How Much Attention Do You Need?

A Granular Analysis of Neural Machine Translation Architectures

Steven Walton
University of Oregon

21 Feb 2019

Overview

Questions:

- If attention is all you need, then how much?
- Where is the attention important?
- What type of attention do we need? Self? LSTM? Transformers?

Overview

Questions:

- If attention is all you need, then how much?
- Where is the attention important?
- What type of attention do we need? Self? LSTM? Transformers?

Answers:

- Source attention on lower encoder layers brings no additional benefits.
- Multiple source attention and residual feed-forward layers are key.
- Self-attention is more important for the source than for the target side.

- *Flexible Neural Machine Translation Architecture Combination*
  - *Neural Machine Translation (NMT)*
  - *Architecture Definition Language (ADL)*
  - *Layer Definitions*
  - *Standard Architectures*
- Related Work
- Experiments
- Conclusion

Neural Machine Translation (NMT)

- NMT is a sequence to sequence prediction task

$$X \mapsto Y$$

$$p(y_t | Y_{1:t-1}, X; \theta) = \text{softmax}(\boldsymbol{W}_o \boldsymbol{z}^L + \boldsymbol{b}_o)$$

- $\boldsymbol{W}_o$ projects a model dependent hidden vector $\boldsymbol{z}^L$ of the L^{th} decoder layer to the dimension of the target vocabulary $\boldsymbol{V}_{trg}$
- Training minimizes cross-entropy loss

Architecture Definition Language (ADL)

- Flexible Neural Machine Translation Architecture Combination
- *Related Work*
- Experiments
- Conclusion

- Flexible Neural Machine Translation Architecture Combination
- Related Work
- *Experiments*
- Conclusion

- Flexible Neural Machine Translation Architecture Combination
- Related Work
- Experiments
- *Conclusion*