**Transformers: Attention is all you need**

**Introduction:**

RNN based Models and, in particular, the Encoder-Decoder architecture has been introduced as the state of the art for sequence modeling tasks such as Language Modeling and Machine Translation.

• The recurrent models are strictly sequential as they move through sequences therefore hindering Parallelism with longer sequences. There have been advancements in making computations more efficient and they include; factorization tricks and conditional computing.

• Sequence modeling is a type of modeling that handles sequential data, and the dependencies can be of any distance in the input or output sequences, and the attention mechanism has become a necessity in sequence modeling. The authors introduce the Transformer, a new model structure that only uses attention and achieves greatly increased parallelism.

• The Transformer, for instance, obtained the new state-of-the-art of translation quality within twelve hours of training on only eight P100 GPUs.

**Background:**

The Transformer tries to decrease the sequential computation like the EN-GPU, Byte Net and ConvS2S. Contrary to these models, which need more operations in order to relate two distant positions, the Transformer employs a constant number of operations.

• The self-attention is the main operation, which enables the model to connect different positions within the same series.

• Self-attention is effective in several of the NLP tasks such as, reading comprehension and abstractive summarization. The Transformer is described as the first of such transduction model that does not use RNNs and/or convolution for the input and the output.

**Model architecture:**

The Transformer also uses an encoder-decoder structure, in which the encoder is made up of multiple self-attention and point wise, fully connected layers.

• Encoder: Composed of N = 6 copies of the same layer and includes Multi-head self-attention existing layer as well as the Position-wise fully connected feed-forward network layer.

• Decoder: Also includes N = 6 same layers as the previous type, but added one more layer for Masked multi-head self-attention.

• All the sub-layers in encoder and decoder blocks use residual connection, and layer normalization as well. Every layer in both encoder and decoder is a fully connected feed forward network applied to each position independent of the others as well as identically. It applies learned embeddings to transform the input tokens vector and the output tokens vector to d\_model dimensions as well. Employs a linear transform as well as the SoftMax function to map the decoder’s output to the probability distribution of the next token.

• The inputs are passed through lookup tables to get embeddings and position embeddings are incorporated into the model to give information about the position of the token in the sequence relative or in absolute terms.

**Why Self-Attention:**

 Last to it, the authors relate the self-attention layers presented to the recurrent and the convolutional layers in sequence transduction tasks. There were three main criteria for comparison:

a) Multiplication of all work load per layer

b) The amount of computation at the quantitative level that can be done in parallel.

c) Long-range dependencies of distance

• Self attention is cheaper computationally in case when the sequence length and the size of each representation is much less than the representation dimension, which is often the case while applying the machine translation models.

• But in these truly huge sequences it might be useful to restrict the self-attention to enhance the model.

• In some convolution layers, it is necessary that at least one of the layers is connected in a fully connected manner and in fully connected layers, all the inputs are connected to all the output neurons which consequently means great path length. This one is cheaper than convolutional layers but could be costly based on application; nonetheless, by the use of sepNet the recurrent layers are eased.

• Similarly, self-attention may lead to the creation of models that could be considered more interpretable since attention heads that are learned might be across different tasks concerning the syntactic and the semantic context.

**Training and Result**:

1. The model was mainly assessed on the WMT 2014 English to German and translation from English to French. The base model was trained was done on 8 NVIDIA P100 GPUs for each model. Some of base models were trained for 100 000 steps (12 hours), the biggest models for 300 000 steps (3, 5 days).

2. To diversify the order of words from the text that come to each network, we combined sentences with approximate lengths in one batch, getting batches of 25000 source words and 25000 target words. Adam optimizer was used.

3. Results: New standards for efficiency: got the highest given BLEU results for English-German and English-French translation with 28. 4 and 41. 8 correspondingly. These works were assessed for their significance in detail by using ablation studies on various components.