Let model dimensions (embedding dims) be 512.
consider an input sentence with 5 words.

$\tilde{I} \in R^{5 \times 512}$

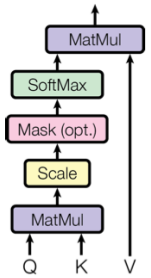
Denote randomly initialized query, key, and value matrices by $\tilde{Q}_w, \tilde{K}_w, \tilde{V}_w$. Each of $\tilde{Q}_w, \tilde{K}_w, \tilde{V}_w \in R^{512 \times 512}$. To obtain queries ($\tilde{Q}$), keys ($\tilde{K}$), and values ($\tilde{V}$) mutiply the input ($\tilde{I}$) with $\tilde{Q}_w, \tilde{K}_w, \tilde{V}_w$. (Assume that we have only one head)

$$\underbrace{\tilde{I}}_{R^{5\times512}} \times \underbrace{\tilde{Q}_w}_{R^{512\times512}} = \underbrace{\tilde{Q}}_{R^{5\times512}}$$

$$\underbrace{\tilde{I}}_{R^{5\times512}} \times \underbrace{\tilde{K}_w}_{R^{512\times512}} = \underbrace{\tilde{K}}_{R^{5\times512}}$$

$$\underbrace{\tilde{I}}_{R^{5\times512}} \times \underbrace{\tilde{V}_w}_{R^{512\times512}} = \underbrace{\tilde{V}}_{R^{5\times512}}$$

Computational graph of the scaled dot product attention is given on the right, we will look into it in depth.



$\tilde{Q} \in R^{5\times 512}$, $\tilde{K} \in R^{5\times 512}$, $\tilde{V} \in R^{5\times 512}$

$$Attn(\tilde{Q},\tilde{K},\tilde{V}) = Softmax\left(\frac{\tilde{Q}\tilde{K}^T, axis=1}{\sqrt{d_k}}\right)\times \tilde{V}, \; d_k = 512$$

$\tilde{Q}\tilde{K}^T =$

$$\begin{bmatrix} q_{11} & \cdots & q_{1,512} \\ q_{5,1} & & q_{5,512} \end{bmatrix}_{R^{5\times 512}} \begin{bmatrix} k_{11} & \cdots & k_{15} \\ \vdots & & \vdots \\ k_{512,1} & & k_{512,5} \end{bmatrix}_{R^{512\times 5}} = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}_{R^{5\times 5}}$$

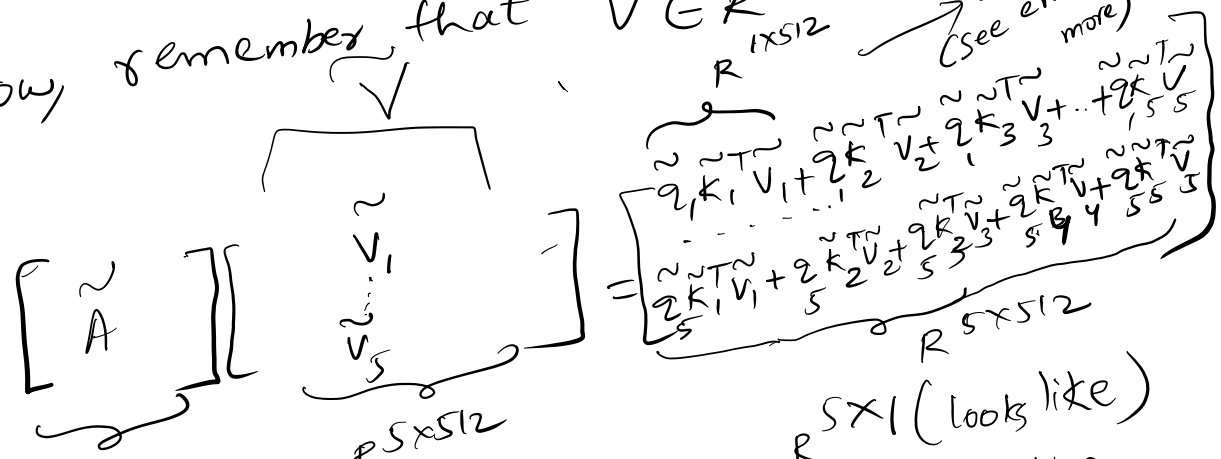
let each row of Q be $\tilde{q}_i$ and each column of K^T be $\tilde{k}_i^T$

$$\begin{bmatrix} \tilde{q}_1 \\ \vdots \\ \tilde{q}_5 \end{bmatrix}_{R^{5\times 512}} \begin{bmatrix} \tilde{k}_1^T & \cdots & \tilde{k}_5^T \end{bmatrix}_{R^{512\times 5}} = \begin{bmatrix} \tilde{q}_1\tilde{k}_1^T & \tilde{q}_1\tilde{k}_2^T & \tilde{q}_1\tilde{k}_3^T & \tilde{q}_1\tilde{k}_4^T & \tilde{q}_1\tilde{k}_5^T \\ \vdots & & & & \\ \tilde{q}_5\tilde{k}_1^T & \tilde{q}_5\tilde{k}_2^T & \tilde{q}_5\tilde{k}_3^T & \tilde{q}_5\tilde{k}_4^T & \tilde{q}_5\tilde{k}_5^T \end{bmatrix}_{R^{5\times 5}} \hookrightarrow \tilde{A}$$

each of the $\tilde{q}_i\tilde{k}_j^T$ yields a scalar as it is a dot

product $1 \le i \le 5, 1 \le j \le 5$

Now, remember that $\tilde{V} \in R^{5 \times 512}$ (row point of view (see end for more))



$R^{5 \times 5} \times$ looks like a $R^{5 \times 1} = R^{5 \times 1}$ (looks like)

Note that we have not done softmax yet on the attention matrix ($\tilde{A}$), $\tilde{A} \in R^{5 \times 5}$ and now we have to do a softmax operation on $\tilde{A}$.

Each row of $\tilde{A}$ linearly scales each row of $\tilde{V}$ to obtain a scaled version of the input, this scaled version of the input ($\tilde{O}$) contains the attention information because $\tilde{V}$ is scaled by the attention matrix. ($\tilde{A}$)

To obtain $\tilde{O}$, we must take softmax ($\tilde{A}$, axis=1), that is softmax operation is performed individually for each of the rows.

To obtain attention matrix $A \in \mathbb{R}^{5\times 5}$ from $\tilde{A}$. We perform

$$\begin{bmatrix} \frac{e^{\tilde{q}_1\tilde{k}_1^T}}{\sum_{j=1}^{5} e^{\tilde{q}_1\tilde{k}_j^T}} & \cdots & \frac{e^{\tilde{q}_1\tilde{k}_5^T}}{\sum_{j=1}^{5} e^{\tilde{q}_1\tilde{k}_j^T}} \\ \vdots & & \\ \frac{e^{\tilde{q}_5\tilde{k}_1^T}}{\sum_{j=1}^{5} e^{\tilde{q}_5\tilde{k}_j^T}} & \cdots & \frac{e^{\tilde{q}_5\tilde{k}_5^T}}{\sum_{j=1}^{5} e^{\tilde{q}_5\tilde{k}_j^T}} \end{bmatrix}$$

Matrix Multiplication Interpretations

I) Column point of view

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{pmatrix} = C$$

Columns of C are linear combination of columns of A with weights given by columns of B.

$$\begin{pmatrix} c_{11} \\ c_{21} \\ c_{31} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} \\ a_{31}b_{11} + a_{32}b_{21} + a_{33}b_{31} \end{pmatrix} = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} b_{11} + \begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \end{pmatrix} b_{21} + \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix} b_{31}$$

II) Row Point of view

Rows of C are linear combinations of the rows of B with weights given by rows of A.

$$\underset{A}{\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}} \underset{B}{\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}} = \underset{C}{\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{pmatrix}}$$

$$(c_{11}\ c_{12}) = (a_{11}b_{11}+a_{12}b_{21}+a_{13}b_{31} \quad a_{11}b_{12}+a_{12}b_{22}+a_{13}b_{32})$$

$$= (b_{11}\ b_{12})a_{11} + (b_{21}\ b_{22})a_{12} + (b_{31}\ b_{32})a_{13}$$