No.

Date

Theory of Deep Learning
Homework 4
Part II

Q1 We have input vectors [1012 11], [1201 11].

$\begin{bmatrix} 1 & 0 & 1 & 2 & 1 & 1 \\ 1 & 2 & 0 & 1 & 1 & 1 \end{bmatrix}$

1. Output vectors of first attention head will be given by

$\text{softmax}\left(\frac{Q_1 K_1^T}{\sqrt{d_k}}\right) V_1$ where Q_1, K_1, V_1

$\Rightarrow Q \times W_1^Q$

$K \times W_1^K$

$V \times W_1^V$

So, $Q_1 = \begin{bmatrix} 1 & 0 & 1 & 2 & 1 & 1 \\ 1 & 2 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 4 & 3 \\ 3 & 5 & 2 \end{bmatrix}_{(2\times 3)}$

$K_1 = \begin{bmatrix} 1 & 0 & 1 & 2 & 1 & 1 \\ 1 & 2 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \\ 2 & 1 & 0 \\ 3 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 7 & 8 & 6 \\ 6 & 8 & 5 \end{bmatrix}$

$V_1 = \begin{bmatrix} 1 & 0 & 1 & 2 & 1 & 1 \\ 1 & 2 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 2 & 1 \\ 1 & 2 & 3 \\ 1 & 2 & 1 \\ 2 & 1 & 2 \\ 1 & 1 & 1 \\ 3 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 12 & 8 & 10 \\ 11 & 9 & 13 \end{bmatrix}$

Now, $Q_1 K_1^T$ values for the two vectors are:

$$[4\ 4\ 3]_{(1\times3)} \begin{bmatrix} 7 \\ 8 \\ 6 \end{bmatrix}_{(3\times1)} = [78] \quad \& \quad [3\ 5\ 2] \begin{bmatrix} 6 \\ 8 \\ 5 \end{bmatrix} = [68]$$

$$Q_1 K_1^T = \begin{bmatrix} 4 & 4 & 3 \\ 3 & 5 & 2 \end{bmatrix}_{(2\times3)} \begin{bmatrix} 7 & 6 \\ 8 & 8 \\ 6 & 5 \end{bmatrix}_{(3\times2)} = \begin{bmatrix} 78 & 71 \\ 73 & 68 \end{bmatrix}$$

(1) Dividing each element by $\sqrt{d_k} \Rightarrow \sqrt{3}$ and running softmax we get output vector as multiplied by V

$$\begin{bmatrix} 0.983 & 0.017 \\ 0.947 & 0.053 \end{bmatrix} \begin{bmatrix} 0.359 & 0.987 \\ 0.640 & 0.017 \end{bmatrix}_{(2\times2)} \begin{bmatrix} 12 & 8 & 10 \\ 11 & 9 & 13 \end{bmatrix}_{(2\times3)}$$

$$= \begin{bmatrix} 11.983 & 8.017 & 10.051 \\ 11.947 & 8.053 & 10.159 \end{bmatrix} \approx \begin{bmatrix} 15.165 & 11.735 & 16.421 \\ 7.867 & 5.273 & 6.621 \end{bmatrix}$$

Now, for head 2:

$$Q_2 = \begin{bmatrix} 1 & 0 & 1 & 2 & 1 & 1 \\ 1 & 2 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 2 \\ 2 & 1 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 6 & 4 \\ 5 & 5 & 7 \end{bmatrix}$$

$$K_2 = \begin{bmatrix} 1 & 0 & 1 & 2 & 1 & 1 \\ 1 & 2 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 3 \\ 2 & 3 & 2 \end{bmatrix}$$

$$V_2 = \begin{bmatrix} 1 & 0 & 1 & 2 & 1 & 1 \\ 1 & 2 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 2 & 4 \\ 4 & 2 & 3 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 3 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 9 & 6 & 9 \\ 16 & 9 & 14 \end{bmatrix}$$

Now, similarly to need 1 we have:

$$Q_2 K_2^T = \begin{bmatrix} 3 & 6 & 4 \\ 5 & 5 & 7 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 3 & 3 \\ 3 & 2 \end{bmatrix} = \begin{bmatrix} 36 & 32 \\ 46 & 39 \end{bmatrix}$$

② After dividing matrix by $\sqrt{3}$ & applying softmax then multiplying by values matrix we get:

$$\begin{bmatrix} 0.909 & 0.091 \\ 0.983 & 0.017 \end{bmatrix} \begin{bmatrix} 0.003 & 0.017 \\ 0.996 & 0.782 \end{bmatrix} \begin{bmatrix} 9 & 6 & 9 \\ 16 & 9 & 14 \end{bmatrix} = \begin{bmatrix} 0.299 & 0.171 & 0.265 \\ 24.676 & 14.814 & 22.712 \end{bmatrix}$$

③ Now, to get final output, we concatenate the two outputs from the 2 needs & multiply with the Wo matrix giving us:

$$\begin{bmatrix} 9.637 & 6.273 & 9.455 \\ 9.119 & 6.051 & 9.085 \end{bmatrix}$$

$$\begin{bmatrix} 15.165 & 11.755 & 16.421 & 0.299 & 0.171 & 0.265 \\ 7.867 & 5.273 & 6.621 & 24.676 & 14.814 & 22.712 \end{bmatrix}$$

$$\begin{bmatrix} 9.637 & 6.273 & 9.455 \\ 9.119 & & \end{bmatrix} \begin{bmatrix} 11.983 & 8.017 & 10.051 & 9.637 & 6.273 & 9.455 \\ 11.947 & 8.053 & 10.159 & 9.119 & 6.051 & 9.085 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 53.854 & 43.615 & 29.325 & 37.944 & 57.723 & 20.102 \\ 52.446 & 42.501 & 28.397 & 37.276 & 56.571 & 20.318 \end{bmatrix}$$

No.

Date

Q2 To explain the complexity expression given above, we'll go step by step to see if we can arrive at the same expression ourselves.

Firstly, we let X be the complete input to the self-attention layer where X would have a dimensionality of (N', D) with N' being the number of input embedding vectors & D the overall dimension of each.

Now, we need to perform a linear transformation to each of the vectors in X with the weight matrices to compute the Q, K, V matrices. This would result in $3N'D^2$ operations / complexity

Next, to compute $Q \times K^t$ & then $(Q \times K^t) \times V$ we need another $2N'^2D$ operations to be led.

After which, we have the final matrix multiplication of this result with the weight matrix W^o which would again be $N'D^2$ complexity.

Ultimately, we have the overall complexity of the MSA as:

$$\Omega(3N'D^2 + 2N'D^2 + N'D^2) \Rightarrow \Omega(4N'D^2 + 2N'^2D)$$

[x] [y] [x] [y]
1 2 3 4 5 6 7

Q3 The EMD of transporting the distribution of masses P_2 to P_3 as shown below would be

$|4-1| + |6-2| \Rightarrow 7$ //

This is the optimal way to transport the blocks of masses hence this is the EMD.

We can also confirm this using the transport plan table

x (From) \ y (To)	4	6
1	1	0
2	0	1

$\Rightarrow$ 7 //. (If you move them)