**Transformers: Attention is all you need**

• The recurrent models are strictly sequential as they move through sequences, hindering parallelism with longer sequences. RNN-based models, and, in particular, the Encoder-Decoder architecture, have been introduced as the state of the art for sequence modeling tasks such as Language Modeling and Machine Translation. Improvements in computational efficiency include conditional computing and factorization techniques. Sequence modeling is a kind of modeling that deals with sequential data; dependencies in the input or output sequences can be of any length, and the attention mechanism has become essential in sequence modeling. The Transformer is a novel model structure that the authors provide. It enables significantly higher parallelism while just requiring attention.   
  
**Background:**

* The Transformer, like the EN-GPU, Byte Net, and ConvS2S, aims to reduce sequential computation. The Transformer uses a fixed number of operations to link two distant places, in contrast to these models that require more actions.
* Several NLP tasks, including text comprehension and abstractive summarization, benefit from self-attention. According to some, the Transformer is the first transduction model of its type without the use of convolution or RNNs for either the input or the output.

**Model architecture:**

The Transformer also has an encoder-decoder architecture, wherein the encoder consists of numerous fully connected layers with self-attention and point-wise structure.

• Encoder: Consists of N = 6 copies of the same layer, which contains the Position-wise completely linked feed-forward network layer and the Multi-head self-attention existing layer.

• Decoder: Added one more layer for masked multi-head self-attention, but otherwise consists of the same N = 6 levels as the preceding kind.

• In the encoder and decoder blocks, layer normalization and residual connection are used by all of the sub-layers. Each layer in the encoder and decoder is a fully connected feed forward network that is applied to each position in a way that is both identical to and independent of the others.

Both the input and output token vectors are transformed to d\_model dimensions using learnt embeddings. uses the SoftMax function in conjunction with a linear transform to translate the output of the decoder to the probability distribution of the subsequent token. • The inputs are run through lookup tables to obtain embeddings, and the model incorporates position embeddings to provide information about the token's relative or absolute location in the sequence.

**Why Self-Attention:**

• Lastly, the authors discuss how the self-attention layers in sequence transduction tasks relate to the recurrent and convolutional layers. Three primary comparison criteria were used:

a) The total workload for each layer multiplied

b) The quantity of quantitative computing that can be completed in parallel.

b) Long-range distance dependencies

• When using machine translation models, self-attention is computationally less expensive when the size of each representation and the sequence length are significantly smaller than the representation dimension. However, restricting the self-attention in these incredibly large sequences could be helpful in improving the model.

Certain convolution layers require that at least one of the layers be fully linked. In fully connected layers, every input neuron is connected to every output neuron, resulting in a long route length. Although it may be more expensive depending on the application, this one is less expensive than convolutional layers; still, the recurrent layers are made easier by the usage of SepNet. In a similar vein, self-attention might result in the development of models that are thought to be easier to understand since learned attention heads may vary according on the syntactic and semantic context of the task.

**Training and Result**:

1. The WMT 2014 English to German and English to French translation tests were used to evaluate the model primarily. For each model, eight NVIDIA P100 GPUs were used to train the basic model. While the largest models were trained for 300,000 steps over three days, other base models were trained for 100,000 steps over twelve hours.   
2. In order to vary the word order that each network receives from the text, we merged sentences of about equal lengths into batches of 25,000 source words and 25,000 target words. The Adam optimizer was applied.   
3. Outcomes: Set new benchmarks for efficiency: achieved the best BLEU scores (28. 4 and 41. 8 respectively) for English-German and English-French translation.