MT Summer Term 2021 Ex14: Attention is all you need

1. In your own words, please explain the following pictorial representation of the transformer NMT system (source: Vaswani et al. 2017):

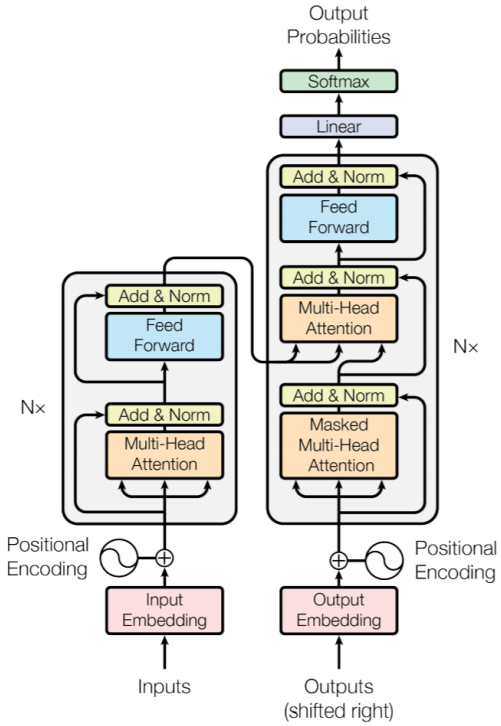


Figure 1: The Transformer - model architecture.

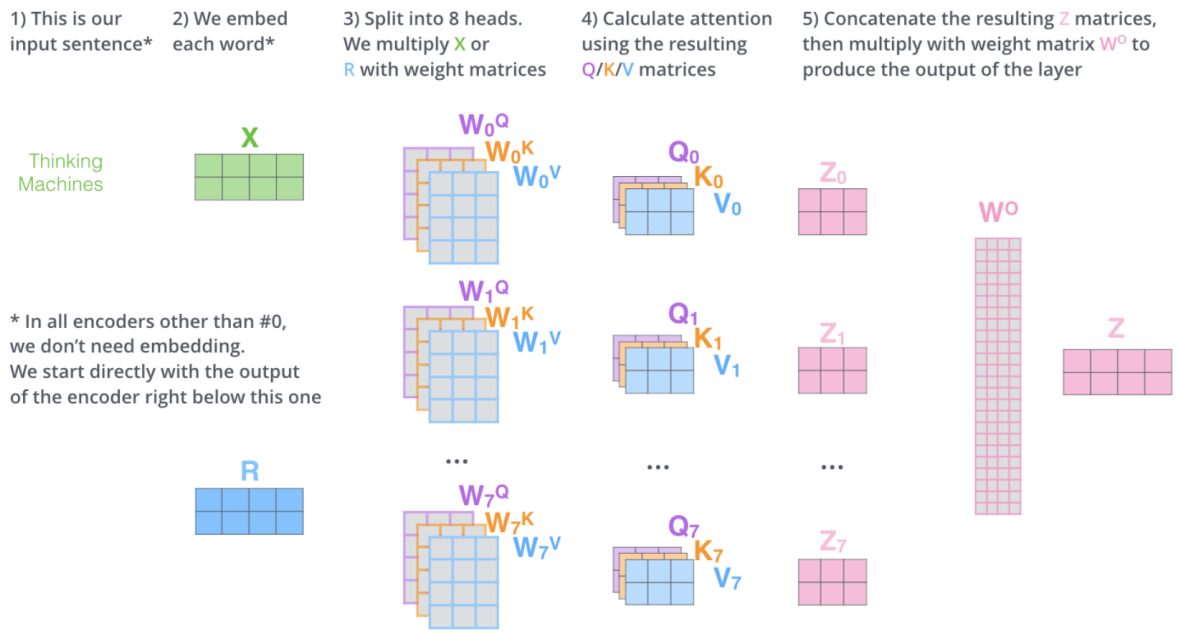
In your discussion, please cover

- Encoder-decoder stacks
- Self-attention
- Cross-attention
- Masked-attention
- Multi-head attention
- Query, key and values in attention computation
- Positional encoding
- Residual connections and layer normalisation
- Static embeddings and contextualised embeddings
- Autoregressive decoding

2. Are parameters shared between encoder blocks in the Transformer? Are parameters shared between decoder blocks in the Transformer? How does this compare to RNNs?

3. Comparing Transformer with simple (!) RNN based encoder-decoder systems, what are their computational complexities?

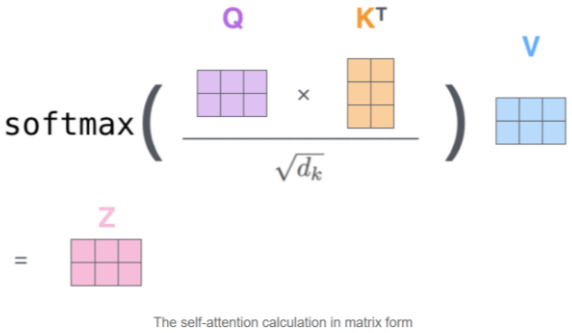
4. Why do you need the query-, key- and value-projection matrices in the picture below:



5. In your own words please explain:

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

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



The self-attention calculation in matrix form

6. Which parts of the Transformer make up BERT? Which parts of the Transformer make up GPT-X?

7. Please read the "Attention is all you need" paper.