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

Suyoun Kim $^{1,2}$, Takaaki Ḥori $^1$, and Shinji Watanabe

$^1$Mitsubishi Electric Research Laboratories (MERL)
$^2$Carnegie Mellon University (CMU)

ICASSP March 6 2017
Presenter: Suyoun Kim
Outline

• Introduction and motivation
• Our proposed model: Joint CTC/Attention
• Experiments and results
• Conclusion
Automatic Speech Recognition (ASR)

- ASR is transcribing speech signal to text
- Conventional ASR system is split into multiple sub-components

**Input**
- Feature Processing
- Feature sequence

**Output**
- Word sequence
- A CAT

---

**Diagram Components**
- Feature sequence
- Adaptation
- GMM-HMM Acoustic Model
- Language Model
- WFST Decoder
- DNN-HMM Acoustic Model
- Pronunciation Model
- Sequence Training
  - sMBR

---

© MERL
Conventional ASR is Complicated

• Many sub-components
  – System development is **complicated**
  – Separate modeling may cause **suboptimal**
  – Decoding algorithm is **complex**

• Many assumptions
  – Assumes future process only depends on current state not previous state (Markovian, Stationary)
    • \( P(s_{t+1}|s_{1:T}) = P(s_{t+1}|s_t) \)
    • \( P(s_{t_1+1} = i|s_{t_1} = j) = P(s_{t_2+1} = i|s_{t_2} = j) \) for any \( t_1 \) and \( t_2 \)
  – Assumes observations are independent given state (Conditional independent)
    • \( P(x_t|x_{1:T}, s_{1:T}) = P(x_t|s_t) \)
  – Assumes all pronunciations can be represented by several phonemes (hand-crafted knowledge)
    • Linguistic expertise is required
End-to-End ASR is transcribing speech signal to text directly with a single model, one step training

- Big data and powerful computational engine
- Deep Learning e.g. Recurrent Neural Network (RNN)
- End-to-End approach

Feature sequence → End-to-End Model → Word sequence

A CAT
Our Joint CTC/Attention model for End-to-End ASR

• Key insight:
  – We can address the weaknesses of two main End-to-End approaches 1) CTC, and 2) Attention model by combining the two, as they have complementary characteristics.

CTC + Attention model → Our Joint CTC/Attention
End-to-End approach 1: Connectionist Temporal Classification (CTC) [Graves(2006)]

- It uses **intermediate label representation** $\pi$ allowing **repetitions** and **blank** labels “_”
- It maximizes the **total probability** of all possible label sequence $\pi$
- It uses **forward-backward algorithm** for the efficient training

**Strength:** There is no need for pronunciation model

**Weakness:** It still relies on conditional independence assumption, typically separate LM is combined
Our Joint CTC/Attention model for End-to-End ASR

- We keep our model **simple**
  - By using Attention model to learn LM jointly

<table>
<thead>
<tr>
<th>CTC</th>
<th>Attention model</th>
<th>Our Joint CTC/Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>❌ It requires separate LM</td>
<td>🔴 It can learn LM jointly</td>
<td>🔵 It can learn LM jointly =&gt; Simple</td>
</tr>
</tbody>
</table>
End-to-End approach 2: **Attention-based Encoder-Decoder** [Chorowski (2014)]

- It uses two RNNs 1) Encoder 2) AttentionDecoder
- For each output step, it estimates weight vector (alignment) over inputs and then decoder uses **weighted sum input**
- Decoder estimates each label **conditioning on previous outputs** (no conditional independent assumption)

**Strength**: It can learn acoustic and language model within a single network

**Weakness**: The alignment can be easily distorted
We regularize input/output alignment of attention

• Unlike CTC, Attention model does not preserve order of inputs
• Our desired alignment in ASR task is **monotonic**
• Not regularized alignment makes the model **hard to learn** from scratch

Example of **monotonic alignment**

Example of **distorted alignment**
Our Joint CTC/Attention model for End-to-End ASR

- We keep our model **simple**
  - By using Attention model to learn inter-character dependencies jointly
- We improve the **learning speed** and **performance**
  - By using CTC to regularize the input/output alignment

### CTC
- ❌ It requires separate LM
- ✔️ It preserves input/output order

### Attention model
- ✔️ It can model LM jointly
- ❌ Input/output alignment is easily distorted
- ❌ It is hard to train from scratch

### Our Joint CTC/Attention
- ✔️ It can learn LM jointly
  ⇒ **Simple**
- ✔️ It can regularize the input/output alignment
  ⇒ **Faster convergence**
  ⇒ **Better performance**
Our Joint CTC/Attention model for End-to-End ASR

- Standard Attention model
Our Joint CTC/Attention model for End-to-End ASR

- Multi-task learning framework
Our Joint CTC/Attention model for End-to-End ASR

1. We share the encoder part
2. We train Attention model with CTC jointly

\[
L_{\text{MTL}} = \lambda L_{\text{CTC}} + (1 - \lambda) L_{\text{Attention}}
\]

Larger \( \lambda \) will give more weight on CTC objective.

<table>
<thead>
<tr>
<th>Global normalization</th>
<th>Local normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ L_{\text{CTC}} \triangleq -\ln P(y^*</td>
<td>x) = -\ln \sum_{\pi \in \Phi(y')} P(\pi</td>
</tr>
</tbody>
</table>

3. We use AttentionDecoder on decoding mode
   - The cost for CTC exists only on training mode
Outline

• Introduction and motivation
• Our proposed model: Joint CTC/Attention
• Experiments and results
• Conclusion
Experiment setup

- Dataset
  - WSJ0 (si84) – 15 hours clean
  - WSJ1 (si284) – 80 hours clean
  - CHiME4 – 18 hours noisy
  - Input – 120d filterbank (+d, +dd)
  - Output – 32 distinct label (+26 char, + apostrophe, period, …, sos/eos)

- Baselines
  - CTC – 4 layer BLSTM (320 cells)
  - Attention – 4 layer BLSTM encoder (320 cells) + 1 layer LSTM decoder (320 cells), location-based attention mechanism

- Our Joint CTC/Attention model
  - 4 layer BLSTM encoder (320 cells) + 1 layer LSTM decoder (320 cells)
    - With $\lambda = \{0.2 \ 0.5 \ 0.8\}$

- Evaluation
  - Character Error Rate (CER)
Faster convergence compared to Attention model

Good! Converge Fast
9.9% relative improvement of CER on WSJ1(80hr)

<table>
<thead>
<tr>
<th></th>
<th>Dev</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTC</td>
<td>11.5</td>
<td>9.0</td>
</tr>
<tr>
<td>Attention</td>
<td>12.0</td>
<td>8.2</td>
</tr>
<tr>
<td><strong>OurModel (λ=0.2)</strong></td>
<td>11.3</td>
<td>7.4</td>
</tr>
<tr>
<td><strong>OurModel (λ=0.5)</strong></td>
<td>12.0</td>
<td>8.3</td>
</tr>
<tr>
<td><strong>OurModel (λ=0.8)</strong></td>
<td>11.7</td>
<td>8.5</td>
</tr>
</tbody>
</table>

WER of our best system was 18.2%
WER of (Bahdanau, et al. ICASSP 2016) was 18.6%
14.6% relative improvement of CER on WSJ0(15hr)

<table>
<thead>
<tr>
<th></th>
<th>Dev</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTC</td>
<td>27.4</td>
<td>20.3</td>
</tr>
<tr>
<td>Attention</td>
<td>25.0</td>
<td>17.0</td>
</tr>
<tr>
<td>OurModel ($\lambda=0.2$)</td>
<td>23.0</td>
<td>14.5</td>
</tr>
<tr>
<td>OurModel ($\lambda=0.5$)</td>
<td>26.3</td>
<td>16.2</td>
</tr>
<tr>
<td>OurModel ($\lambda=0.8$)</td>
<td>32.2</td>
<td>21.3</td>
</tr>
</tbody>
</table>

Lower is Better!

$\lambda=0.2$ performs best!

Larger $\lambda$ gives more weight on CTC

14.6% improvement
5.4% relative improvement of CER on CHiME4(18hr)

Lower is Better!

<table>
<thead>
<tr>
<th></th>
<th>Dev</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTC</td>
<td>37.6</td>
<td>48.8</td>
</tr>
<tr>
<td>Attention</td>
<td>35.0</td>
<td>47.6</td>
</tr>
<tr>
<td><strong>OurModel</strong> ($\lambda=0.2$)</td>
<td><strong>32.1</strong></td>
<td><strong>45.0</strong></td>
</tr>
<tr>
<td><strong>OurModel</strong> ($\lambda=0.5$)</td>
<td>34.6</td>
<td>46.5</td>
</tr>
<tr>
<td><strong>OurModel</strong> ($\lambda=0.8$)</td>
<td>35.4</td>
<td>48.3</td>
</tr>
</tbody>
</table>

$\lambda=0.2$ performs best!

Larger $\lambda$ gives more weight on CTC

5.4% improvement
More robust input/output alignment of attention

- Alignment of one selected utterance from CHiME4
Outline

• Introduction and motivation
• Our proposed model: Joint CTC/Attention
• Experiments and results
• Conclusion
Conclusion

• Joint CTC/Attention model
  – does not use any linguistic information
  – shows 5.4 – 14.6 % relative improvements in CER, compared to Attention-based Encoder-Decoder
  – speeds up learning process
  – requires small additional computational cost but only in training mode, not in decoding mode.

• Our framework can be applied to other seq2seq tasks where its alignment is monotonic
Current research

- Further experimental results on Corpus of Spontaneous Japanese (CSJ) – 581hr
  - Achieved comparable performance to state-of-the-art

<table>
<thead>
<tr>
<th></th>
<th>task1</th>
<th>task2</th>
<th>task3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention (581h)</td>
<td>11.5</td>
<td>7.9</td>
<td>9.0</td>
</tr>
<tr>
<td>OurModel (581h)</td>
<td>10.9</td>
<td>7.8</td>
<td>8.3</td>
</tr>
<tr>
<td><strong>OurModel2 (581h)</strong></td>
<td>9.5</td>
<td>7.0</td>
<td>7.8</td>
</tr>
<tr>
<td>DNN/sMBR-hybrid (236h for AM/ 581h for LM)</td>
<td>9.0</td>
<td>7.2</td>
<td>9.6</td>
</tr>
<tr>
<td>CTC-syllable (581h)</td>
<td>9.4</td>
<td>7.3</td>
<td>7.5</td>
</tr>
</tbody>
</table>
Thank you!

Questions & Answers