CS 335: Introduction to Large Language Models

Habib University

Activity Sheet 05

Name: _____ ID: _____

Question 01: Multihead Self Attention

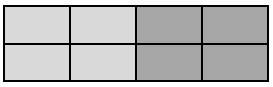
Multi-head self-attention is computed using the following formulas:

$$\mathbf{Q}^i = \mathbf{X}\mathbf{W}_i^{\mathbf{Q}} \qquad \mathbf{K}^i = \mathbf{X}\mathbf{W}_i^{\mathbf{K}} \qquad \mathbf{V}^i = \mathbf{X}\mathbf{W}_i^{\mathbf{V}}$$

$$\text{head}_i = \text{Self}-\text{Attention}\ (\mathbf{A}) = \left(\text{softmax}\left(\frac{\mathbf{Q}^i {\mathbf{K}^i}^{\mathbf{T}}}{\sqrt{d_k}}\right)\right)\mathbf{V}^i$$

$$\text{MultiHead Attention}\ (\mathbf{M}) = (\text{head}_1 \oplus \text{head}_2\ \ldots \oplus \text{head}_h)\mathbf{W}^0$$

The concatenation simply joins the results of attention heads together



head_1 head_2

As an example, consider the following results for 2 attention heads

$$\text{head}_1 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \qquad \text{head}_2 = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$$

The concatenation will be

$$\text{head}_1 \oplus \text{head}_2 = \begin{bmatrix} 1 & 2 & 5 & 6 \\ 3 & 4 & 7 & 8 \end{bmatrix}$$

Given the input matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

where each row represents a word in the input sequence, and the weight matrices for two attention heads are:

Head 1:

$$\mathbf{W}_1^{\mathbf{Q}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{W}_1^{\mathbf{K}} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \mathbf{W}_1^{\mathbf{V}} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

Head 2:

$$\mathbf{W}_2^{\mathbf{Q}} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \mathbf{W}_2^{\mathbf{K}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{W}_2^{\mathbf{V}} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

Final Projection Matrix:

$$\mathbf{W}^{o} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}$$

Compute the Multi-Head Self-Attention output (M).