Natural Language Processing

Lab 6

March 4, 2024

This lab sheet is to practice the concepts taught last week: encoder-decoder, and attention.

1. What is an encoder decoder model and what tasks is it used to perform?
    a. Bonus: can you use an encoder-decoder model to do text classification?

2. What neural architectures can be used for the encoder model? What about the decoder model?

3. Are encoder-decoder and sequence-to-sequence always the same terms?

4. Why do we need attention in neural networks? Which problems does it solve? How is it calculated?

5. Your encoder model has the following 6 “hidden states”:

    [4,6,7,2,5], [7,9,4,2,5], [9,8,4,6,6], [2,5,6,6,7], [7,3,6,2,3], [1,4,7,5,3]

    Your decoder current state is [3,1,7,1,1]

    Can you calculate the context representation for the decoder state, using dot-product attention?

6. What is the difference between causal and bidirectional attention?
    a. When do we use each of them?

7. Which of the following networks can be parallelized and why (not)?
    a. FFN
    b. CNN
    c. LSTM
    d. Attention
    e. RNN

8. What is the role of query, key, value?
    a. Can we have self-attention without them?

9. Is this a correct representation of the original transformer? Can you explain how it works:

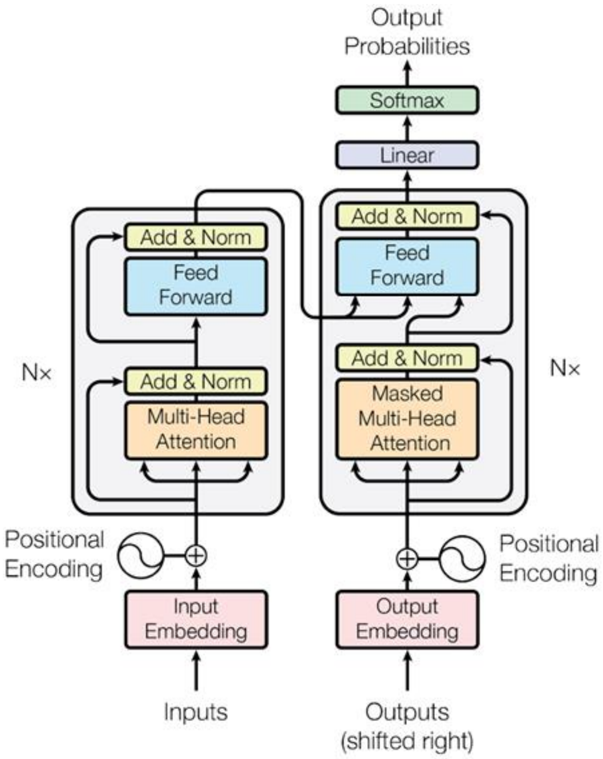

Figure 1: The Transformer - model architecture.