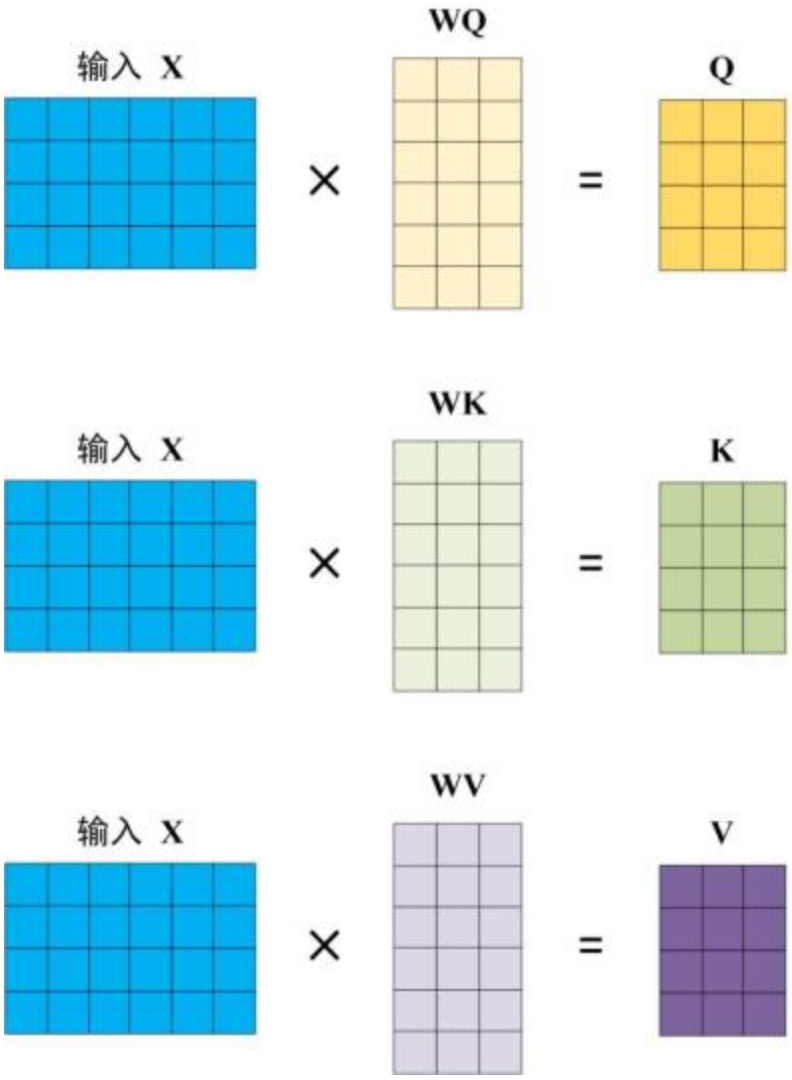
Attention

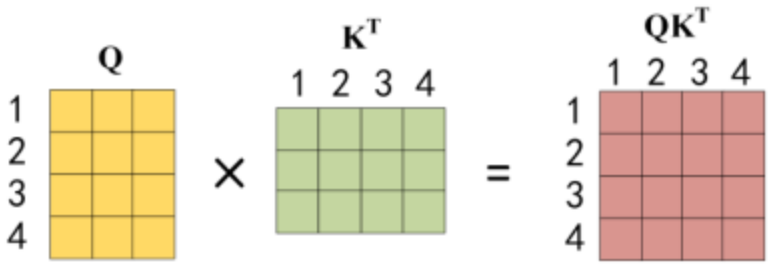
Attention机制用于引入全局视野，使得输入整体对于部分都有一定的影响。

首先使用输入x与参数矩阵相乘得到Q，K，V三个矩阵。（Q，K，V并非需要都来自同一个输入，但如果输入全部相同，可以称为self-attention）

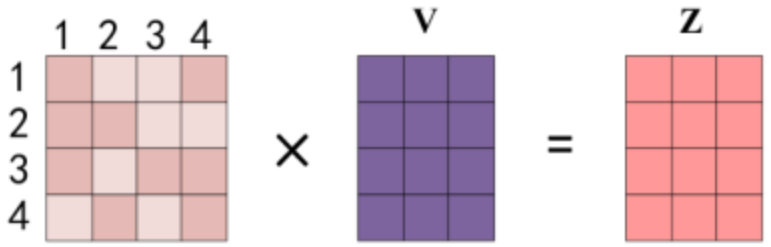
且在输入X中，每一行代表了一个独立的单元，在求得的QKV中也是如此。



将Q，K相乘，得到QK矩阵（为了防止数值过大，往往要除以 $\sqrt{d}$,d为Q，K，V的列数），此时得到的矩阵，每一行代表一个单元，而每一列则代表对应单元对本单元的影响程度。



经过softmax（归一化，即每一行最后的和都为1）后，与V相乘，即可得到最后的输出Z



对于多头注意力机制，则可以得到多个输出Z，将其链接并线性变换，即可得到与输入相同大小的矩阵作为最终输出