..., n\}\) are the original unmasked FBank coefficients, \(B\) is the batch size, and \(T\) is the time step of the FBank coefficients. \(L_{\text{Huber}}\) is the Huber loss function with \(\delta = 0.5\).
tion of the SPC model. Note that reconstruction loss is back-propagated only when acoustic features are masked. In this case, the masked acoustic features are seen as one kind of data augmentations of speech. We add another auxiliary task after the self-attention of the decoder, which uses the same loss function as the TPC model for language modeling. Through multi-task semantic knowledge learning, Transformer continues to learn about acoustic and linguistic semantic knowledge during fine-tuning. The loss function of the fine-tuning process with multi-task semantic knowledge learning is defined as follows:

\[ L_{MTSL} = \alpha L_{CTC} + (1 - \alpha) L_{TRA} + \lambda_1 L_{SPC} + \lambda_2 L_{TPC} \] (4)

where \( L_{CTC} \) is the CTC loss, \( L_{TRA} \) is the cross entropy loss of Transformer, and \( L_{SPC} \) and \( L_{TPC} \) are the SPC loss and the TPC loss respectively. \( \alpha \) and \( \lambda_1, \lambda_2 \) are hyper parameters used to balance every losses. We set \( \alpha = 0.3, \lambda_1 = 0.2, \) and \( \lambda_2 = 0.1 \) in this paper.

3. Experimental setup

3.1. Data sets

In our experiments, for the sake of universality, both Mandarin and English applications are considered. Mandarin datasets are used including AISHELL-1 [24], AISHELL-2 [25], THCHS30 [26] and the Datatang dataset: Chinese500. LibriSpeech [27] and Tedlium2 [28] are used as the English datasets. For the AISHELL-2 dataset, only 681 hours of transcription are used. AISHELL-1 training set includes about 151 hours of speech data, TPC is more useful than SPC. We further introduce two representative benchmarks: the first one is a TDNN-HMM model optimized with the lattice free maximum mutual information (LF-MMI) objective, the second one is a LSTM-attention model with CTC and CE objectives. We also conducted two unsupervised pre-training methods named 1PC [20] and APC [19] and fine-tuned on the same fine-tuning datasets.

We first conducted experiments with SPC and TPC using only AISHELL-1 training set as pre-training data. The results listed in Table 2 show that compared to the baseline, 0.8% relative CER reduction is obtained with only SPC pre-training or both SPC and TPC pre-training, 0.5% with TPC pre-training. This indicates that the our proposed pre-training strategies are useful even without any additional data.

To further verify the effects of various amounts of pre-training data size on the fine-tuning results, we first merged THCHS30 and AISHELL-1 training set as a new pre-training dataset. The results listed in Table 2 show that using SPC alone achieves 6.51% CER, while using TPC alone further reduces CER to 6.41%. So, for a small amount of additional pre-training data, TPC is more useful than SPC. We further increased the pre-training data by combining AISHELL-1 training set, THCHS30, and Chinese500 to create a new pre-training dataset called Combine. The results show that CER decreased over the baseline relatively by 5.9% using only TPC, 7.0% using only SPC, and 7.3% using both SPC and TPC. We find that when there is more pre-training data, SPC plays a more important role.

### 4. Results

#### 4.1. Unsupervised pre-training

Our baseline system is the Transformer-CTC hybrid structure used in [7], except that the position-wise fully-connected feed-forward network layers are replaced with Conv1D layers, which has been proven to achieve state-of-the-art performance in all current end-to-end speech recognition systems. In addition, we conducted two representative benchmarks: the first one is a TDNN-HMM model optimized with the lattice free maximum mutual information (LF-MMI) objective, the second one is a LSTM-attention model with CTC and CE objectives. We also conducted two unsupervised pre-training methods named M-PC [20] and APC [19] and fine-tuned on the same fine-tuning datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Speech (hours)</th>
<th>Text transcription (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AISHELL-1</td>
<td>151</td>
<td>151</td>
</tr>
<tr>
<td>THCHS30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Chinese500</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>AISHELL-2</td>
<td>681</td>
<td>681</td>
</tr>
<tr>
<td>LibriSpeech</td>
<td>960</td>
<td>960</td>
</tr>
</tbody>
</table>

Table 1: Detailed Unsupervised Pre-training Dataset Information

Footnote: 1A Chinese data provider (https://www.datatang.com/)

used to replace the position-wise fully-connected feed-forward network. In addition, a convolutional front-end was used to sub-sample the acoustic features by a factor of 4.
4.2. Effects of multi-task semantic knowledge learning

The results listed in Table 2 indicate that the proposed multi-task semantic knowledge learning is beneficial for CER reduction even without unsupervised pre-training. We obtained the best CER for AISHELL-1 test set: 5.9% when we used our proposed unsupervised pre-trained strategies with the Combine dataset as the pre-training data and with the integration of MTSL, which achieved 10.6% reduction for CER. We conducted APC and MPC with the same unsupervised pre-training data set, and the CER of AISHELL-1 only reduced by 5.3% and 4.9% respectively, which indicates that our methods are better than MPC and APC using the same unsupervised pre-training data size. We owe the effects of multi-task semantic knowledge learning to the fact that time and frequency masking has a similar effect to dropout [33], which prevents overfitting in the training set. In addition, the mask strategies of SPC for input features is similar to SpecAugment [34], which can achieve the effect of data augmentation and improve the generalization of speech recognition systems. Finally, since the two mask strategies are equivalent to introduce some noise to input features, the reconstruction loss strengthens the anti-noise performance of the Transformer’s encoder.

4.3. Randomness of unpaired data

In order to prove the effectiveness of our strategies are not due to some inherent connections in the pre-training data (originally paired), we used the speech data of Combine dataset to pretrain SPC and used the same-scale text of AISHELL-2, which is not paired with the Combine dataset, to pre-train TPC. The results show that we can still achieve the similar results as our best results, which indicates that our strategies are useful for arbitrary unpaired data.

4.4. Low-resource case

To prove that our proposed unsupervised pre-training strategies and fine-tuning method are still effective in low-resource case, we used Combine and LibriSpeech datasets to perform unsupervised pre-training using TPC and SPC strategies, and then fine-tuned on the 50-hour AISHELL-1 dataset and Tedlium2 dataset with MTSL, respectively. As shown in Table 3, the CERs of AISHELL-1 and Tedlium2 have been reduced relatively by 20.3% and 17.8%, which indicates that our proposed methods are able to learn useful semantic information from a large amount of unlabeled speech and text data so as to improve the performance of low-resource speech recognition systems.

Table 2: The test set CER(%) of AISHELL-1 with different pre-training strategies and pre-training data size

<table>
<thead>
<tr>
<th>Model Strategy</th>
<th>Unsupervised Pre-training Data Size</th>
<th>CER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>151 (AISHELL-1)</td>
<td>16.5</td>
</tr>
<tr>
<td>Transformer-Conv1d-MTSL SPC+TPC Combine(681 hours)+AISHELL-2(681 hour text)</td>
<td>9.0</td>
<td></td>
</tr>
<tr>
<td>Transformer-Conv1d-MTSL SPC+TPC Combine(681 hours)</td>
<td>9.5</td>
<td></td>
</tr>
<tr>
<td>Transformer-Conv1d-MTSL SPC+TPC AISHELL-1+THCHS30(181 hours)</td>
<td>10.6</td>
<td></td>
</tr>
<tr>
<td>Transformer-Conv1d-MTSL SPC+TPC AISHELL-1(151 hours)</td>
<td>11.1</td>
<td></td>
</tr>
<tr>
<td>Transformer-Conv1d-MTSL SPC+TPC AISHELL-1(50 hour)</td>
<td>11.5</td>
<td></td>
</tr>
<tr>
<td>Transformer-Conv1d-MTSL SPC+TPC AISHELL-1 LibriSpeech(960 hours)</td>
<td>11.6</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: The test set CER(%) of AISHELL-1 and Tedlium2 Dataset in low-resource case

<table>
<thead>
<tr>
<th>Fine-tuning Dataset</th>
<th>Pre-training Data Size (hours)</th>
<th>CER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AISHELL-1(50 hours)</td>
<td>Combine(681 hours)</td>
<td>15.4</td>
</tr>
<tr>
<td>AISHELL-1(50 hours)</td>
<td>LibriSpeech(960 hours)</td>
<td>12.2</td>
</tr>
<tr>
<td>AISHELL-1(50 hours)</td>
<td>Tedlium2(50 hours)</td>
<td>20.8</td>
</tr>
<tr>
<td>Tedlium2(50 hours)</td>
<td>LibriSpeech(960 hours)</td>
<td>17.1</td>
</tr>
</tbody>
</table>

4.5. Cross-lingual case

For some low-resource languages, we may not have enough data for pre-training. Inspired by [35], we assume that SPC pre-training in other languages is also helpful for fine-tuning on Mandarin dataset, so we used LibriSpeech dataset to pretrain the SPC model and then fine-tuned on the low resource AISHELL-1 set. As we can see from Table 3 that CER decreased relatively by 4.6% over the baseline, which indicates that there are some commonalities for speech in different languages. We attribute this improvement to some phonemic features shared between different languages at a certain level.

4.6. Loss and ACC

We observed the effectiveness of the proposed pre-training strategies for model convergence and accuracy improvement. The loss and the accuracy rate (ACC) curves are shown in Fig.3. The curve ending with _std represents the baseline without pre-training, while the curve ending with _Com represents the case where the pre-training strategies are used with the Combine data set. We can find that the pre-training strategies proposed in this paper provide a better initial position for model training, so that the model converges faster, and the accuracy increases faster. It’s obviously that rich acoustic and linguistic knowledge obtained from pre-trained SPC and TPC models benefits downstream automatic speech recognition (ASR) tasks.

Figure 3: Loss and ACC curves with and without our proposed methods.

5. Conclusions

In this paper, we propose two unsupervised pre-training strategies, named speech predictive coding (SPC) and text predictive coding (TPC). These strategies use a large amount of unlabeled speech and text data for pre-training, and provide rich acoustic and linguistic semantic information for downstream tasks. We also propose a new semi-supervised fine-tuning method, named multi-task semantic knowledge learning, which helps Transformer to strengthen the learning capability of semantic knowledge during the fine-tuning process. Through our unsupervised pre-training strategies and the fine-tuning method, the performance of Transformer-based speech recognition system is improved, which is suitable for the low-resource and cross-lingual speech recognition applications.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No.61876160).
7. References


