**Transformers: Attention is all you need**

**Introduction:**

Sequence modeling activities like language modeling and machine translation have been shown to benefit greatly from RNN based models, and specifically from the Encoder-Decoder architecture.   
• Because the recurrent models proceed through sequences in a strictly sequential fashion, parallelism with larger sequences is hampered. Improvements in the efficiency of computations have been made possible via conditional computing and factorization techniques.  
• Sequence modeling is a kind of modeling that deals with data in a sequential fashion. Dependencies within input or output sequences can be of any length, thus sequence modeling now requires the use of an attention mechanism. The Transformer, a novel model structure that generates significantly higher parallelism solely by using attention, is presented by the authors.   
During training on just eight P100 GPUs, The Transformer achieved the new translation quality state-of-the-art in just twelve hours.

**Background:**

Like the EN-GPU, Byte Net, and ConvS2S, the Transformer seeks to reduce sequential computation. The Transformer uses a fixed number of operations, in contrast to these methods, which require more operations to link two distant places.

• Effective in several NLP tasks including reading comprehension and abstractive summarization, self-attention is the primary operation that allows the model to connect multiple places within the same series. It is said that the Transformer is the first transduction model of this type that uses convolution and/or RNNs for both the input and the output.

**Model Architecture:**

The encoder of the Transformer is composed of numerous self-attention and point-wise, fully connected layers. The Transformer also employs an encoder-decoder construction.

Position-wise completely connected feed-forward network layer and the Multi-head self-attention existent layer are both included in the encoder, which is made up of N = 6 copies of the same layer.

• Decoder: Contains the same N = 6 layers as the preceding kind, plus an additional layer for masked multi-head self-attention.

• Residual connections and layer normalization are used by every sub-layer in the encoder and decoder blocks. Each layer in the encoder and decoder is a completely connected feed forward network that is applied to each position independently and uniformly from the others.

Both the input tokens vector and the output tokens vector are transformed to d\_model dimensions using learnt embedding techniques. applies the SoftMax function along with a linear transform to map the output of the decoder to the token's probability distribution.

• The location of the token in the sequence, either relative or absolute, is indicated via the incorporation of position embeddings into the model, which are obtained by passing the inputs through lookup tables.

**Why Self-Attention:**

• In the final section, the authors discuss how the self-attention layers are related to the convolutional and recurrent layers in sequence transduction tasks. Three primary criteria were used for the comparison:   
a) The total workload multiplied by each layer   
b) The quantity of parallelizable quantitative computation that is possible.   
b) Distance-dependent long-range relationships   
• Self attention is computationally less expensive when using machine translation models, as is frequently the case when the length of the sequence and the size of each representation are significantly smaller than the representation dimension.

To improve the model, it could be beneficial to limit the self-attention in these incredibly large sequences. • It is required for some convolution layers to have at least one completely connected layer. In fully connected layers, every input neuron is coupled to every output neuron, resulting in a long route length. The recurrent layers are made easier by the usage of SepNet, even though it is less expensive than convolutional layers depending on the application. Similarly, as learnt attention heads may vary in terms of syntactic and semantic context across tasks, self-attention may result in the development of models that are thought to be more interpretable.

**Training and Result**:

1. The WMT 2014 English to German and English to French translation tests served as the primary evaluation tool for the model. For each model, the base model was trained using eight NVIDIA P100 GPUs. Largest base models were trained for 300,000 steps (3, 5 days), whereas some base models were trained for 100,000 steps (12 hours).   
2. We grouped sentences of roughly equal lengths into a single batch, resulting in batches of 25,000 source words and 25,000 target words, in order to diversity the word order that enters each network. Adam optimizer was applied.

3. Findings: Set new benchmarks for effectiveness: achieved the best BLEU scores (28. 4) and (41. 8) for English-German and English-French translation, respectively. Ablation studies on different components were used to thoroughly evaluate the importance of these research.