Scaled Dot-Product Attention

Self-attention is the core mechanism behind Transformer models [1], which have provided state-of-the-art results in various scientific fields (i.e. Natural Language Processing). Self-attention enables models to weigh the significance of different elements (tokens) within a sequence, concerning each other and capturing their dependencies.

Unlike recurrent neural networks (RNNs) or even convolutional neural networks (CNNs) the attention mechanism allows the model to process every element in a sequence simultaneously.

Scaled Dot-Product Attention Explained

Self-attention mechanism utilizes the matrices Q (queries), K (keys), V (values) which derive directly from the input sequence.

Initially, the input sequence of size $\boldsymbol{n}$ passes through an embedding layer of dimension $\boldsymbol{d}$, where each element is converted into a high-dimensional vector representation.

Subsequently, the weight matrices $\boldsymbol{W_Q} \in R^{d \times d_q}$, $\boldsymbol{W_K} \in R^{d \times d_k}$, $\boldsymbol{W_V} \in R^{d \times d_v}$, which are also learnable parameters during training, are utilized to project the embedded elements into the Queries ($Q \in R^{n \times d_q}$), Keys ($K \in R^{n \times d_k}$) and Values ($V \in R^{n \times d_v}$) components. Here, $d_q = d_k$, while d_v may differ but will always depict the output dimension.

The unnormalized attention weights are computed by the dot product between the Q and K matrices. The results are then scaled by the factor $\left(\frac{1}{\sqrt{d_k}}\right)$ in order to ensure that the dot products won't grow too large and consequently avert the vanishing gradient problem.

Following, the normalized attention scores are obtained by applying the Softmax function on the unnormalized attention weights. The scores now depict a probability distribution, assigning higher weights to more relevant elements of the sequence and lower weights to less relevant ones, always in the range of $[0,1]$.

Finally, the attention scores are used to compute the weighted sum of the Value matrices (V). This process outputs the final representation of the query element, considering its relationship with all the other elements in the sequence.

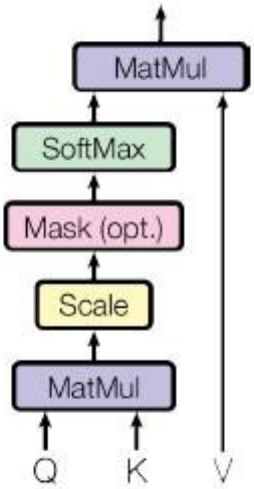

Figure 1. Scaled Dot-Product Attention.

The overall expression of scaled dot-product attention:

$$Attention(Q,K,V) = softmax\left(\frac{K \cdot Q^T}{\sqrt{d_k}}\right) \cdot V$$

[1] Ashish Vaswani et al. «Attention is all you need». In: Advances in neural information processing systems. 2017, pp. 5998–6008.