Joint CTC-Attention based End-to-end Speech Recognition using Multi-task Learning

Suyoun Kim, Takaaki Hori, and Shinji Watanabe

Presenter: Desh Raj
Outline

• CTC and attention—the good and the bad
• The joint CTC-attention model
• Experimental results
End-to-end ASR

- Several issues with hybrid DNN-HMM models
- Several independent moving components—acoustic model, language model, lexicon, etc.
- Make conditional independence assumptions and approximations
- End-to-end models learn acoustic frames to character mapping
End-to-end ASR

- Two main approaches:
  1. Connectionist temporal classification (CTC)
  2. Attention-based encoder decoder
CTC

- Uses intermediate label representation—allows repetitions of labels and occurrence of a blank label

\[ P(y|x) = \sum_{\pi \in \Phi(y')} P(\pi|x), \]

Sum over all possible intermediate label representations
CTC

- Uses intermediate label representation—allows repetitions of labels and occurrence of a blank label

\[
P(y|x) = \sum_{\pi \in \Phi(y')} P(\pi|x),
\]

Conditional independence of output labels

\[
P(\pi|x) \approx \prod_{t=1}^{T} P(\pi_t|x) = \prod_{t=1}^{T} q_t(\pi_t)
\]
CTC

• Uses intermediate label representation—allows repetitions of labels and occurrence of a blank label

\[ P(y|x) = \sum_{\pi \in \Phi(y')} P(\pi|x), \]

• Can just use forward-backward to compute
Attention

- No conditional independence assumptions

\[ P(y|x) = \prod_u P(y_u|x, y_{1:u-1}) \]

\[ h = \text{Encoder}(x) \]

\[ y_u \sim \text{AttentionDecoder}(h, y_{1:u-1}). \]
Attention

Decoder Network

Output Sequence

s0 \rightarrow s1 \rightarrow s2 \rightarrow \cdots \rightarrow sj

Attention Mechanism

v0 \rightarrow v1 \rightarrow v2 \rightarrow \cdots \rightarrow vj

Encoder Network

Input Sequence

x1 \rightarrow x2 \rightarrow x3 \rightarrow \cdots \rightarrow xT

h1 \rightarrow h2 \rightarrow h3 \rightarrow \cdots \rightarrow hT
Attention

- No conditional independence assumptions

\[ P(y|x) = \prod_u P(y_u|x, y_{1:u-1}) \]

\[ h = \text{Encoder}(x) \]

\[ y_u \sim \text{AttentionDecoder}(h, y_{1:u-1}) \]

Can be content-based or location-based
Attention

• So what’s the problem?

• Too much flexibility—easily affected by noise.

• Also hard to train from scratch on long input sequences.
Attention

- So what’s the problem?
  - Too much flexibility—easily affected by noise.
  - Also hard to train from scratch on long input sequences.

CTC models don’t have these problems since they impose left-to-right constraints.
Joint CTC-Attention
Joint CTC-Attention

\[ \mathcal{L}_{\text{MTL}} = \lambda \mathcal{L}_{\text{CTC}} + (1 - \lambda) \mathcal{L}_{\text{Attention}} \]
Experiments

- 3 datasets—WSJ1 (81 hours), WSJ0 (15 hours), and Chime-4 (18 hours)

- 40 Mel-scale filterbank coefficients + first and second order temporal derivatives = 120 feature values

- No LM or lexicon used
Experiments

- Encoder—4-layer Bi-LSTM
- Top 2 layers perform sequence contraction by half each
- Decoder—1-layer LSTM
Results

<table>
<thead>
<tr>
<th>Model (train)</th>
<th>CER (valid)</th>
<th>CER (eval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ-train_si284 (80hrs)</td>
<td>dev93</td>
<td>eval92</td>
</tr>
<tr>
<td>CTC</td>
<td>11.48</td>
<td>8.97</td>
</tr>
<tr>
<td>Attention (content-based)</td>
<td>13.68</td>
<td>11.08</td>
</tr>
<tr>
<td>Attention (location-based)</td>
<td>11.98</td>
<td>8.17</td>
</tr>
<tr>
<td>MTL (λ = 0.2)</td>
<td><strong>11.27</strong></td>
<td><strong>7.36</strong></td>
</tr>
<tr>
<td>MTL (λ = 0.5)</td>
<td>12.00</td>
<td>8.31</td>
</tr>
<tr>
<td>MTL (λ = 0.8)</td>
<td>11.71</td>
<td>8.45</td>
</tr>
<tr>
<td>WSJ-train_si84 (15hrs)</td>
<td>dev93</td>
<td>eval92</td>
</tr>
<tr>
<td>CTC</td>
<td>27.41</td>
<td>20.34</td>
</tr>
<tr>
<td>Attention (content-based)</td>
<td>28.02</td>
<td>20.06</td>
</tr>
<tr>
<td>Attention (location-based)</td>
<td>24.98</td>
<td>17.01</td>
</tr>
<tr>
<td>MTL (λ = 0.2)</td>
<td><strong>23.03</strong></td>
<td><strong>14.53</strong></td>
</tr>
<tr>
<td>MTL (λ = 0.5)</td>
<td>26.28</td>
<td>16.24</td>
</tr>
<tr>
<td>MTL (λ = 0.8)</td>
<td>32.21</td>
<td>21.30</td>
</tr>
<tr>
<td>CHiME-4-tr05_multi (18hrs)</td>
<td>dt05_real</td>
<td>et05_real</td>
</tr>
<tr>
<td>CTC</td>
<td>37.56</td>
<td>48.79</td>
</tr>
<tr>
<td>Attention (content-based)</td>
<td>43.45</td>
<td>54.25</td>
</tr>
<tr>
<td>Attention (location-based)</td>
<td>35.01</td>
<td>47.58</td>
</tr>
<tr>
<td>MTL (λ = 0.2)</td>
<td><strong>32.08</strong></td>
<td><strong>44.99</strong></td>
</tr>
<tr>
<td>MTL (λ = 0.5)</td>
<td>34.56</td>
<td>46.49</td>
</tr>
<tr>
<td>MTL (λ = 0.8)</td>
<td>35.41</td>
<td>48.34</td>
</tr>
</tbody>
</table>

Clean environment—possible that CTC improved generalization since its training does not use character inter-dependencies.
## Results

<table>
<thead>
<tr>
<th>Model(train)</th>
<th>CER(valid)</th>
<th>CER(eval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ-train_si284 (80hrs)</td>
<td>dev93</td>
<td>eval92</td>
</tr>
<tr>
<td>CTC</td>
<td>11.48</td>
<td>8.97</td>
</tr>
<tr>
<td>Attention(content-based)</td>
<td>13.68</td>
<td>11.08</td>
</tr>
<tr>
<td>Attention(location-based)</td>
<td>11.98</td>
<td>8.17</td>
</tr>
<tr>
<td>MTL($\lambda = 0.2$)</td>
<td><strong>11.27</strong></td>
<td><strong>7.36</strong></td>
</tr>
<tr>
<td>MTL($\lambda = 0.5$)</td>
<td>12.00</td>
<td>8.31</td>
</tr>
<tr>
<td>MTL($\lambda = 0.8$)</td>
<td>11.71</td>
<td>8.45</td>
</tr>
<tr>
<td>WSJ-train_si84 (15hrs)</td>
<td>dev93</td>
<td>eval92</td>
</tr>
<tr>
<td>CTC</td>
<td>27.41</td>
<td>20.34</td>
</tr>
<tr>
<td>Attention(content-based)</td>
<td>28.02</td>
<td>20.06</td>
</tr>
<tr>
<td>Attention(location-based)</td>
<td>24.98</td>
<td>17.01</td>
</tr>
<tr>
<td>MTL($\lambda = 0.2$)</td>
<td><strong>23.03</strong></td>
<td><strong>14.53</strong></td>
</tr>
<tr>
<td>MTL($\lambda = 0.5$)</td>
<td>26.28</td>
<td>16.24</td>
</tr>
<tr>
<td>MTL($\lambda = 0.8$)</td>
<td>32.21</td>
<td>21.30</td>
</tr>
<tr>
<td>CHiME-4-tr05_multi (18hrs)</td>
<td>dt05_real</td>
<td>et05_real</td>
</tr>
<tr>
<td>CTC</td>
<td>37.56</td>
<td>48.79</td>
</tr>
<tr>
<td>Attention(content-based)</td>
<td>43.45</td>
<td>54.25</td>
</tr>
<tr>
<td>Attention(location-based)</td>
<td>35.01</td>
<td>47.58</td>
</tr>
<tr>
<td>MTL($\lambda = 0.2$)</td>
<td><strong>32.08</strong></td>
<td><strong>44.99</strong></td>
</tr>
<tr>
<td>MTL($\lambda = 0.5$)</td>
<td>34.56</td>
<td>46.49</td>
</tr>
<tr>
<td>MTL($\lambda = 0.8$)</td>
<td>35.41</td>
<td>48.34</td>
</tr>
</tbody>
</table>

Noisy environment—much better than attention-based model
Results

CTC trains quickly but low accuracy

Fig. 2: Comparison of learning curves: CTC, location-based attention model, and MTL with $\lambda = 0.2, 0.5, 0.8$. The character accuracy on the validation set of CHiME-4 is calculated by edit distance between hypothesis and reference. Note that the reference history were used in the attention and our MTL models.
Results

Fig. 2: Comparison of learning curves: CTC, location-based attention model, and MTL with ($\lambda = 0.2, 0.5, 0.8$). The character accuracy on the validation set of CHiME-4 is calculated by edit distance between hypothesis and reference. Note that the reference history were used in the attention and our MTL models.

Attention-based model reaches same accuracy as MTL but takes twice as much time
Results

Fig. 2: Comparison of learning curves: CTC, location-based attention model, and MTL with ($\lambda = 0.2, 0.5, 0.8$). The character accuracy on the validation set of CHiME-4 is calculated by edit distance between hypothesis and reference. Note that the reference history were used in the attention and our MTL models.
Results

Attention alignments between characters and acoustic frames

(a) Attention 1 epoch  
(b) Attention 3 epoch  
(c) Attention 5 epoch  
(d) Attention 7 epoch  
(e) Attention 9 epoch  
(f) MTL 1 epoch  
(g) MTL 3 epoch  
(h) MTL 5 epoch  
(i) MTL 7 epoch  
(j) MTL 9 epoch
Results

Does not learn desired alignments even after 9 epochs
Results

Learns desired alignment after 5 epochs
Key takeaways

• Combining CTC and attention performs better on both clean and noisy data

• Speeds up training significantly

• Also gives desired alignments unlike attention
Thank you!

Questions? Comments?