Habib University - City Campus
CS 335: Introduction to Large Language Models
Spring 2024
Activity Sheet # _____
Date _____

Name _____ Student ID: _____

Question 1: [0 points]

In the previous week, we worked with a single attention head. In this activity, we will working with multiple attention heads. Computing self-attention is comprised of two stages:

- In the first stage, we compute the matrices Q,K,and V which can be seen in the figure below.

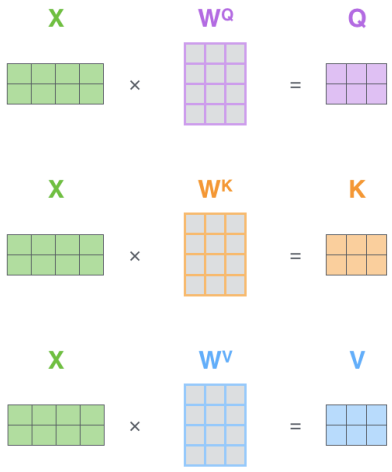

Figure 1: Computing the matrices Q,K,and V

- In the second stage, we compute the normalized attention scores. After normalisation, we apply SoftMax and multiply with V to get Z. Stage 2 is summarised in the figure below.

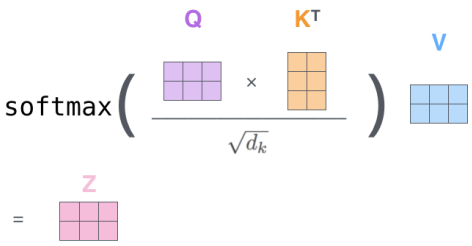

Figure 2: Computing the matrix Z

In this example, we will working with two attention heads and for each we will be calculating attention separately. This is highlighted in the figure below.

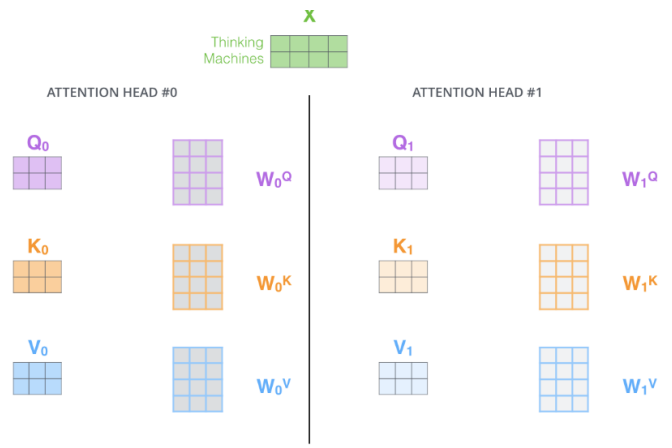

Figure 3: Computing attention with multiple heads.

After we have computed the attention for each head, we concatenate the result and multiply with W^o to compute the final output which is then passed onto the FFN.

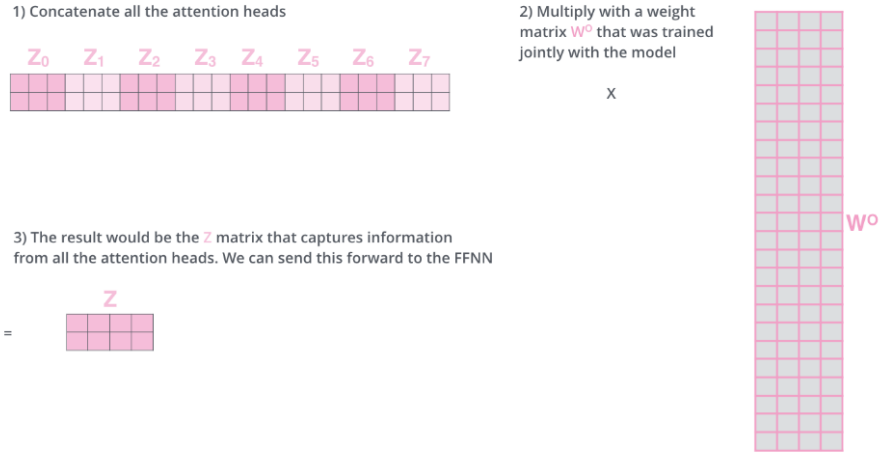

Figure 4: Computing attention with multiple heads.(8)

Assume that our input sequence is "Thinking Machine" and the embedding matrix X is as follows:

$$X = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}$$

where each row of X represents a word in the sequence.

The weight matrices for the first attention head are as follows:

$$W_{0^Q} = \begin{bmatrix} 0.1 & 0.2 & 0.3 \\ 0.4 & 0.5 & 0.6 \\ 0.7 & 0.8 & 0.9 \\ 1.0 & 1.1 & 1.2 \end{bmatrix}, W_{0^K} = \begin{bmatrix} 0.3 & 0.2 & 0.1 \\ 0.6 & 0.5 & 0.4 \\ 0.9 & 0.8 & 0.7 \\ 1.2 & 1.1 & 1.0 \end{bmatrix}, W_{0^V} = \begin{bmatrix} 0.7 & 0.8 & 0.9 \\ 0.4 & 0.5 & 0.6 \\ 0.1 & 0.2 & 0.3 \\ 1.3 & 1.4 & 1.5 \end{bmatrix}$$

The weight matrices for the second attention head are as follows:

$$W_{1^Q} = \begin{bmatrix} 0.2 & 0.3 & 0.4 \\ 0.5 & 0.6 & 0.7 \\ 0.8 & 0.9 & 1.0 \\ 1.1 & 1.2 & 1.3 \end{bmatrix}, W_{1^K} = \begin{bmatrix} 0.4 & 0.3 & 0.2 \\ 0.7 & 0.6 & 0.5 \\ 1.0 & 0.9 & 0.8 \\ 1.3 & 1.2 & 1.1 \end{bmatrix}, W_{1^V} = \begin{bmatrix} 0.9 & 1.0 & 1.1 \\ 0.6 & 0.7 & 0.8 \\ 0.3 & 0.4 & 0.5 \\ 1.4 & 1.5 & 1.6 \end{bmatrix}$$

(a) Compute $Q_0, K_0, V_0, Q_1, K_1, V_1$

(b) For each attention head, compute the following equation:

$$Z_i = softmax(\frac{Q_i \cdot K_i^\top}{\sqrt{d_k}}) \times V_i$$

where $\sqrt{d_k}$ is the dimension of the key matrix.

(c) Obtain Z by concatenating Z_0 and Z_1

(d) Assume that W_o is defined as below, compute the final output that will be passed onto the FFN.

$$W_0 = \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 \\ 0.5 & 0.6 & 0.7 & 0.8 \\ 0.9 & 1.0 & 1.1 & 1.2 \\ 1.3 & 1.4 & 1.5 & 1.6 \\ 1.7 & 1.8 & 1.9 & 2.0 \\ 2.1 & 2.2 & 2.3 & 2.4 \end{bmatrix}$$

Note: Since the calculations are quite cumbersome to be performed by hand, you may use Python or Mathlab to perform calculations

All the figures are taken from http://jalammar.github.io/illustrated-transformer/