**Transformers: Attention is all you need**

**Introduction:**

RNN-based models, specifically the Encoder-Decoder architecture, are widely used for sequence modeling tasks like language modeling and machine translation.

However, these models are strictly sequential and do not support parallelism with longer sequences. Advancements in computation efficiency include factorization and conditional computing. Sequence modeling, which handles sequential data with dependencies of any distance, requires attention mechanisms. The authors introduce the Transformer, a novel model structure that relies just on attention and achieves much enhanced parallelism.

The Transformer, for example, achieved the new state-of-the-art translation quality after only twelve hours of training on eight P100 GPUs.

**Background:**

Similar to the EN-GPU, Byte Net, and ConvS2S, the Transformer attempts to reduce sequential processing. Unlike these models, which require more operations to connect two distant points, the Transformer uses a fixed number of operations.   
• The model's core operation is self-attention, which connects locations within a series. • Self-attention is successful in NLP tasks like reading comprehension and abstractive summarization. The Transformer is characterized as the first transduction model that does not employ RNNs or convolution for both input and output.

**Model architecture:**

The Transformer also has an encoder-decoder architecture, with the encoder consisting of numerous self-attention and point-wise, completely linked layers.   
• Encoder: Made up of N = 6 copies of the same layer, which contains the Multi-head self-attention existent layer and the Position-wise completely linked feed-forward network layer.   
• Decoder: Includes the same N = 6 layers as the preceding kind, but with one additional layer for masked multi-head self-attention.   
• All sublayers in the encoder and decoder blocks employ residual connection, as well as layer normalization. Every layer in the encoder and decoder is a fully linked feed forward network that applies to each point independently and identically. It uses learnt embeddings to convert the input token vector and output tokens vector to the d\_model dimensions as well. Uses a linear transform and the SoftMax function to convert the decoder output to the probability distribution of the following token. The inputs are processed via lookup tables to get embeddings, and position embeddings are used in the model to provide information about the token's location in the sequence, either relative or absolute.

**Why Self-Attention:**

Finally, the authors discuss how the self-attention layers compare to the recurrent and convolutional layers in sequence transduction tasks. There were three primary criteria for comparison:   
a) Multiplying all work loads by layer   
b) The quantity of quantitative calculation that can be completed in parallel.   
b) Long-distance dependencies   
• Self-attention is more computationally efficient when the sequence length and size of each representation are smaller than the representation dimension, which is common when using machine translation models. However, in these extremely long sequences, restricting self-attention may be beneficial in improving the model.

Some convolution layers require at least one layer to be completely linked, while fully connected layers connect all inputs to all output neurons, resulting in a long route length. While sepNet is less expensive than convolutional layers, it may be pricey depending on the application. However, it can reduce the need for recurrent layers. Similarly, self-attention may result in the development of more interpretable models since learned attention heads may be spread across diverse activities in the syntactic and semantic context.

**Training and Result**:

1. The model was evaluated mostly based on the WMT 2014 English to German and English to French translations. The basic model was trained on eight NVIDIA P100 GPUs for each model. Some basic models were trained for 100,000 steps (12 hours), whereas the largest models were trained for 300,000 steps (3 to 5 days).   
2. To vary the sequence of words in the text that arrives to each network, we grouped sentences of approximate lengths into a single batch, yielding batches of 25000 source words and 25000 target words. Adam optimizer was used.   
3. values: Set new efficiency norms by achieving the greatest BLEU values for English-German and English-French translation, with 28. 4 and 41. 8 respectively.