$$\begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1d} \\ q_{21} & q_{22} & \cdots & q_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N1} & q_{N2} & \cdots & q_{Nd} \end{pmatrix} \times \begin{pmatrix} k_{11} & k_{21} & \cdots & k_{N1} \\ k_{12} & k_{22} & \cdots & k_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ k_{1d} & k_{2d} & \cdots & k_{Nd} \end{pmatrix} \longrightarrow \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1N} \\ s_{21} & s_{22} & \cdots & s_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{NN} \end{pmatrix}$$

$Q\ (N \times d)$ $K^T\ (d \times N)$ $QK^T\ (N \times N)$

$$s_{11} = q_{11}k_{11} + q_{12}k_{12} + \cdots + q_{1d}k_{1d}$$

Attention score between Query and Key of Token 1

$$\begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1d} \\ q_{21} & q_{22} & \cdots & q_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N1} & q_{N2} & \cdots & q_{Nd} \end{pmatrix} \times \begin{pmatrix} k_{11} & k_{21} & \cdots & k_{N1} \\ k_{12} & k_{22} & \cdots & k_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ k_{1d} & k_{2d} & \cdots & k_{Nd} \end{pmatrix} \longrightarrow \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1N} \\ s_{21} & s_{22} & \cdots & s_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{NN} \end{pmatrix}$$

$Q\ (N \times d)$ $K^T\ (d \times N)$ $QK^T\ (N \times N)$

$$s_{12} = q_{11}k_{21} + q_{12}k_{22} + \cdots + q_{1d}k_{2d}$$

Attention score between Query of Token 1 and Key of Token 2

$$\begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1d} \\ q_{21} & q_{22} & \cdots & q_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N1} & q_{N2} & \cdots & q_{Nd} \end{pmatrix} \times \begin{pmatrix} k_{11} & k_{21} & \cdots & k_{N1} \\ k_{12} & k_{22} & \cdots & k_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ k_{1d} & k_{2d} & \cdots & k_{Nd} \end{pmatrix} \longrightarrow \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1N} \\ s_{21} & s_{22} & \cdots & s_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{NN} \end{pmatrix}$$

$Q\ (N \times d)$ $K^T\ (d \times N)$ $QK^T\ (N \times N)$

$$s_{21} = q_{21}k_{11} + q_{22}k_{12} + \cdots + q_{2d}k_{1d}$$

Attention score between Query of Token 2 and Key of Token 1

	Key1	Key2	Key j	KeyN
Query1	s_{11}	s_{12}	$\cdots$	s_{1N}
Query2	s_{21}	s_{22}	$\cdots$	s_{2N}
Query i	$\vdots$	$\vdots$	s_{ij}	$\vdots$
QueryN	s_{N1}	s_{N2}	$\cdots$	s_{NN}

Full Attention Score (QK^T matrix) for one head only