VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION

The Transformer is a simple network architecture based solely on attention mechanisms that is superior in quality and parallelizable, requiring significantly less time to train.

Recurrent neural networks, long short-term memory, and gated recurrent neural networks have been established as state-of-the-art approaches in sequence modelling and equal contribution.

The Transformer stands out as it's the first-ever transduction model which relies solely on self-attention. This revolutionary model can compute representations of its input and output without sequence-aligned RNNs or convolution, thereby saving time & resources.

Transformer utilizes an encoder-decoder structure combined with cascading self-attention and point-wise, densely connected layers.

Encoder and decoder employ residual connections, layer normalization, and masking to ensure predictions for position I can only depend on known outputs.

Attention functions map query and key-value pairs to an output, which is weighted by a compatibility function.

Additive attention outperforms dot-product attention for small values of dk, while dot-product attention outperforms additive attention for larger values.

Multi-head attention lets models attend to different subspaces of data simultaneously, allowing for efficient computation. This greatly reduces the computing cost associated with traditional models.

The Transformer uses multi-head attention in three ways: encoder-decoder attention, self-attention layers, and scaling dot-product attention. We use learned embedding and softback functions to convert input and output tokens to vectors of dimension d model. Positional encoding is needed to make use of the order of the sequence. This work uses sine and cosine functions of different frequencies to add positional encodings to input embedding at the bottoms of the encoder and decoder stacks, allowing the model to learn to attend by relative positions.

When the sequence length is shorter than the representation dimensionality, self-attention layers are more efficient than recurrent layers. This makes them time-saving and a practical choice when dealing with short sequences. We trained our models on a single machine with O(n/k) convolutional layers, with separable convolutions decreasing complexity to O(k · n · d + n · d2).

The Transformer (big) model stands head and shoulders above its competitors in terms of performance and requires less training than any of them. This opens up a world of opportunities to use it for various applications that can benefit from its efficiency. We measured the change in performance on English-to-German translation on the Transformer architecture by varying the number of attention heads and the attention key and value dimensions, keeping the amount of computation constant.

The Transformer is the first sequence transduction model based entirely on attention, outperforming recurrent or convolutional layers for translation tasks. We are grateful to Nal Kalchbrenner and Stephan Gouws for their support.