Aeden Jameson
9810912
EE 596 Autumn 2021
Homework 5

(1) Motivation – Why do the authors want to work on this problem

The authors want to advance the state of the art in deep learning techniques in sequence modeling and transduction problems such as language modeling and machine translation[1].

(2) Contributions – What are the accomplishments they achieved in this paper (others did not achieve)?

Through use of transformers the authors were able to advance the state of the art by significantly improving parallelization, beating results on standard WMT data sets by about 1 BLEU point and doing so by training for as little as twelve hours on eight P100 GPUs.

(3) Formulations – How do they solve the problems as mentioned/discussed in the introduction or related literature?

The authors achieved their contributions by eschewing traditional RNN architectures. And constructing an overall model architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder[1]; combined with temporal embeddings, layer norm. Lastly a variant of dot-product attention with multiple heads is then used.

(4) Justification – Do the experiments/simulations support their claimed accomplishments?

Support for the claim of reduced training time is well supported by the complexity analysis in Table 1. Additionally there experimental setup looks replicable and so I don't see a sufficient reason to doubt the model performance numbers they present in Table 2.

(5) Your Own Thoughts – What are you most impressed with in this paper?

What I found most impressive overall was the authors achieving state-of-the-art results by not trying more complicated RNN architectures and or tricks with them, but by taking what existed and using it more effectively. Thus boiling it all down to attention is all you need.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems* 30 (NIPS 2017)