On the Relation between Position Information and Sentence Length in Neural Machine Translation

Masato Neishi
The University of Tokyo
neishi@tkl.i.u-tokyo.ac.jp

Naoki Yoshinaga
Institute of Industrial Science, the University of Tokyo
ynaga@iis.u-tokyo.ac.jp

Code available: https://github.com/nem6ishi/conll19_relative_transformer

Summary

Problem: NMT has difficulty in translating long sentences.
Hypothesis: Word position encoding significantly affects the performance.
Relative (ex. RNN) vs. Absolute (ex. Positional Encodings)
Conclusion: Relative position is better and prevents overfitting to the sentence length.

1. Background

- Long sentence: A major problem in NMT
  - Attention mechanism helps RNN-based NMT model to mitigate this problem. (Bahdanau, 2015; Luong, 2015)
  - RNN-based NMT < Phrase-based SMT in translating very long (>80) sentences. (Koehn and Knowles, 2017)

- Does Transformer (Vaswani 17), NMT model superior to RNN-based one, work well for long sentences? - No, it is worse. (cf. §5)

2. Preliminary: Type of position information

Transformer and RNN-based NMT differ in position information to handle variable-length input.

- Relative position
  - Ex. Encoder of RNN-based model
  - No explicit position representations to learn.

- Absolute position
  - Ex. Encoder of Transformer
  - Need to learn to process the position vector.
  - Less chance to learn large positions.

Hypothesis

The type of position information significantly affects the translation of long sentences.

3. Approach: Transformer with Relative Position

- Compare the types of position information using Transformer. Position information customizable!

- [Shaw+, 2018]: Self-attention with relative position
  - Introduce relative position vectors into self-attention process (and remove positional encoding layer).
  - Need to learn the process the position vector, but more chance to learn large positions.

- [The modified self-attention process]
  \[ z_i = \sum_{j=1}^{d} a_{ij}(y_j^V - y_i^V), \quad a_{ij} = \exp c_{ij}, \quad c_{ij} = x_i W^Q (x_j W^K + w^K_j)^T \]

- Proposal: RNN as a Relative Positional Encoder
  - Replace positional encoding layers by RNN.
  - [Original] \( w_0^i = w_i + \text{PositionalEncoding}(i) \)
  - [Proposed] \( w_0^i = h_i = \text{GRU}(w_i, h_{i-1}) \)

- Variants of Transformer for comparison

4. Experimental Settings

- Models and their types of position information:
  - RNN-NMT [Luong+, 2015], (Relative)
  - Transformer (Absolute) and its three variants (Relative)
    - The number of parameters set to be almost equal.

- Datasets (preprocessed):
  - WMT2014 English-to-German (3.7M sentences)
  - ASPEC English-to-Japanese (1.2M sentences)
    - Sentences longer than 49 tokens are filtered out.

5. Result & Analysis

- BLEU score [Papineni, 2002]

<table>
<thead>
<tr>
<th></th>
<th>WMT2014 En-De</th>
<th>ASPEC En-Ja</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN-NMT</td>
<td>19.95</td>
<td>36.67</td>
</tr>
<tr>
<td>Transformer</td>
<td>21.00</td>
<td>39.84</td>
</tr>
<tr>
<td>Rel-Transformer</td>
<td>22.51</td>
<td>39.58</td>
</tr>
<tr>
<td>RNN-Transformer</td>
<td>23.35</td>
<td>39.17</td>
</tr>
<tr>
<td>RR-Transformer</td>
<td>23.01</td>
<td>40.34</td>
</tr>
</tbody>
</table>

- Among Transformers, Relative beats Absolute.
- RR-Transformer performs the best.

[Analysis on WMT2014] (See our paper on ASPEC)

- Evaluation on test data split by input length

  - Transformer fails to translate long sentences, and overfits to short input sentences in the training data.
  - Relative position avoids this overfitting.

- Does Transformer overfit to short input sentences only?

- Results when trained on length-controlled data
  - Input sentence length: 34-49

  - Transformer overfits to the lengths of input sentences in the training data.

6. Conclusion

- Relative position shows better translation quality while Absolute position causes overfitting.

**TAKE AWAY:** Use Relative position in NMT.