PAY LESS ATTENTION WITH LIGHTWEIGHT AND DYNAMIC CONVOLUTIONS

Abraham Jose
ID :5068109, CAP6614
abraham@knights.ucf.edu

1 Summary

The author was able to bring a very lightweight convolution technique to replace the self-attention models. They introduces convolution kernels, Dynamic convolution and Lightweight convolution for getting the attention with a fixed context window. The number of operations required by this approach scales linearly in the input length which is less than the self-attention models which runs in quadratic operations.

2 Strengths of the proposal

1. The number of operations required for the convolution kernels implemented is linear to the input length, whereas the self-attention model requires quadratic operations to the input length.
2. Operates on the context window efficiently and effectively in the language model tasks machine translation, language modeling and abstractive summarization. They outperforms the best strong self-attention models.
3. They were able to prove that the self-attention can be achieved without incorporating the self-attention module.

3 Weaknesses of the proposal

1. The architecture of the model is similar to the Transformer Big of Vaswani et al., except that self-attention modules are swapped with either fixed or dynamic convolutions. From the results they are not able to explain the variation of the output in both case except for the cost of operation. Such inferences about trade offs are missing in the experiments.
2. The definition and implementation of Dynamic convolution is not clearly mentioned.

4 Results

The author tested the model across various language model task including machine translation, language modeling and abstractive summarization on respective large-scale data-set such as WMT En-De, En-Fr IWSLT, CNN-DailyMail etc.. They were able to achieve state of the art result in the WMT'14 English-German test set for machine translation. Also they were able to prove that the self attention models are not necessary for building generative models for language, which can be replaced by dynamic convolutions or lightweight convolution.

5 Discussion

How can we extent this work using the dynamic convolution to create attention mechanism for the images? How we can sequence an image effectively for context based information?

Why does self-attention models with quadratic complexity has similar performance with a network that has linear operation complexity while using convolutions? The *"Attention Is All You Need"* from Vaswani et al. is rooted in the importance of attention which is the state of the art language model called Transformers.

CAP 6614: Current Topics in Machine Learning (Fall 2019), UCF.