Transformer Attention Mechanism

1. What are Key, Query, and Value in the attention mechanism?

The Query matrix is the matrix representing the question for our current token, that we are trying to find information for. The Key matrix defines how well each input token satisfies the question asked by the query token. It shows how well each input token attends to the Query matrix of the other input token. Lastly the Value(s) matrix are the actual values which will be retrieved if the query matrix matches closely the key matrix. It will embed a different meaning in the original embedded vector, so that it has a more nuanced position in the vector space. These matrices are created by multiplying the initial input embedding by three different weight matrices, learned during training:

$$W^Q, \quad W^K, \quad W^V$$

2. At which stage of the Transformer are they used?

These matrices are used both in the encoding and decoding process of our transformer architecture. In the encoding process it is used in the self-attention process, where each token is able to understand its relationship with every other token in the input sentence. In the decoder, these matrices are used for two distinct types of attention layers:

- **Masked self-attention:** where the decoder output is fed through an attention layers which masks future tokens, to avoid the model relying on future input data to make predictions (this attention layer is mostly used for model training).
- **Encoder-Decoder attention:** Here the Key and Value matrices are the output of the encoder, while the Query matrices are the ones from the previous decoder. So the decoder "talks" with the encoder to find which original tokens are most relevant for generating the next output token.

3. How can they be intuitively interpreted?

The query matrix is equivalent to the question each token asks to every other token in the input. It could be any question regarding its relationship with every other token in the input, such as: "Do I have an article before me", for text inputs, or "Are my sorrounding pixel tokens brighter or darker than me?", for image inputs. The key matrix is the response each token gives to the query matrix, it's bigger the better the answer to the query is. The value matrix could be interpreted as the actual response given by the token, which is taken into consideration, only if the key and query match closely.