Fast decoding in neural machine translation with Ray

MAREK STRELEC
<s> For them it was a fascinating game
Time cost statistics for decoding

<table>
<thead>
<tr>
<th>Calculation Units</th>
<th>GPU</th>
<th></th>
<th>CPU</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time(s)</td>
<td>Percentage</td>
<td>Time(s)</td>
<td>Percentage</td>
</tr>
<tr>
<td>Eq. (6): ( s_j = f(e^*<em>{y</em>{j-1}}, s_{j-1}, c_j) )</td>
<td>551.07</td>
<td>75.73%</td>
<td>1370.92</td>
<td>19.42%</td>
</tr>
<tr>
<td>Eq. (7): ( t_j = g(e^*<em>{y</em>{j-1}}, c_j, s_j) )</td>
<td>88.25</td>
<td>12.13%</td>
<td>277.76</td>
<td>3.93%</td>
</tr>
<tr>
<td>Eq. (8): ( o_j = W o t_j )</td>
<td>25.33</td>
<td>3.48%</td>
<td>2342.53</td>
<td>33.18%</td>
</tr>
<tr>
<td>Eq. (9): ( D_j = \text{softmax}(o_j) )</td>
<td>63.00</td>
<td>8.66%</td>
<td>3069.25</td>
<td>43.47%</td>
</tr>
</tbody>
</table>

Ray

- “A flexible, high-performance distributed execution framework”
- Implements a dynamic task graph computation model
- Global Control Store
- Bottom-up distributed scheduler
- Actor abstraction
Steps

- Implement an NMT model in TensorFlow
- Train the model on a subset of parallel data (Europarl)

Experiments
- Distributed batched translation
- Distributed Beam Search
- Dynamic Beam Search
- Heterogeneous environment

- Compare times and BLEU score
Thank you!