The paper is about new simple network architecture, the Transformer which is based on only attention mechanism without using sequence aligned RNNs or CNNs. As we know, the sequence transduction models are based on complex recurrent or convolutional neural networks including a encoder and a decoder. That is why the authors presented Transformer as the first sequence transduction model based entirely on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention. In this work, they proposed the Transformer, architecture for a model that completely forgoes recurrence in favor of drawing global relationships between input and output. The Transformer can achieve a new level of excellence in translation quality after only twelve hours of training on eight P100 GPUs and offers substantially higher parallelization.

The Transformer follows the overall encoder-decoder architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder. The encoder is composed of a stack of N = 6 identical layers. The decoder is also composed of a stack of N = 6 identical layers. A query, a collection of key-value pairs, and an output, all of which are vectors, can be mapped to one another by an attention function. The result is calculated as a weighted sum of the values, with each value's weight determined by the query's compatibility function with its associated key. In "encoder-decoder attention" layers, the memory keys and values are derived from the encoder's output, while the queries are derived from the preceding decoder layer. As a result, the decoder's locations may all pay attention to every position in the input sequence. Self-attention layers are present in the encoder. All of the keys, values, and queries in a self-attention layer originate from the same source, in this example, the encoder's output from the previous layer. Every place in the encoder may service every position in the layer below it. Similar to this, the decoder's self-attention layers enable each position to pay attention to all positions up to and including it. To maintain the auto-regressive characteristic of the decoder, we must prohibit leftward information flow. By masking away (setting to) all values in the softmax's input that correspond to illicit connections, they implemented this within scaled dot-product attention.

Over 4.5 million phrase pairs from the standard WMT 2014 English-German dataset served as their training data. Using byte-pair encoding, which has a common source target vocabulary of around 37,000 tokens, sentences were encoded. They utilized the much bigger WMT 2014 English-French dataset for English-French, which consists of 36M phrases and divided tokens into a 32000 word-piece vocabulary. Sentence pairs were grouped together based on their approximate length of sequence. A collection of phrase pairings with around 25000 source tokens and 25000 target tokens were included in each training batch.

For the WMT 2014 English-to-German translation challenge, the model scores 28.4 BLEU, outperforming the previous best outcomes, including ensembles, by more than 2 BLEU. The model generates a new single-model state-of-the-art BLEU score of 41.8 on the WMT 2014 English to French translation problem after training for 3.5 days on eight GPUs, a small portion of the training expenses of the best models from the literature. By effectively using the Transformer for English constituency parsing with both big and small amounts of training data, they demonstrate how well it generalizes to different tasks.

**Paper Name:** Attention Is All You Need

**Authors Name:** Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

**Published On:** 6 December, 2017