Self-Attentive Residual Decoder for Neural Machine Translation

Lesly Miculicich†, Nikolaos Pappas*, Dhananjay Ram†, Andrei Popescu-Belis‡

*Idiap Research Institute, †EPFL, ‡HEIG-VD / HES-SO

Motivation

Limitations of the RNN-based decoder for NMT
- The RNN’s internal memory is shared across words and is prone to a recency bias.
- Does not fully capture the structure of language.

Proposed approach:
- Enhance the RNN memory with direct and selective access to past.
- The residual connections facilitate the flow of information.
- The self-attention allows selective use of previously predicted words.

Other Self-Attentive Networks

Memory RNN
RNN with memory cells of previous representations
[Cheng et al., EMNLP 2016]

Self-Attentive RNN
RNN with a summary vector from past predictions
[Danilk et al., ICLR 2016]

Self-attention Matrices:

Mean Residual

Baseline NMT decoder
p(y_t | y_1, ..., y_{t-1}, c_t) \approx g(h_t, c_t, y_{t-1})

Self-attentive residual decoder
p(y_t | y_1, ..., y_{t-1}, c_t) \approx g(h_t, c_t, y_{t-1})

h_t = f(h_{t-1}, y_{t-1})

Self-attentive residual connections
\alpha_t^i = \text{attention}(h_{t-1}, h_t)

Experimental Setup

Datasets: En-ZH UN Corpus 0.5M, Es-En WMT 2.1M, En-De WMT 4.5M
Architecture: Attention-based NMT with GRUs of dimension 1024, 500 for word embeddings, and vocabulary of 50K.

Results

<table>
<thead>
<tr>
<th>Models</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT baseline</td>
<td>21.6</td>
</tr>
<tr>
<td>NMT transformer (comparable model)</td>
<td>109.0M</td>
</tr>
<tr>
<td>NMT baseline</td>
<td>108.7M</td>
</tr>
<tr>
<td>Memory RNN</td>
<td>106.7M</td>
</tr>
<tr>
<td>Self-attentive RNN</td>
<td>110.2M</td>
</tr>
<tr>
<td>Mean residual connections</td>
<td>108.7M</td>
</tr>
<tr>
<td>Self-attentive residual connections</td>
<td>108.9M</td>
</tr>
</tbody>
</table>

BLEU on tokenized text. | is the number of parameters.

- Self-attentive residual connections outperform other models, while using fewer parameters than other self-attentive methods.

Code at: https://github.com/idiap/Attentive_Residual_Connections_NMT

Conclusion

- We proposed self-attentive residual learning framework.
- Improvements over a standard baseline, and two variants of self-attention.
- Analysis of the attention shows syntactic-like structures.
- It can be applied to other tasks based on RNNs.

Acknowledgements

Supported by the European Union Horizon 2020 SUMMA project (grant 688139.

www.summa-project.eu)