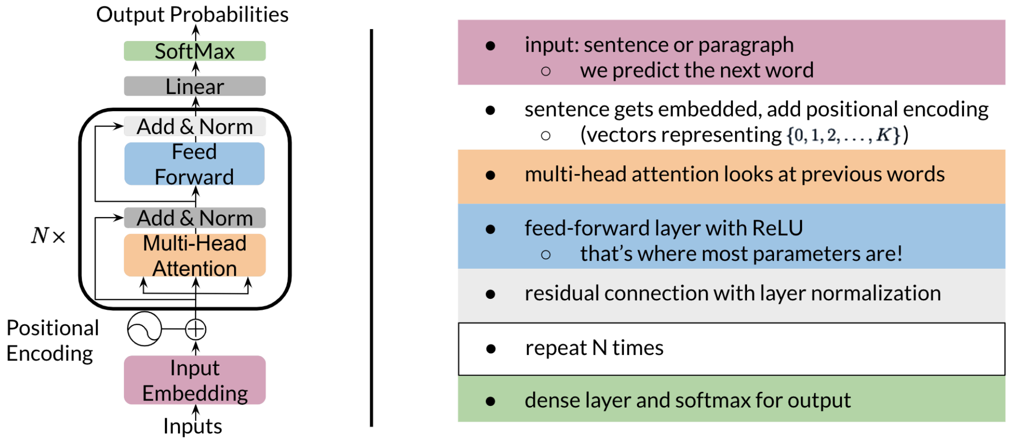
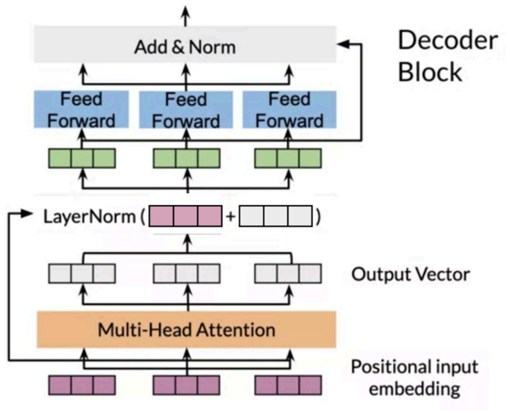
That was a lot of information! Don't worry if you did not understand everything, we will go over everything step by step.

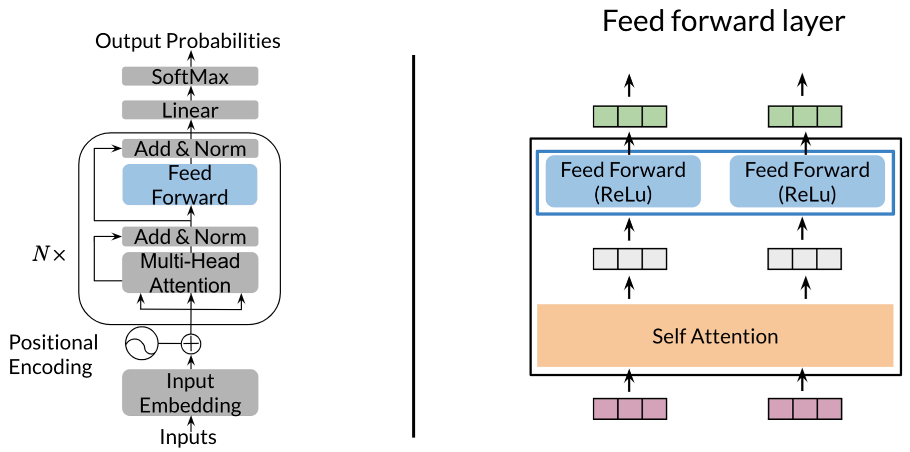


- input: sentence or paragraph
  - we predict the next word
- sentence gets embedded, add positional encoding
  - (vectors representing $\{0, 1, 2, \ldots, K\}$)
- multi-head attention looks at previous words
- feed-forward layer with ReLU
  - that's where most parameters are!
- residual connection with layer normalization
- repeat N times
- dense layer and softmax for output

Once you get the embeddings, you add the positional encoding, which you can think of just a number that tells you information about the position of the word. Then, you do multi-head attention as explained in the previous video/reading. There is a feedforward layer (blue) with a ReLU (and optionally some linear layers) after this, then a residual connection with layer normalization (repeat the block shown above N times), finally a linear layer with a softmax. Zooming into the decoder block that gets repeated N times, you get the following:



Now for the feedforward block, you can just implement the following:



You get the input, (red vector) run it through self-attention and then a feedforward with ReLU. At the end of the decoder, you can just run a linear layer and a softmax.