LEARNING NEURAL TRANS-DIMENSIONAL RANDOM FIELD LANGUAGE MODELS WITH NOISE-CONTRASTIVE ESTIMATION

Bin Wang, Zhijian Ou
Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University, Beijing, China.
wangbin12@mails.tsinghua.edu.cn, ozj@Tsinghua.edu.cn

Introduction

Trans-dimensional random field (TRF) LMs
◆ To fit the joint probability \( p(x_1, \ldots, x_l) \) directly
◆ Support both discrete features and neural network features
◆ Inference is fast but training is slow

To improve the training efficiency and the performance of neural TRF LMs:
✓ Define the TRF in the form of exponential tilting of a reference distribution
✓ Introduce the noise-contrastive estimation (NCE) to train TRF LM.
✓ Marry the deep CNN and the bi-directional LSTM

Model Definition

\[
p_m(x^l; \theta, \zeta) = \pi_l q(x^l) e^{\phi(x^l; \theta) - \zeta_l}
\]

\( x^l = (x_1, \ldots, x_l) \) a word sequence of length \( l \)
\( \pi_l \) the prior length probability
\( q(x^l) \) an LSTM language model
\( \zeta_l \) the normalization constant of length \( l \) need to be estimated
\( \phi(x^l; \theta) \) potential function with parameter \( \theta \)

Noise-contrastive Estimation (NCE)

\[
\text{model distribution: } C = 0 \quad P(C = 0 | x^l) = \frac{p_m}{p_m + \nu p_n}
\]

\[
\text{noise distribution: } C = 1 \quad P(C = 1 | x^l) = 1 - P(C = 0 | x^l)
\]

\[
\max_{\theta, \zeta} \frac{1}{|D|} \sum_{x^l \in D} \log P(C = 0 | x^l) + \frac{\nu}{|B|} \sum_{x^l \in B} \log P(C = 1 | x^l)
\]

Experiments

Speech recognition WERs on CHiME-4 Challenge data.

<table>
<thead>
<tr>
<th>model</th>
<th>Dev real</th>
<th>Dev simu</th>
<th>Test real</th>
<th>Test simu</th>
</tr>
</thead>
<tbody>
<tr>
<td>KN5</td>
<td>5.03</td>
<td>4.79</td>
<td>7.38</td>
<td>5.78</td>
</tr>
<tr>
<td>LSTM (i.e. ( q(x^l) ))</td>
<td>3.63</td>
<td>3.24</td>
<td>5.70</td>
<td>4.53</td>
</tr>
<tr>
<td>TRF</td>
<td>3.53</td>
<td>3.20</td>
<td>5.68</td>
<td>4.36</td>
</tr>
<tr>
<td>KN5+LSTM</td>
<td>3.56</td>
<td>3.29</td>
<td><strong>5.71</strong></td>
<td>4.18</td>
</tr>
<tr>
<td>KN5+TRF</td>
<td>3.53</td>
<td>3.22</td>
<td>5.54</td>
<td>4.20</td>
</tr>
<tr>
<td>KN5+LSTM+TRF</td>
<td>3.42</td>
<td>3.10</td>
<td><strong>5.44</strong></td>
<td>4.13</td>
</tr>
</tbody>
</table>

◆ KN5: 5gram LM with modified Kneser-Ney smoothing
◆ LSTM: 2 hidden layers, 512 hidden units per layer
◆ “Dev” denotes the development set and “Test” denotes the test set.
◆ “+” denotes the log-linear interpolation with equal weights

Conclusion:
✓ On a 40x larger training set use only \( 1/3 \) training time
✓ Achieve a **4.7%** relative WER reduction on the top of a strong LSTM LM baseline.