Transformer

Attention Is All You Need (2017) 62k

RNN: $h_{t-1} \rightarrow h_t$

Issue: sequential computation hard for parallelization, may forget very early information

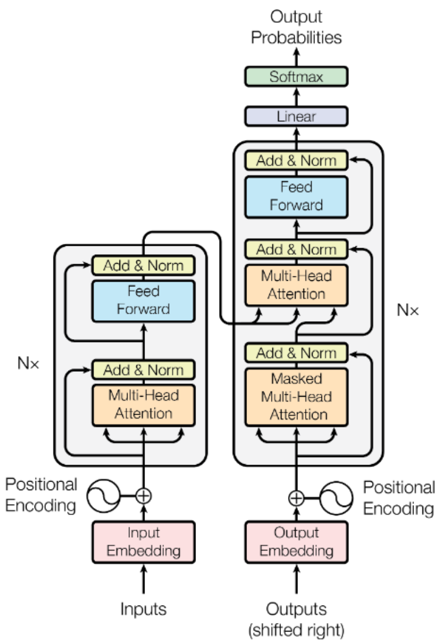

Figure 1: The Transformer - model architecture.

Layernorm

Residual connections

Decoder: masked self-attention, to ensure that the predictions for position i can depend only on the known outputs at positions less than i

Attention: query, keys, values $\rightarrow$ output, weighted sum of values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key

Scaled dot-product attention: queries and keys of dimension d_k

$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$

Multi-head attention:

$\mathrm{Multihead}(Q, K, V) = \mathrm{concat}(\mathrm{head}_1, ..., \mathrm{head}_h)W^O$ where $\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, W are parameter matrices

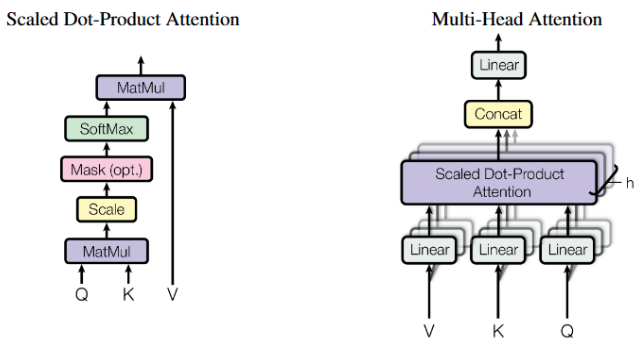

Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.