

Scaled Attention Scores $\mathbf{QK}^T / \sqrt{d_k}$

	\mathbf{k}_1	\dots	\mathbf{k}_n
\mathbf{q}_1	$Score_{q_1 k_1} / \sqrt{d_k}$	\dots	$Score_{q_1 k_n} / \sqrt{d_k}$
\vdots	\vdots	\ddots	\vdots
\mathbf{q}_n	$Score_{q_n k_1} / \sqrt{d_k}$	\dots	$Score_{q_n k_n} / \sqrt{d_k}$

Attention Weights $\text{softmax}(\mathbf{QK}^T / \sqrt{d_k})$



	\mathbf{k}_1	\dots	\mathbf{k}_n	
\mathbf{q}_1	$Weight_{q_1 k_1}$	\dots	$Weight_{q_1 k_n}$	$\Sigma = 1$
\vdots	\vdots	\ddots	\vdots	$\Sigma = 1$
\mathbf{q}_n	$Weight_{q_n k_1}$	\dots	$Weight_{q_n k_n}$	$\Sigma = 1$