End-to-End Speech Recognition with Local Monotonic Attention
Andros Tjandra, Sakriani Sakti, Satoshi Nakamura
Nara Institute of Science and Technology (NAIST), Japan

**Motivation**

- Most attentional mechanism in encoder-decoder neural network has a “global” property
  → Attend whole input sequence and calculate the expected context

- However, it does not fit with monotonous nature in ASR

**We need:**

Attention Mechanism with Local & Monotonicity Properties

- Monotonicity: current expected position ≥ than previous expected position
- Locality: attention score outside certain distance will be ignored

**Designing Local & Monotonic Attention**

1. Monotonicity-based Prediction of Central Position
   - Predict \( \Delta p_t, \lambda_t \) with an MLP(\( h^d_t \))
   - New position \( p_t = p_{t-1} + \Delta p_t \)
   - Generate Gaussian attention \( a^N_t(s) = \lambda_t \exp\left(-\frac{(s-p_t)^2}{2\sigma^2}\right) \)
   Two different formulations for \( \Delta p_t \):
   a. Constrained
   \[
   \Delta p_t = c_{max} \cdot f\left(\text{MLP}(h^d_t)\right)
   \]
   b. Unconstrained
   \[
   \Delta p_t = f\left(\text{MLP}(h^d_t)\right)
   \]
   \( f: \mathbb{R} \rightarrow [0,1] \) (e.g., sigmoid)
   \( c_{max} \) is hyperparameter (max. jump range)

2. Locality-based Alignment Generation
   - Select a subset of encoder states \( h^s_t \) where \( \forall s \in [p_t - 2\sigma, p_t + 2\sigma] \)
   - Calculate the scorer attention \( a^s_t = \text{Align}(h^s_t, h^d_t) \) only based on a subset of \( h^s_t \)
   \[
   \text{Align}(h^s_t, h^d_t) = \left(h^d_t W h^s_t + \text{bilinear}\left(W_1 \sigma(W_2[h^s_t, h^d_t])\right)\right) \text{MLP}
   \]
   Advantage: reduce the complexity from \( O(T \ast S) \rightarrow O(T \ast \sigma) \)

3. Context Calculation
   Combine attention \( a^N_t, a^s_t \) and calculate weighted sum \( c_t \)
   \[
   c_t = \sum_{s=(p_t-2\sigma)}^{(p_t+2\sigma)} (a^N_t(s) \ast a^s_t(s)) \ast h^s_t
   \]

**Experimental Results**

<table>
<thead>
<tr>
<th>Setup:</th>
<th>Model</th>
<th>Test PER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Att Enc-Dec Global MLP Scorer</td>
<td>23.8</td>
<td></td>
</tr>
<tr>
<td>Att Enc-Dec local-m (Loung et al., 2015)</td>
<td>(not converged)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Proposed</th>
<th>Monotonicity</th>
<th>Locality (Alignment)</th>
<th>Scorer</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Constrained (sigmoid)</strong></td>
<td>No</td>
<td>-</td>
<td>23.2</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>Bilinear</td>
<td>21.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MLP</td>
<td>21.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unconstrained (exp)</strong></td>
<td>No</td>
<td>-</td>
<td>23.1</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>Bilinear</td>
<td>20.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MLP</td>
<td>21.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conclusion**

- Demonstrated a novel attention mechanism that ensures monotonicity & locality properties
- Explained various ways to control those properties
- Experimental results showed that unconstrained position prediction + local alignment produced best result
- Can be applied to other tasks: G2P & MT on certain language pairs

Link to full paper: [https://goo.gl/GGqYiT](https://goo.gl/GGqYiT)
Contact: andros.tjandra.ai6@is.naist.jp