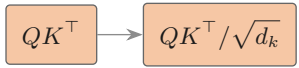
$$\text{Attention}(Q, K, V) = \text{softmax}(QK^\top/\sqrt{d_k})V$$

缩放来自方差，不是装饰



§3.2.1，避免 softmax 饱和

思路	点积方差随 d_k 线性涨
改写	缩放是改写，不是新算子
注意	没有缩放，softmax 进饱和区，梯度消失