3) Calculate the number of parameters in a single Transformer encoder layer given the following:
Input dimension: 64
Number of heads in multi-head attention: 4
Dimension of each head: 16
Dimension of feed-forward network: 120
Assume the weight matrices for the linear transformations in multi-head attention and the feed-forward network are the primary contributors to the parameter count.

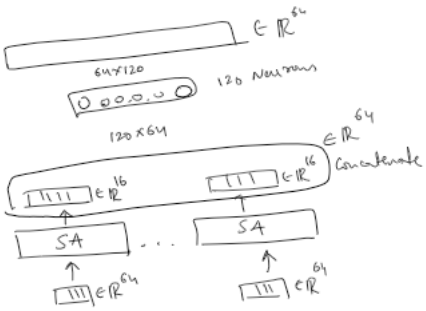
27648

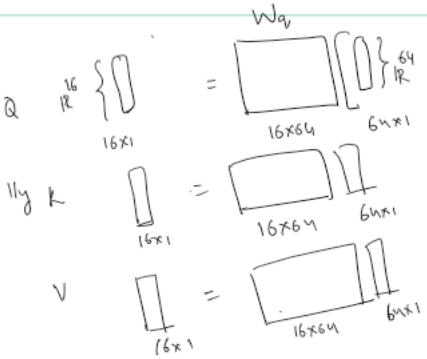
Yes, the answer is correct.
Score: 1

Accepted Answers:

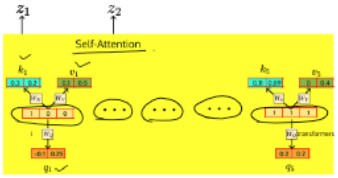
(Type: Numeric) 27648

parameter W_q, W_k, W_v $\left[3(16\times64)\right]\times4 + (64\times120)\times2$ encoder





Let's focus on first calculating the first output from self-attention layer



Fixed variable

$score(s_{t-1}, h_j)$
$score(q_2, k_j)$

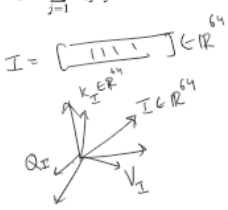
Score func: dot product

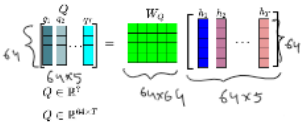
$e_{2j} = [q_2 \cdot k_1, \quad q_2 \cdot k_2, \quad \cdots, \quad \cdots \quad q_2 \cdot k_5]$
$\alpha_{2j} = softmax(e_{2j})$
$z_2 = \sum_{j=1}^{5} \alpha_{2j} v_j$

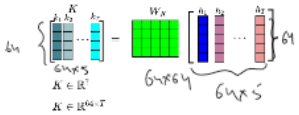
attention score

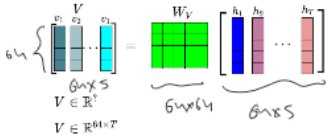
Repeat the procedure for all other z.

alignment score









lets say each word $\in \mathbb{R}^{64}$

- Sequence Length : t = 32
- Number of Heads : h = 2
- Embedding dimension : d_{model}
- Input $X \in \mathbb{R}^{d_{\text{model}} \times t}$
- $d_k = d_q = \frac{d_{\text{model}}}{h}$ = 32
- $W_Q \in \mathbb{R}^{d_q \times d_{\text{model}}}$
- $W_K \in \mathbb{R}^{d_k \times d_{\text{model}}}$
- $W_V \in \mathbb{R}^{d_v \times d_{\text{model}}}$
- $W_o \in \mathbb{R}^{d_{\text{model}} \times (h \times d_v)}$ 64 × (2×16) 64×32

Suppose $t = 32$, $d_{\text{model}} = 64$, $h = 2$ and $d_v = 16$. What will be the shape of the output of the scaled dot-product attention operation for a single head, given by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q^T K}{\sqrt{d_k}}\right) V^T$$

Compute the resulting output dimension and report the **total number of elements** in the resulting attention output. 512