**SEARNN: Training RNNs with Global-Local Losses**

Rémi Leblond*, Jean-Baptiste Alayrac*, Anton Osokin Simon Lacoste Julien

INRIA ENS Paris France

MILA DIRO UdeM

* equal contribution

**Learning to Search**

**Structured prediction**
Learn a mapping \( f \) between inputs \( X \) and structured outputs \( Y \) made of interrelated parts often subject to constraints

**Learning To Search (L2S)**
Reduces the structured problem down to cost sensitive classification with theoretical guarantees

**How does it work?**
A unique shared classifier makes predictions on one by one conditioned on the input and the previous tokens. This classifier is trained on an intermediate dataset

**Links between Learning to Search and RNNs**
- Decomposition of structured tasks in sequential predictions on the past conditional
- Unique shared classifier for all decisions using predecessors

**SEARNN [3]**

**Overview:**
Integrate roll-outs in the decoder to compute the cost of ever possible action at each step

**Leverage** these costs to enable better training losses

\[
\begin{align*}
& h_t = f(h_{t-1}, y_{t-1}) \\
& s_t = \text{proj}(h_t) \\
& o_t = \text{softmax}(s_t)
\end{align*}
\]

**Probabilistic interpretation:**

\[
\begin{align*}
& o_t = P(Y_t | X, Y_1, \ldots, Y_{t-1}) \\
& \prod_{t=1}^{T} o_t = P(Y_1, \ldots, Y_T | X)
\end{align*}
\]

**MLE:**

\[
\begin{align*}
& \text{max}_y \sum_{t=1}^{T} \log(P_y(Y | Y_X | X))
\end{align*}
\]

**Known problems:**
- **test loss**
  - edit distance (bleu score)
- **0/1 flavour**
  - Probability mass aggregated on one sample bad for structured prediction
- **exposure bias**
  - Training scenario is different from what happens at test time

**Experiments**

**Full algorithm:**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( A )</th>
<th>( T )</th>
<th>Cost</th>
<th>MLE</th>
<th>\text{mix}</th>
<th>Learned</th>
<th>\text{ref}</th>
<th>Learned</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCR</td>
<td>26</td>
<td>15</td>
<td></td>
<td></td>
<td>2.8</td>
<td>1.9</td>
<td>2.4</td>
<td>1.9</td>
</tr>
<tr>
<td>CoNLL</td>
<td>22</td>
<td>70</td>
<td>4.2</td>
<td>2.7</td>
<td>1.9</td>
<td>1.9</td>
<td>2.0</td>
<td>2.4</td>
</tr>
<tr>
<td>Under</td>
<td>0.5</td>
<td>43</td>
<td></td>
<td></td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Spelling</td>
<td>0.5</td>
<td>43</td>
<td></td>
<td></td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>

**Scaling approach:**

- Significant improvements over MLE on all 4 tasks
- The harder the task the bigger the improvement
- Learned/mixed is the best strategy for roll in out
- The best performing losses for most are those structurally close to MLE
- SEARNN does not require warm start
- The proposed sampling strategy works maintaining improvements at a fraction of the cost

**Machine Translation (in progress): ISWT 14 Ger/Eng**

<table>
<thead>
<tr>
<th>Depth</th>
<th>MLE</th>
<th>MIXER</th>
<th>SEA-RNN ref mix</th>
<th>SEA-RNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**References**