

$\bar{x}_i \rightarrow \bar{q}_i, \bar{k}_i, \bar{v}_i$

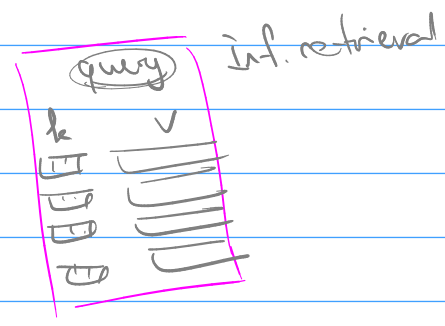
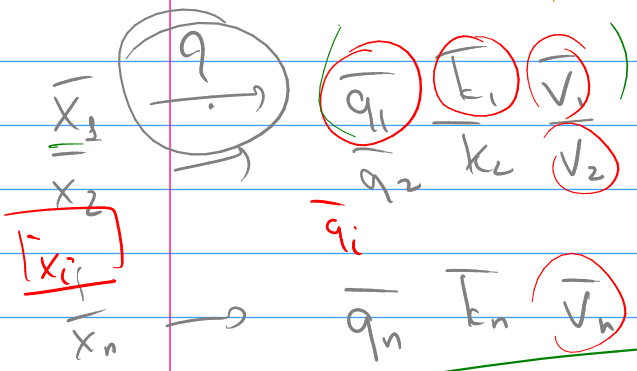


Diagram showing the dot product calculation for attention weights. A query vector  $\bar{q}_i$  is compared against key vectors  $\bar{k}_1, \dots, \bar{k}_n$  to produce attention weights  $\frac{\bar{q}_i^T \bar{k}_j}{\bar{q}_i^T \bar{k}_i}$ . These weights are then multiplied by value vectors  $\bar{v}_j$  to produce the attended output  $\bar{z}_i$ .

$$\bar{z}_i = \sum_{j=1}^n \text{softmax} \left( \frac{\bar{q}_i^T \bar{k}_j}{\bar{q}_i^T \bar{k}_i} \right) \bar{v}_j$$

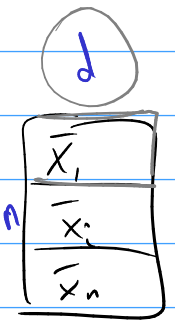


Diagram showing the matrix representation of the attention mechanism. The input matrix  $X$  is transformed into query matrix  $Q$  (dimension  $n \times m$ ), key matrix  $K$  (dimension  $n \times m$ ), and value matrix  $V$  (dimension  $n \times l$ ). The output is  $Z$  (dimension  $n \times l$ ).

$$Q \quad K \quad V \quad = \quad Z$$

Attention head

$s = 0.7$

Multi-head attention

Diagram showing the matrix multiplication for multi-head attention. The query matrix  $Q$  (dimension  $n \times n$ ) is multiplied by the key matrix  $K^T$  (dimension  $n \times n$ ) to produce a matrix of dimension  $n \times n$ . This matrix is then multiplied by the value matrix  $V$  (dimension  $n \times l$ ) to produce the final output  $Z$  (dimension  $n \times l$ ).

$$\text{softmax} \left( \frac{1}{\sqrt{m}} (QK^T) \right) V = \sum_{j=1}^n \text{softmax} \left( \bar{q}_i^T \bar{k}_j \right) \bar{v}_j$$

X

$d \times m$   
 $W_s^Q, W_s^K, W_s^V$

$$\bar{x}_i \quad \bar{q}_i = \bar{x}_i W_s^Q$$

$$\bar{k}_i = \bar{x}_i W_s^K$$

$$\bar{v}_i = \bar{x}_i W_s^V$$

$$Q_s = X W_s^Q$$

$$K_s = X W_s^K$$

$$V_s = X W_s^V$$

weights

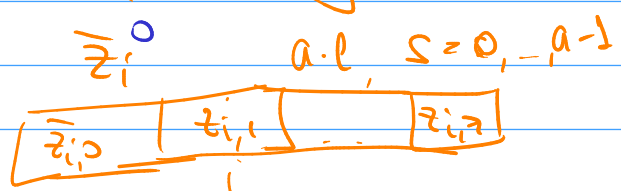
multi-heads weights

$$\begin{bmatrix} W_0^Q & \dots & W_7^Q \\ W_0^K & \dots & W_7^K \\ W_0^V & \dots & W_7^V \end{bmatrix}$$

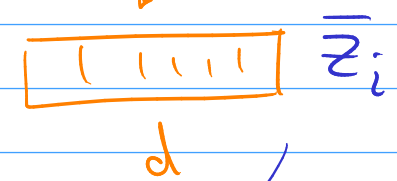
$$d \rightarrow W_0^- X \rightarrow \dots \rightarrow z_{i,0}$$

$$\bar{x}_i \rightarrow W_1^- X \rightarrow \dots \rightarrow z_{i,1}$$

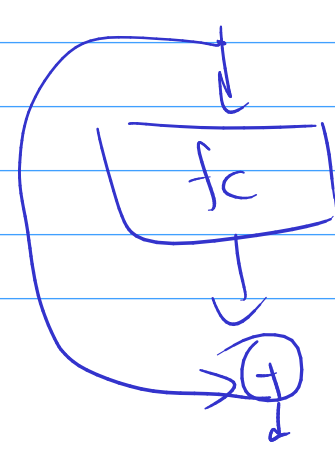
$$W_7^- X \rightarrow \dots \rightarrow z_{i,7}$$



$$W^0 (a \cdot l) \times d$$



$$\text{Norm}(\bar{x}_i + \bar{z}_i) = \bar{x}_i'$$



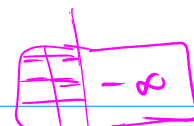
Decoder

$n \times d$   
z

$W^k$   $W^v$

Attention.

$K, V$   $z \rightarrow Q$   
 $W^Q$



mask

Self-attention

$$\frac{1}{\sqrt{m}}(QK^T)$$

