Week 3 Questions

Attention is All You Need

1. Why do they separate their model into an encoder and a decoder?
2. What is the softmax function?
3. What are residual connections?
4. What is layer normalization?
5. Write down the dimensions of the keys, queries, values, and go through the operations in the attention blocks to make sure that these dimensions are actually compatible with the operations in the attention block.
6. Does the encoder information go to the decoder as keys, values, or queries? Why wouldn’t it work to send encoder outputs as queries and have the decoder outputs from the previous layer go to the keys and values?
7. Why do we need a mask in the self-attention blocks in the decoder? How does the mask work (be specific about what the paper refers to as ‘illegal connections’)? Why don’t we need the mask in the encoder-decoder attention blocks or the encoder attention blocks? Why are the masked values $-\infty$ and not 0 or some other number?
8. (Difficult) Why do the authors decide to use positional embeddings? What would happen if we try removing these? Would the effect of removing these be different for an encoder transformer vs a decoder transformer?

The Illustrated GPT-2

1. What is the difference between the decoder blocks from the Attention is All You Need paper and the decoder blocks in GPT-2?
2. The transformer in Attention is All You Need has an encoder and a decoder while GPT-2 is only a decoder. What differences might this make in terms of what tasks they perform well at? Are there certain problems an encoder-decoder transformer can solve that a decoder transformer cannot?