Attention Is All You Need

Transduction problems, which include language modelling and machine translation, have historically been implemented through highly complex recurrent or convolutional neural networks - using an encoder and a decoder. This paper presents the idea of the Transformer, a simple network architecture which is based purely on attention processes, avoiding any needs for recurrence or convolutions. The Transformer model is autoregressive, which means that each output is dependent on both previous outputs and previous inputs. Sequential computation is constrained in these types of recurrence networks - having limitations when dealing with extended sequences and being slow to train.

The motivation behind using attention mechanisms is given by three criteria:
The computational complexity per self-attention layer, the number of computations that can be parallelized, and the path length between long range dependencies in the network.

An additional benefit of attention-based models is that they are easier to interpret than those used in the past. The Transformer model allows for significantly higher parallelization capabilities and can achieve a new state of the art in translation quality.

Results obtained from the 2014 WMT (Workshop on Machine Translation) English-German and English-French translation tasks showed that the transformer model outperforms others when comparing the BLEU score and total training time. The BLEU (BiLingual Evaluation Understudy) is a metric for automatically evaluating machine-translated text. The transformer model achieved 28.4 BLEU, an improvement of nearly 2 BLEU over the previous best results, including ensemble methods.

Significance

This paper presents a revolutionary new addition to the world of transduction, which should help things like machine translation models to perform better and train faster than before.

Limitations

The researchers intend to apply the new Transformer model to challenges requiring input and output modalities other than text, as well as to examine restricted attention techniques for efficiently handling huge inputs and outputs including pictures, audio, and video. The lack of these implementations and features are the only real limitations of this new architecture.

A Question

How could these transformer models be improved upon even more?