Quiz 10 CS 725 2023

1. Consider a transformer model with an input sequence of length 4, an embedding dimension of 8, and 2 attention heads. Calculate the total number of parameters in the self-attention mechanism, including weight matrices and biases (assume that the keys, queries and value vectors are of dimension 4) [2 marks]

   Answer: 192, For the self-attention mechanism, there are weight matrices for each head, including the query, key, and value matrices. Each of these matrices will have dimensions (num-embed , dim-of-vectors), in this case (8,4). So, for 2 attention heads, the total number of parameters is 2 * (8 * 4 * 3) = 192 parameters.

   If bias is included, we would 216 total parameters. We have given credit for this too.

   Some of the original implementations also include an additional linear layer to combine the attention heads. This should bring your answer to 256. We have given credit for this answer too.

2. Bhupendra Jogi is training a Seq-to-Seq model with attention. The encoder states are as given below:

$$h_1 = \begin{bmatrix} 0.6 \\ -0.2 \\ 0.2 \end{bmatrix}$$

$$h_2 = \begin{bmatrix} 0.4 \\ 0.2 \\ 0.2 \end{bmatrix}$$

$$h_3 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

   Suppose he gets the decoder hidden state at timestep 4 as

$$\begin{bmatrix} 0.2 \\ 0.1 \\ -0.5 \end{bmatrix}$$

   Help him calculate the score for each of the encoder state and calculate the final context vector for timestep 4. (Take metric to measure similarity as cosine similarity) [4 marks]

   Fill in the values for the softmax scores for the embeddings and the final context vector. Take ln 2 as 0.7 and ln 3 as 1.1

   $Scores = [a, b, c]$ and $d_4 = \begin{bmatrix} d \\ e \\ f \end{bmatrix}$

   Answer : [0.25,0.25,0.5], [0.75,0,-0.4]

   Solution: To get the scores, we first take the dot product of vector of decoder hidden state with each of encoder states. This gives $[0, 0, 0.7]$ Taking softmax of this, we get(using $\ln 2 = 0.7$, i.e. $0.7 = e^2$) $Scores = [0.25, 0.25, 0.5]$

   Now, multiplying each vector by it's softmax score and taking the sum of these weighted vectors, we get the required answers

   i.e.

$$0.25 * h_1 + 0.25 * h_2 + 0.5 * h_3 = [0.75, 0, -0.4]$$

3. Elvis bhai is training a transformer model. The key, query and value matrices are as given below:

$$K = \begin{bmatrix} 0 & 1 & 0 \\ 1.1 & 0 & 0 \end{bmatrix}$$

$$Q = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.6 & -0.1 \\ 0.2 & 3 \end{bmatrix}$$

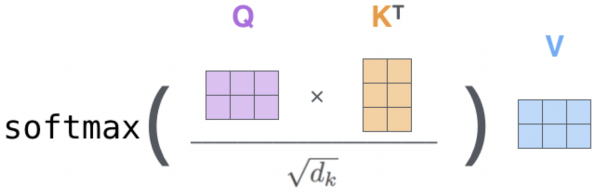
Help him calculate the final output of the self attention layer (Take $d_k = 1$) [3 marks] Take ln 2 as 0.7 and ln 3 as 1.1.

$$Output = \begin{bmatrix} w & x \\ y & z \end{bmatrix}$$

Answer:

$$Output = \begin{bmatrix} 0.3 & -0.05 \\ 0.15 & 0.75 \end{bmatrix}$$

Solution: Since we are already given all of the Key, Query and Value matrices, we simply need to do:



$$QK^T = \begin{bmatrix} 0 & 1.1 \\ 0 & 0 \end{bmatrix}$$

Taking the softmax scores, (using ln 3 = 1.1, i.e. $3 = e^{1.1}$) we get

$$\begin{bmatrix} 0.5 & 0.75 \\ 0.5 & 0.25 \end{bmatrix}$$

(The softmax score would be along the last dimension)

Taking it's matrix product with the value matrix, we get

$$\begin{bmatrix} 0.45 & 2.2 \\ 0.35 & 0.7 \end{bmatrix}$$

4. Which of the following is **not** a capability of BERT?

- Text generation
- Text classification
- Relabelling words in a sentence
- Generate good text embeddings

Answer Text generation