Department of Computer Science
Databases and Information Systems

Universitätsstr. 1 D–40225 Düsseldorf

Heinrich Heine
University Düsseldorf

Paper Summary

Attention Is All You Need

Zeynep Boztoprak

The paper Attention Is All You Need written by Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin in 2017 presents a sequence transduction model called Transformer which is used in tasks like abstractive summarization or reading comprehension. Unlike previous work the Transformer relies entirely on the attention mechanism without using any reccurence or convolution.

The attention mechanism computes a representation of the whole sequence by relating positions of different distances to each other.
As a first step learned embeddings are used to convert the input and output tokens to vectors to which they add positional encoding. This encoding allows to inject some position informations of the tokens in the sequence.
These are fed to the model which is split into the encoder and decoder:

The encoder has stacked layers each containing first a multi-head attention mechanism and second a positionwise fully connected feed-forward network. The encoder maps the input sequence to a sequence of continuous representations.
The decoder is similar to the encoder except that it has an additional sublayer which performs multi-head attention over the representations of the encoder. Also the self-attention sublayer is modified. At the end softmax is applied to predict next-token probabilities.
It is important to mention that at each step the model is auto-regressive.

Given queries, keys and values the attention mechanism outputs a weighted sum of the values. The weights are computed by a compatibility function of the queries with the corresponding keys.
This kind of attention is called Scaled Dot-Product Attention which is used in the paper instead of the other attention functions because of the speed and memory gains since it can be implemented using highly optimized matrix multiplication.

Furthermore multi-head attention is introduced. They linearly project the queries, keys and values h (number of attention heads) times with different learned linear projections and perform on each of them the attention function in parallel. After that each projection will be concatenated and once again projected to the final values.
This approach allows to jointly attend informations from different representation subspaces at different positions using parallel attention heads. It is observed that each head seems to learn to perform a different task and also that each head is able to capture semantic and syntactic structures of the senteces.

In the Machine Translation task both their best and even their base model outperforms the best previously presented models which led the model to be the state-of-the-art model in translation quality. Here they observed that too many heads lead to quality drop.

Due to the fact that the number of operations to connect all positions is limited to a constant the transformer architecture is faster than previously presented models based on convolution and recurrence.