Phoneme-based Neural Transducer for Large Vocabulary Speech Recognition

Weichao Zhou, Simon Berger, Ralf Schlüter, Hermann Ney

Phoneme-based Neural Transducer (TLv2)

### Motivation

- Classical hybrid hidden Markov model (HMM)
  - **pros:** flexibility (modularity), scalability to low-resource tasks
  - **cons:** complexity, inconsistency of modeling
- End-to-end automatic speech recognition (ASR)
  - **pros:** simplicity, consistent training & inference
  - **cons:** flexibility, scalability, amount of data & training time

**Goal:** join the advantages of both approaches

### Phoneme-Based Neural Transducer

**Model definition**

\[ p(a_i^T | x^T) = \sum_{(a_j^u, a_j^v)} p(y_j^u, s_j^v | h_i^T) \]

- input feature sequence \( x^T \)
- alignment sequence \( a_i^T \) \( \rightarrow \) \( h_i^T \)
- encoder output \( y_j^u \) \( \rightarrow \) \( s_j^v \) \( \rightarrow \) \( y_j^u \) \( \rightarrow \) \( a_j^u \)
- output label sequence \( a_j^u \)

- context size \( k \) (default 1): local dependency (co-articulation)

**HMM alignment label topology**

- \( a_i^u \): each \( a_i \) can loop for multiple steps and no blank \( a_i^u \)

**Decision & Decoding**

- external word-level language model (LM) and lexicon
- no internal LM (2019) applied: suppressed negative effect

### Simplification and Aachment

**Simplified NN architecture**

- recurrent neural network transducer (RNN-T) [Graves 2012]
- encoder: \( 6 \times 512 \) bidirectional long short-term memory (BiLSTM) with subsampling of factor 2 using max-pooling
- feed-forward neural network (FFNN)-based prediction network
- joint network (element-wise addition) and a final softmax
- footprint: about 30M parameters

**Viterbi training**

- full-sum (FS) over all alignments: time and memory consuming
- frame-wise cross-entropy (CE) loss w.r.t. \( p(y_j^u, s_j^v | h_i^T) \) and a fixed external alignment
- enable more training techniques for speed and performance

**Word boundary-based phoneme label augmentation**

- end-of-word (EOW) phonemes: \( 2 \times |V| \)
- start-of-word (SOW) + EOW phonemes: \( 4 \times |V| \)

### Experiments and Word Error Rate (WER) Results

**Setup**

- TED-LIUM Release 2 (TLv2)
- 300 Switchboard (SWBD): Hub5’00 (dev) and Hub5’01 (test)
- recognition: full-sum decoding with a 4-gram word-level LM

**Label unit & topology**

<table>
<thead>
<tr>
<th>Phonetone Label</th>
<th>TLv2-dev</th>
<th>Hub5’00</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA HMM</td>
<td>RNA HMM</td>
<td></td>
</tr>
<tr>
<td>original</td>
<td>7.6</td>
<td>9.3</td>
</tr>
<tr>
<td>EOW-awarded</td>
<td>6.9</td>
<td>8.8</td>
</tr>
<tr>
<td>+ SOW-awarded</td>
<td>7.3</td>
<td>9.0</td>
</tr>
<tr>
<td>Hybrid HMM</td>
<td>7.4</td>
<td>7.3</td>
</tr>
<tr>
<td>segEnd</td>
<td>6.3</td>
<td>13.4</td>
</tr>
<tr>
<td>CTC</td>
<td>7.2</td>
<td>13.4</td>
</tr>
</tbody>
</table>

**Alignment & Label position**

- \( \nu_l \): positions in \( y_j^u \) where \( s_j^v \) occurs
- stable training procedure: various alignment properties

### Conclusion

A simple and competitive phoneme-based neural transducer approach

- advantages of both classical and end-to-end approaches
- utilize local dependency of phonemes: simplified NN with small footprint and straightforward LM integration
- stable and efficient training using frame-wise CE loss
- RNA topology: better than HMM topology for transducer modeling
- EOW-awarded phonemes: consistent improvement
- phonetic context size of one + chunk-wise Viterbi training: best performance

### References

- [Sak 2015](#) Hailin Sak et al., “Recurrent Neural Aligner: An Encoder-Decoder Neural Network Model for Sequence-to-Sequence Mapping”, Interspeech 2017
- [Tüske 2017](#) Zoltan Tümke et al., “Attention based Sequence-to-Sequence Model for State-of-the-Art Results on Switchboard”, Interspeech 2017
- [Variani 2020](#) Ehsan Variani et al., “Hybrid Autoregressive Transducer”, CSTR 2020
- [Zhou 2020](#) Wei Zhou, Simon Berger, Ralf Schlüter, Hermann Ney

### Further WER Results

<table>
<thead>
<tr>
<th>Work</th>
<th>Epoch</th>
<th>Modeling</th>
<th>Label</th>
<th>LM</th>
<th>TLV2-dev (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="#">Karita 2019</a></td>
<td>100</td>
<td>Attention</td>
<td>Label</td>
<td>RNN</td>
<td>9.3</td>
</tr>
<tr>
<td><a href="#">Han 2017</a></td>
<td>50</td>
<td>Transducer</td>
<td>Label</td>
<td>RNN</td>
<td>9.4</td>
</tr>
<tr>
<td><a href="#">Zhou 2020</a></td>
<td>50</td>
<td>Transducer</td>
<td>Label</td>
<td>RNN</td>
<td>9.4</td>
</tr>
<tr>
<td><a href="#">Variani 2020</a></td>
<td>50</td>
<td>Transducer</td>
<td>Label</td>
<td>RNN</td>
<td>9.4</td>
</tr>
</tbody>
</table>

### Acknowledgements

This work has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No 645952, project 'SEQCLAS') and from a Simple Focus Award. The work reflects only the authors’ views and none of the funding parties is responsible for any use that may be made of the information it contains.