Interspeech 2018
End-to-End based ASR
Part 1-2

Lu Huang
2018-09-19
Papers for Mandarin Chinese ASR

• Alibaba: Acoustic Modeling with DFSMN-CTC and Joint CTC-CE Learning

• Bo Xu: Extending Recurrent Neural Aligner for Streaming End-to-End Speech Recognition in Mandarin

• Bo Xu: Syllable-Based Sequence-to-Sequence Speech Recognition with the Transformer in Mandarin Chinese
Acoustic Modeling with DFSMN-CTC and Joint CTC-CE Learning

Shiliang Zhang, Ming Lei

Machine Intelligence Technology, Alibaba Group
{sly.zsl, lm86501}@alibaba-inc.com
Motivations

• CTC based ASR system’s latency
  • CTC: an output target is detected can be arbitrarily delayed
  • (B)LSTM: a huge amount of memory when the sequence is very long; BLSTM’s latency

• How to:
  • Using DFSMN to replace (B)LSTM
  • Joint CTC-CE training to improve stability
CTC

\[
\begin{align*}
F(a, -, b, c, -, -) \\
F(\ldots, a, -, b, c) \\
F(a, b, b, b, c, c) \\
F(a, -, b, -, c, c)
\end{align*}
\] => \((a, b, c)\)

\[
P(z|x) = \sum_{\pi \in \Phi(z)} P(\pi|x)
\]

\[
\mathcal{L}_{ctc}(x) = -\log P(z|x)
\]
CTC training

• CTC training is not stable

• How to:
  • by using **two output layers** with CTC and the conventional CE loss during the training
  • **initializing from a CE** loss pre-trained model.

• It is found that even with CE pre-trained networks as initialization, CTC training can sometime still fail to converge.

• CTC training with CI-Phones is more stable than CD-Phones.
  • The searching space of CD-Phones alignments is more huge than that of CI-Phones.
Joint CTC-CE Learning

• Difference between CTC CE:
  • loss function
  • additional CTC blank

• Joint CTC-CE
  • a single softmax output layer
    \[ \mathcal{L}_{ctcce}(x) = \mathcal{L}_{ctc}(x) + \alpha \cdot \mathcal{L}_{ce}(x) \]

\[
\mathcal{L}_{ce}(x) = -\sum_{i=2}^{K} (1 - p(y_1|x)) t_i \log p(y_i|x)
\]

\[ T = \{t_2, t_3, \ldots, t_K\} \text{ denotes the frame-level target labels.} \]

• Need frame-level alignment
  • Still End-to-End?
Experiments

• Data: 1k, 4k, 20k hours
  • a normal test set and a fast speed test set

• Feature: 80-dim FBK
  • stack the consecutive frames(±5)
  • Subsample with 3
## Results

- **Baseline**

<table>
<thead>
<tr>
<th>Method</th>
<th>Label</th>
<th>Model Size (MB)</th>
<th>Time/Epoch (Hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLSTM-CE</td>
<td>CD-Phone</td>
<td>155</td>
<td>3.67</td>
</tr>
<tr>
<td>DFSMN-CE</td>
<td>CD-Phone</td>
<td>114</td>
<td>0.50</td>
</tr>
<tr>
<td>DFSMN-CTC</td>
<td>CD-Phone</td>
<td>114</td>
<td>0.58</td>
</tr>
<tr>
<td>DFSMN-CTC</td>
<td>CI-Phone</td>
<td>97</td>
<td>0.43</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Label</th>
<th>Data (Hours)</th>
<th>Test set (WER %)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Normal</td>
<td>Fast</td>
</tr>
<tr>
<td>BLSTM-CE</td>
<td>CD-Phone</td>
<td>1k</td>
<td>19.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4k</td>
<td>16.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20k</td>
<td>13.97</td>
</tr>
<tr>
<td>DFSMN-CE</td>
<td>CD-Phone</td>
<td>1k</td>
<td>18.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4k</td>
<td>14.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20k</td>
<td>12.10</td>
</tr>
<tr>
<td>DFSMN-CTC</td>
<td>CI-Phone</td>
<td>1k</td>
<td>17.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4k</td>
<td>13.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20k</td>
<td>11.46</td>
</tr>
<tr>
<td>DFSMN-CTC</td>
<td>CD-Phone</td>
<td>1k</td>
<td>16.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4k</td>
<td>13.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20k</td>
<td>11.71</td>
</tr>
</tbody>
</table>
Results

• Joint CTC-CE
• CD-Phone

<table>
<thead>
<tr>
<th>Method</th>
<th>Alpha</th>
<th>Test set (WER %)</th>
<th>Normal</th>
<th>Gain</th>
<th>Fast</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>-</td>
<td>12.10</td>
<td>-</td>
<td>29.79</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>CTC</td>
<td>-</td>
<td>11.71</td>
<td>3.2%</td>
<td>24.04</td>
<td>19.3%</td>
<td></td>
</tr>
<tr>
<td>Joint CTC</td>
<td>0.1</td>
<td>10.92</td>
<td>9.8%</td>
<td>21.68</td>
<td>27.2%</td>
<td></td>
</tr>
<tr>
<td>CTC CE</td>
<td>0.5</td>
<td>10.67</td>
<td>11.8%</td>
<td>21.98</td>
<td>26.2%</td>
<td></td>
</tr>
<tr>
<td>CTC CE</td>
<td>1.0</td>
<td>10.77</td>
<td>11.0%</td>
<td>20.80</td>
<td>30.1%</td>
<td></td>
</tr>
<tr>
<td>CTC CE</td>
<td>2.0</td>
<td>11.03</td>
<td>8.8%</td>
<td>22.86</td>
<td>23.3%</td>
<td></td>
</tr>
</tbody>
</table>
Results

• Joint CTC-CE
  • accurate alignment
Extending Recurrent Neural Aligner for Streaming End-to-End Speech Recognition in Mandarin

Linhao Dong\textsuperscript{1,2}, Shiyu Zhou\textsuperscript{1,2}, Wei Chen\textsuperscript{1}, Bo Xu\textsuperscript{1}

\textsuperscript{1}Institute of Automation, Chinese Academy of Sciences, China
\textsuperscript{2}University of Chinese Academy of Sciences, China
Motivations

• English->Chinese

• Recurrent Neural Aligner (RNA)
  • streaming recognition

• Improve by:
  • redesign the temporal down-sampling and introduce a powerful convolutional structure.
  • In the decoder, we utilize a regularizer to smooth the output distribution and conduct joint training with a language model.
RNA

- $e_u$ is the encoded vector of $z_u$
- Diff with CTC in
  - the conditional distribution
    \[
    p(z|x) = \prod_u p(z_u | z_{u-1}^u, x)
    \]
    \[
    p(z|x) = \prod_u p(z_u | x)
    \]
  - RNA obtains the predicted output sequence by simply removing the blanks from alignment, while the CTC model needs to remove first the repeated labels and then the blanks

\[
\begin{align*}
  h &= \text{encoder}(x) \\
  z_u &= \arg \max_{l \in [1, L+1]} (\text{decoder}(h_u, e_{u-1})) \\
  p(y|x) &= \sum_z p(z|x)
\end{align*}
\]
Temporal down-sampling

- Pooling between LSTMs
- Strided convolutional layers
Multiplicative Units

\[ g_1 = \sigma(W_1 \ast I + b_1) \]
\[ g_2 = \sigma(W_2 \ast I + b_2) \]
\[ g_3 = \sigma(W_3 \ast I + b_3) \]
\[ u = \tanh(W_4 \ast I + b_4) \]
\[ MU(h; W) = g_1 \odot \tanh(g_2 \odot h + g_3 \odot u + b_5) \]
Confidence Penalty

- Label Smoothing
- Obtain better generalization

\[ H(p(\theta(z|x))) = - \sum_{u \in [1, U]} \sum_{z_u \in [1, L+1]} p(\theta(z_u|x)) \log(p(\theta(z_u|x))) \]

\[ L(\theta) = \sum_{(x, y)} -\log(p(\theta(y|x))) - \lambda \sum_{x} H(p(\theta(z|x))) \]
Joint training with RNN-LM

• Difficult:
  • If we use the shallow fusion in, it’s hard to obtain accurate alignments containing blank for training the LM.
  • If we use the mechanism of joint training with RNN-LM, the blank label hampers the synchronism between the outputs of RNA and the RNN-LM
Joint training with RNN-LM

- Let $h^{\{LM\}}_u$ represents the LM state
  - uses the current output of LM-RNN if $z_{\{u-1\}}$ is non-blank
  - uses the previous output of LM-RNN if $z_{\{u-1\}}$ is blank

\[
g_u = \sigma(W_1 \cdot [s_u; h_u^{LM}] + b_1) \\
s_u^F = [s_u; g_u \odot h_u^{LM}] \\
p(z_u|z_{1}^{u-1}, x) = \text{softmax}(W_2 \cdot s_u^F + b_2)
\]
Exp

- HKUST
Exp

• Temporal down-sampling

<table>
<thead>
<tr>
<th>Down-sampling mechanism</th>
<th>Rate</th>
<th>CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>frame stacking and sub-sampling [5]</td>
<td>1/3</td>
<td>43.19</td>
</tr>
<tr>
<td>pooling{2,4}-width{2,2}</td>
<td>1/4</td>
<td>39.80</td>
</tr>
<tr>
<td>pooling{2,4}-width{3,2}</td>
<td>1/6</td>
<td>34.07</td>
</tr>
<tr>
<td>pooling{1,2,4}-width{2,2,2}</td>
<td>1/8</td>
<td>31.94</td>
</tr>
<tr>
<td>pooling{1,2,4}-width{3,2,2}</td>
<td>1/12</td>
<td>33.53</td>
</tr>
<tr>
<td>pooling{1,2,3,4}-width{2,2,2,2}</td>
<td>1/16</td>
<td>36.63</td>
</tr>
<tr>
<td>conv-stride{2,2,2}</td>
<td>1/8</td>
<td>34.78</td>
</tr>
<tr>
<td>conv-stride{2,2} + pooling{2}-width{2}</td>
<td>1/8</td>
<td>32.62</td>
</tr>
<tr>
<td>conv-stride{2} + pooling{2,4}-width{2,2}</td>
<td>1/8</td>
<td>30.86</td>
</tr>
</tbody>
</table>
Exp

- further extensions on RNA

<table>
<thead>
<tr>
<th>Model-ID</th>
<th>Model</th>
<th>CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$</td>
<td>RNA with the best down-sampling</td>
<td>30.86</td>
</tr>
<tr>
<td>$M_2$</td>
<td>$M_1 + 1 \times \text{MU}$</td>
<td>29.89</td>
</tr>
<tr>
<td></td>
<td>$M_1 + 1 \times \text{ConvLSTM}$</td>
<td>30.55</td>
</tr>
<tr>
<td></td>
<td>$M_1 + 1 \times \text{GLU}$</td>
<td>30.36</td>
</tr>
<tr>
<td>$M_3$</td>
<td>$M_2 + \text{Confidence Penalty (} \lambda = 0.2 \text{)}$</td>
<td>29.06</td>
</tr>
<tr>
<td>$M_4$</td>
<td>$M_3 + \text{Joint training with RNN-LM}$</td>
<td>28.32</td>
</tr>
</tbody>
</table>
Syllable-Based Sequence-to-Sequence Speech Recognition with the Transformer in Mandarin Chinese

Shiyu Zhou\textsuperscript{1,2}, Linhao Dong\textsuperscript{1,2}, Shuang Xu\textsuperscript{1}, Bo Xu\textsuperscript{1}

\textsuperscript{1}Institute of Automation, Chinese Academy of Sciences
\textsuperscript{2}University of Chinese Academy of Sciences
Motivation

• Transformer achieves a state-of-the-art BLEU on NMT
• Extend it to speech as the basic architecture of sequence-to-sequence attention-based model on Mandarin Chinese ASR
• Investigate a comparison between syllable based model and context-independent phone based model
  • syllables have the advantage of avoiding OOV problem
• A greedy cascading decoder with the Transformer is proposed for mapping CI-phoneme sequences and syllable sequences into word sequences
Transformer model

- the same as sequence-to-sequence attention-based models except relying entirely on self-attention and position-wise
  - Encoder
  - Decoder
  - Multi-head attention

\[
\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V
\]

\[
\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O
\]

where \(\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)\)
Transformer model

- MHA and position-wise, fully connected layers for both the encode and decoder
- positional encodings
Greedy cascading decoder

• First, the best sub-word unit sequence $s$ is calculated by the Transformer from observation $X$ to sub-word unit sequence with beam size $\beta$.

• Then, the best word sequence $W$ is chosen by the Transformer from sub-word unit sequence to word sequence with beam size $\gamma$.

$$\bar{W} = \arg\max_w Pr(W|X)$$

$$= \arg\max_w \sum_s Pr(W|s)Pr(s|X)$$

$$\approx \arg\max_w Pr(W|s)Pr(s|X)$$
Exp

- HKUST
- CI-phoneme: 122
- syllable: 1388
Exp

• CI-phoneme and syllable based model

Table 2: Comparison of CI-phoneme and syllable based model with the Transformer on HKUST datasets in CER (%).

<table>
<thead>
<tr>
<th>sub-word unit</th>
<th>model</th>
<th>CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI-phonemes</td>
<td>D512-H8</td>
<td>32.94</td>
</tr>
<tr>
<td></td>
<td>D1024-H16</td>
<td>30.65</td>
</tr>
<tr>
<td></td>
<td>D1024-H16 (speed perturb)</td>
<td>30.72</td>
</tr>
<tr>
<td>Syllables</td>
<td>D512-H8</td>
<td>31.80</td>
</tr>
<tr>
<td></td>
<td>D1024-H16</td>
<td>29.87</td>
</tr>
<tr>
<td></td>
<td>D1024-H16 (speed perturb)</td>
<td>28.77</td>
</tr>
</tbody>
</table>
Exp

- Comparison with previous works

Table 3: CER (%) on HKUST datasets compared to previous works.

<table>
<thead>
<tr>
<th>model</th>
<th>CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTMP-9\times800P512-F444 [24]</td>
<td>30.79</td>
</tr>
<tr>
<td>CTC-attention+joint dec. (speed perturb., one-pass)</td>
<td>28.9</td>
</tr>
<tr>
<td>+VGG net</td>
<td>28.0</td>
</tr>
<tr>
<td>+RNN-LM (separate) [9]</td>
<td></td>
</tr>
<tr>
<td>CI-phonemes-D1024-H16</td>
<td>30.65</td>
</tr>
<tr>
<td>Syllables-D1024-H16 (speed perturb)</td>
<td>28.77</td>
</tr>
</tbody>
</table>
Exp

• Comparison of different frame rates

Table 4: Comparison of different frame rates on HKUST datasets in CER (%).

<table>
<thead>
<tr>
<th>model</th>
<th>frame rate</th>
<th>CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI-phonemes-D1024-H16 (speed perturb)</td>
<td>30ms</td>
<td>30.72</td>
</tr>
<tr>
<td></td>
<td>50ms</td>
<td>31.68</td>
</tr>
<tr>
<td></td>
<td>70ms</td>
<td>33.96</td>
</tr>
<tr>
<td>Syllables-D1024-H16 (speed perturb)</td>
<td>30ms</td>
<td>28.77</td>
</tr>
<tr>
<td></td>
<td>50ms</td>
<td>29.36</td>
</tr>
<tr>
<td></td>
<td>70ms</td>
<td>32.22</td>
</tr>
</tbody>
</table>