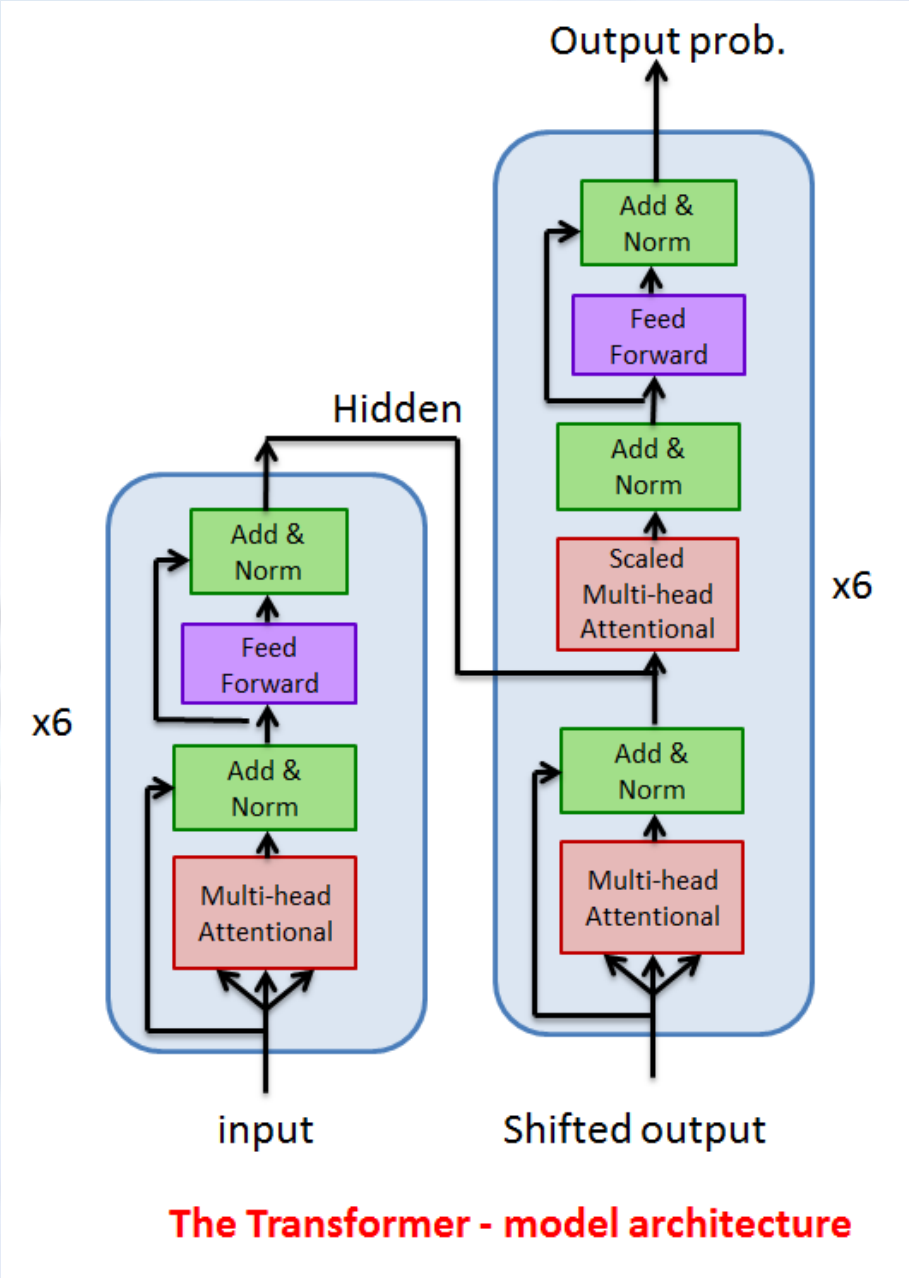
Multi-head attention linearly projects the queries, keys and values h times with different, learned linear projections. Different d_v dimensional output values are achieved by the various projections of queries, keys and values. The final output of the multi-head attention is a concatenation of all the various projections. See figure below.



The Transformer - model architecture

Multi-head attention is used at three different locations in Transformer:

1. The first location is the "encoder-decoder attention" layers; it takes the queries from the preceding decoder layer, and the keys and values from the encoder output. This allows each decoder position to attend to every input sequence position, resembling a typical encoder-decoder attention mechanism in sequence-to-sequence models.

2. The second location is at the self-attention layers of the encoder. Self-attention refers to the fact that in these layers all of the keys, values and queries are from the same source: in this case, the output of the preceding encoder layer.