Adaptive Multi-pass Decoder for Neural Machine Translation

EMNLP 2018

http://aclweb.org/anthology/D18-1048
Neural Machine Translation (NMT)

- The encoder-decoder are widely used in neural machine translation
  - the encoder transforms the source sentence into continuous vectors
  - the decoder generates the target sentence according to the vectors
  - the alternatives of the encoder/decoder can be RNN/CNN/SAN
Motivation

• Traditional attention-based NMT adopts one-pass decoding to generate the target sentence

• Recently, the polishing mechanism-based approaches demonstrate their effectiveness
  – these approaches first create a complete draft using the conventional models
  – and then polish this draft based on the global understanding of the whole draft

• Divided into two categories
  – **post-editing** - > a source sentence $e$ is first translated to $f$, and then $f$ is refined by another model
  – with respect to post-editing, the generating and refining are two separate processes

  – **end-to-end approaches** - > most relevant to our work
Related Work

- Deliberation Networks (Xia et al. NIPS 2017)
  - consist of two decoders: a first-pass decoder generates a draft, which is taken as input of second-pass decoder to obtain a better translation
  - The second-pass decoder has the potential to generate a better sequence by looking into future words in the raw sentence

- ABDNMT (Zhang et al. AAAI 2018)
  - adopt a backward decoder to capture the right-to-left target-side contexts
  - assist the second-pass forward decoder to obtain a better translation

- the idea of multi-pass decoding is not well explored
Adaptive Multi-pass Decoder

- Consist of three components -> encoder, multi-pass decoder and policy network
  - multi-pass decoder -> polish the generated translation with decoding over and over
  - policy network -> choose the appropriate decoding depth (the number of decoding passes)
Multi-pass Decoder

• Similar to the conventional decoder, the multi-pass decoder leverages an attention model to capture the source context from the source sentence.
• Towards considering the context information from generated translation, another attention model is utilized to achieve this target.
• The attended hidden states are derived from the inference using the previous decoder.
Policy Network

- The policy network determines to continue decoding or halt -> two actions
- The hidden states of policy network are computed with RNN to model the difference between the consecutive decoding
- We use attention model to capture useful information and take the output as input of RNN
- We use REINFORCE algorithm to train the policy network, and take BLEU as the reward

\[
\begin{align*}
    s_l^{\text{policy}} &= f^{\text{policy}}(s_{l-1}^{\text{policy}}, m_l) \\
    \pi(a_l|s_l^{\text{policy}}; \theta_p) &= \text{softmax}(W_p s_l^{\text{policy}} + b_p) \\
    J^{\text{policy}}(\theta_p) &= \mathbb{E}_{\pi(a_l|s_l^{\text{policy}}; \theta_p)} r(\hat{y}_L(x,y)) \\
    r(\hat{y}_L(x,y)) &= \text{BLEU}(\hat{y}_L(x,y), y)
\end{align*}
\]
Experiments

• Chinese-English translation task
  – 1.25M sentence pairs from LDC corpora
  – use NIST02 as development dataset and NIST03, NIST04, NIST05, NIST06 and NIST08 as testing datasets
  – take BLEU as evaluation metric

<table>
<thead>
<tr>
<th>System</th>
<th>#Params</th>
<th>Speed</th>
<th>MT02</th>
<th>MT03</th>
<th>MT04</th>
<th>MT05</th>
<th>MT06</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses</td>
<td>–</td>
<td>–</td>
<td>33.79</td>
<td>30.86</td>
<td>32.71</td>
<td>30.02</td>
<td>30.49</td>
<td>31.02</td>
</tr>
<tr>
<td>RNNSearch</td>
<td>83.99M</td>
<td>87</td>
<td>39.68</td>
<td>36.51</td>
<td>40.20</td>
<td>36.87</td>
<td>36.43</td>
<td>37.50</td>
</tr>
<tr>
<td>Deliberation Network</td>
<td>125.16M</td>
<td>162</td>
<td>40.98</td>
<td>37.82</td>
<td>40.56</td>
<td>37.67</td>
<td>37.20</td>
<td>38.31</td>
</tr>
<tr>
<td>ABDNMT</td>
<td>122.86M</td>
<td>132</td>
<td>41.12</td>
<td>38.01</td>
<td>41.20</td>
<td>38.07</td>
<td>37.59</td>
<td>38.71</td>
</tr>
<tr>
<td>2-pass decoder</td>
<td>87.81M</td>
<td>–</td>
<td>41.18</td>
<td>37.76</td>
<td>41.06</td>
<td>38.02</td>
<td>37.41</td>
<td>38.56</td>
</tr>
<tr>
<td>3-pass decoder</td>
<td>87.81M</td>
<td>245</td>
<td>41.28</td>
<td>37.99</td>
<td>40.72</td>
<td>37.86</td>
<td>37.63</td>
<td>38.55</td>
</tr>
<tr>
<td>4-pass decoder</td>
<td>87.81M</td>
<td>293</td>
<td>41.05</td>
<td>37.86</td>
<td>40.87</td>
<td>38.18</td>
<td>37.57</td>
<td>38.62</td>
</tr>
<tr>
<td>5-pass decoder</td>
<td>87.81M</td>
<td>322</td>
<td>40.88</td>
<td>37.70</td>
<td>40.84</td>
<td>38.06</td>
<td>37.97</td>
<td>38.64</td>
</tr>
<tr>
<td>Adaptive multi-pass decoder</td>
<td>96.01M</td>
<td>180</td>
<td><strong>41.42</strong></td>
<td><strong>38.39</strong></td>
<td><strong>41.43</strong></td>
<td><strong>38.54</strong></td>
<td><strong>37.86</strong></td>
<td><strong>39.05</strong></td>
</tr>
</tbody>
</table>

• The average decoding depth is 2.12
## Case Study

<table>
<thead>
<tr>
<th>Source</th>
<th>xinhua news agency, beijing, april 5, bill gates, the all-famous microsoft chairman, was duped by a canadian radio station the other day and fell a victim to a big prank on april fools’ day.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>xinhua news agency, beijing, april 5, a fews days ago, microsoft ’s president, microsoft corporation, was ” UNK ” by a radio station in canada.</td>
</tr>
<tr>
<td>1st-pass</td>
<td>xihua news agency report of april 5th from beijing (by staff reporter UNK UNK) - the president of microsoft ’s microsoft corporation, gates, was recently ” UNK ” by a radio station in canada and was hit by a UNK day on the day of the day.</td>
</tr>
<tr>
<td>2nd-pass</td>
<td>xinhua news agency, beijing, april 5, microsoft ’s president bill gates, the president of microsoft, was ” UNK ” by a radio station in canada in few days ago.</td>
</tr>
<tr>
<td>3rd-pass</td>
<td>xinhua news agency, beijing, april 5, microsoft ’s president bill gates, the president of microsoft, was ” UNK ” by a radio station in canada in few days ago.</td>
</tr>
<tr>
<td>4th-pass</td>
<td>xinhua news agency, beijing, april 5, microsoft ’s president bill gates, the president of microsoft, was ” UNK ” by a radio station in canada in few days ago.</td>
</tr>
<tr>
<td>5th-pass</td>
<td>xinhua news agency, beijing, april 5, microsoft ’s president bill gates, the president of microsoft, was ” UNK ” by a radio station in canada in few days ago.</td>
</tr>
</tbody>
</table>
Conclusion

- We first explore to generate the translation with fixed decoding depth
- Further we leverage policy network to determines continuing decoding or halt and train this network using reinforcement learning
- We demonstrate its effectiveness on Chinese-English translation task
Thanks & QA