Tutorial 9

Question 1

Consider a question-answering system where an attention-based seq2seq model is given a passage: "The Nile is the longest river in the world. It flows through northeastern Africa." The system needs to answer questions based on this passage.

1) Describe the step-by-step decoding process that the seq2seq model with attention uses to generate each token to answer the question "Which river is the longest?". What is the input for this task? (Assume the decoder hidden states are given as $s_1, s_2, ..., s_T$. The encoder hidden states for the input sequence are $h_1, h_2, ..., h_N$.)

2) Assume when the decoder is generating the first token, the attention distribution is:

The	Nile	is	the	longest	river	in	the	world	.	It
0.4	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0
flows	through	north-eastern	Africa	.	Which	river	is	the	longest	?
0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0

And the attention distribution for generating the second token is:

The	**Nile**	is	the	longest	river	in	the	world	.	It
0.0	0.6	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0
flows	through	north-eastern	Africa	.	Which	river	is	the	longest	?
0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0

Discuss what these attention distributions suggest? What could possibly be the two tokens being generated?

3) Now suppose the question has been changed to "Where does the longest river flow?". Will that affect the attention distributions when performing the decoding steps? Why? What's the token(s) with the highest attention score(s) if the model is well-trained?

4) Imagine the passage is much longer and contains detailed descriptions of several rivers. How might this additional information impact the attention mechanism during decoding? What are some strategies for modifying the attention mechanism to maintain accuracy and focus in long passages?

Question 2

Consider a self attention model that is processing the sentence "The cat sat on the mat, and it was happy, accompanied by another cat." This sentence is used as an input to a self-attention layer to obtain a hidden representation for each word.

1) Describe how self-attention captures the context of each word in the sentence. Specifically, detail how self-attention helps the model understand the relationship between "it" and the first "cat."

2) Walk through the computation of self-attention weights for the word "it" in the given sentence. Explain how the self-attention scores between "it" and all other words in the sentence are calculated.

3) There are two occurrences of the word "cat" in the sentence. How might the self-attention model differentiate between these instances when processing the word "it"? What mechanisms or additional model adjustments could help disambiguate references in such cases?

4) In self-attention, each word's hidden representation is influenced by other words in the sentence. After self-attention is applied, how might the representation for the word "cat" differ between its first and second occurrence?

Question 3 (Advanced)

Imagine using a Transformer model with a single encoder block and decoder block to translate the English sentence "The quick brown fox jumps over the lazy dog" into French.

1) Discuss how information is transferred from the encoder to the decoder in a Transformer model. How does self-attention function in this scenario (i.e., crossing the encoder and decoder)?

2) Explain the role of multi-head attention in the context of translating the phrase "jumps over." How might different attention heads capture various aspects of this phrase's meaning and grammar?

3) Explain the function of the "Add and Norm" layers found after each sub-layer in the encoder and decoder. Why are these layers critical to the Transformer's architecture?

4) In cases of translation ambiguity, such as the word "over" with multiple meanings, how might the Transformer model disambiguate the correct sense of the word during translation?

Coding exercises: Attention
https://colab.research.google.com/drive/1nAUqAP4BQF2xFSq5V-50Bn_xEzx_m0Jm?usp=sharing