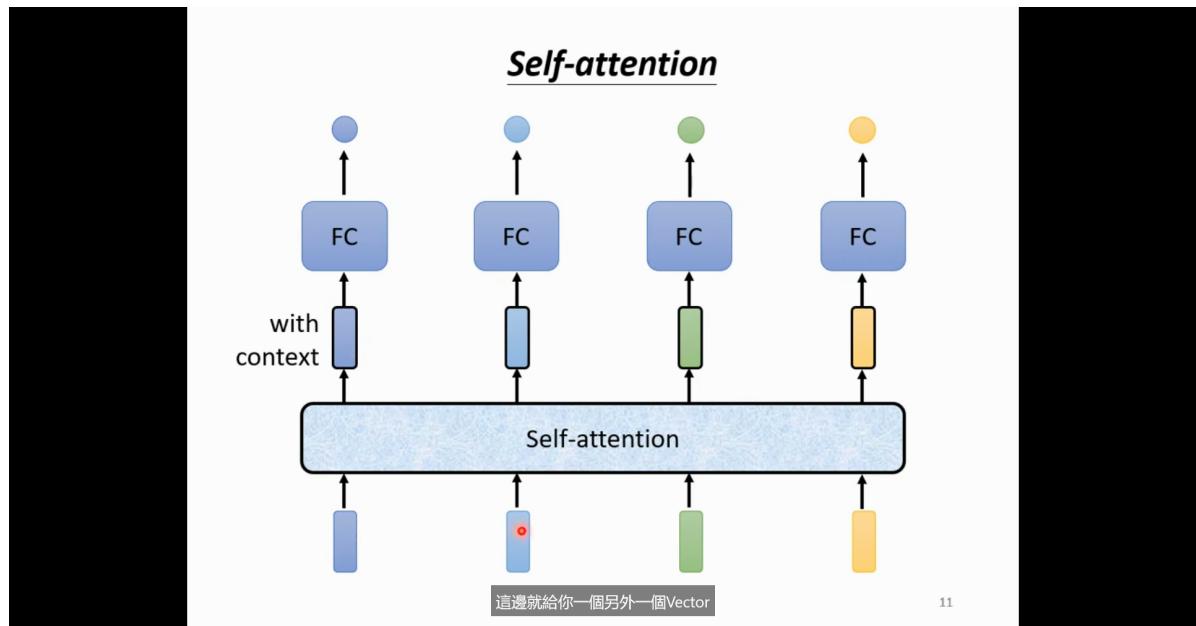


输入几个向量输出几个label

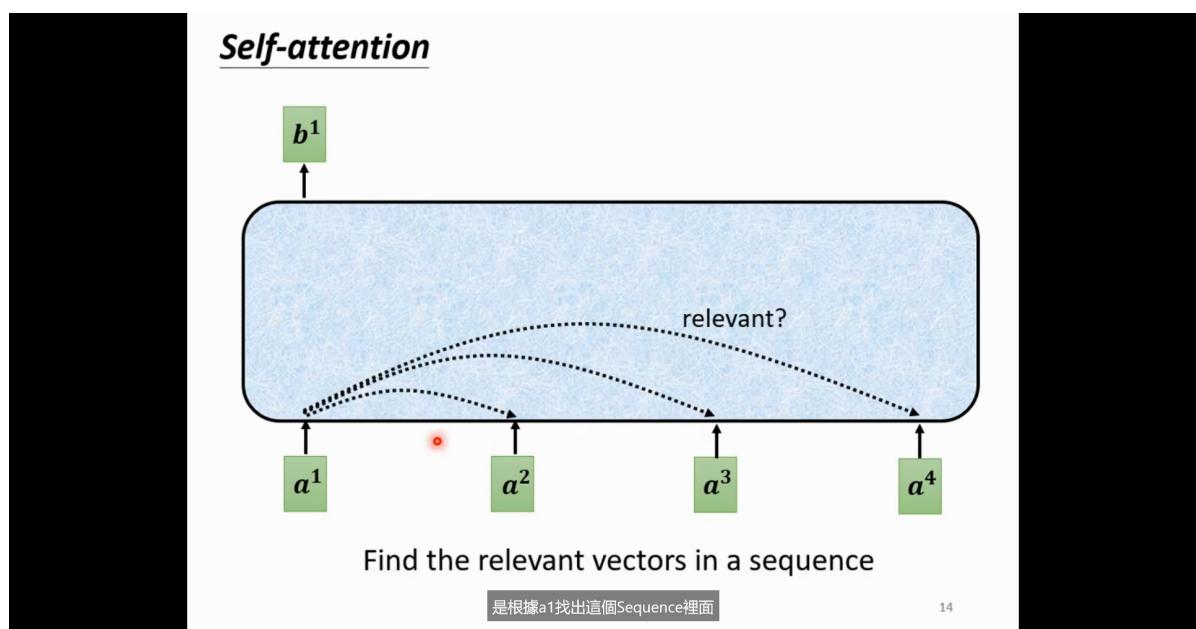


输出的4个都考虑了整个输入序列

可以叠好多层

怎样计算 $b_1$ 呢?

## Self-Attention



1. 根据 $a^1$ 找到输入序列中跟 $a^1$ 相关的向量, 每个其他的输入跟 $a^1$ 相关的程度用 $\alpha$ 来表示

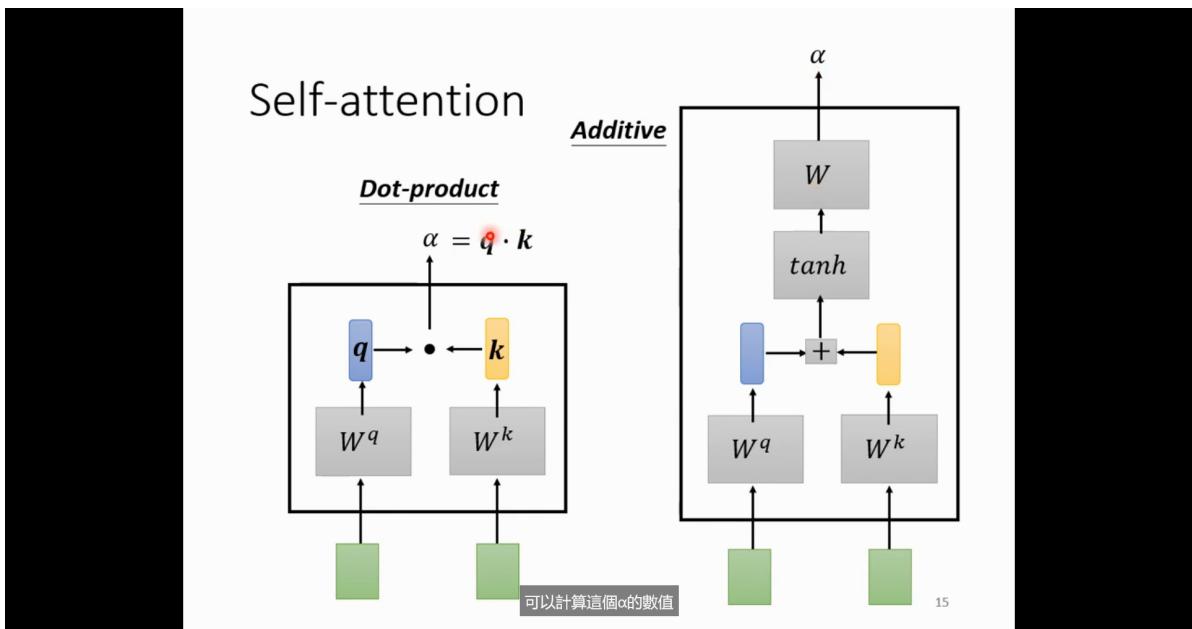
下图为计算权重系数的方法

两个输入向量 分别乘权重矩阵, 得到 $q, k$

$$q = a * W^q$$

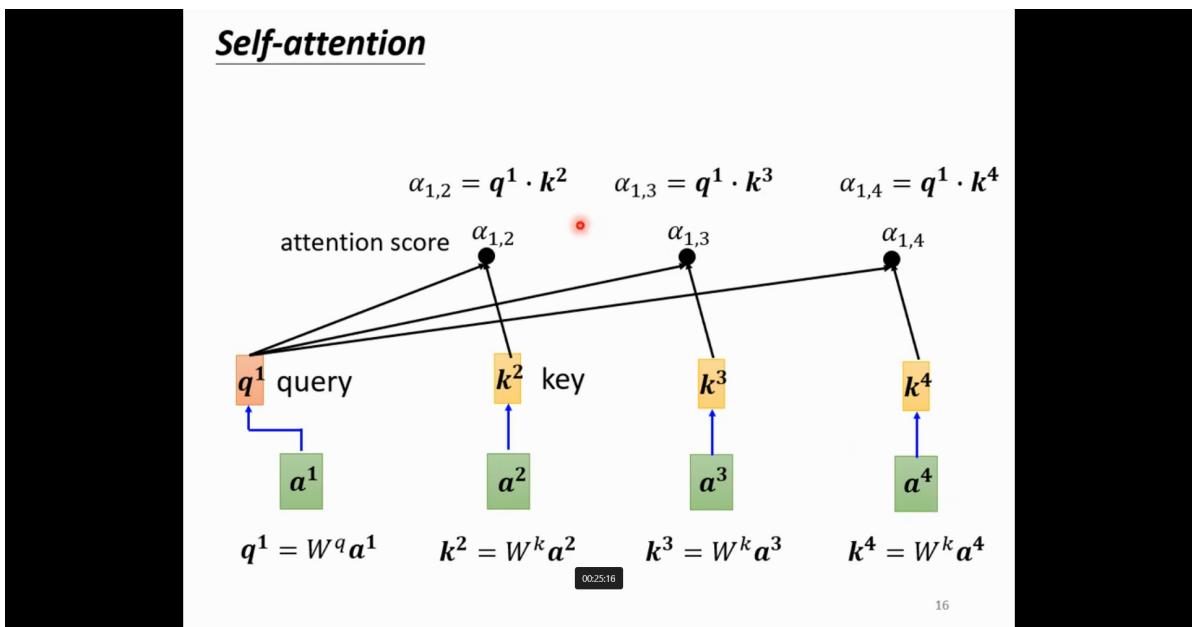
$$v = a * W^v$$

$q, k$ 再内积则为权重系数



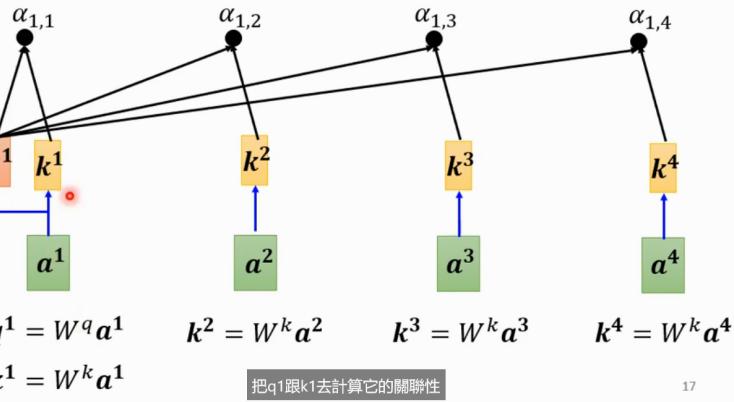
$\alpha_{1,2}$ 表示 $q$ 是1提供的 $k$ 是2提供的

用 $a_1$ 分别去和 $a_2, a_3, a_4$ 内积, 算出3个注意力系数



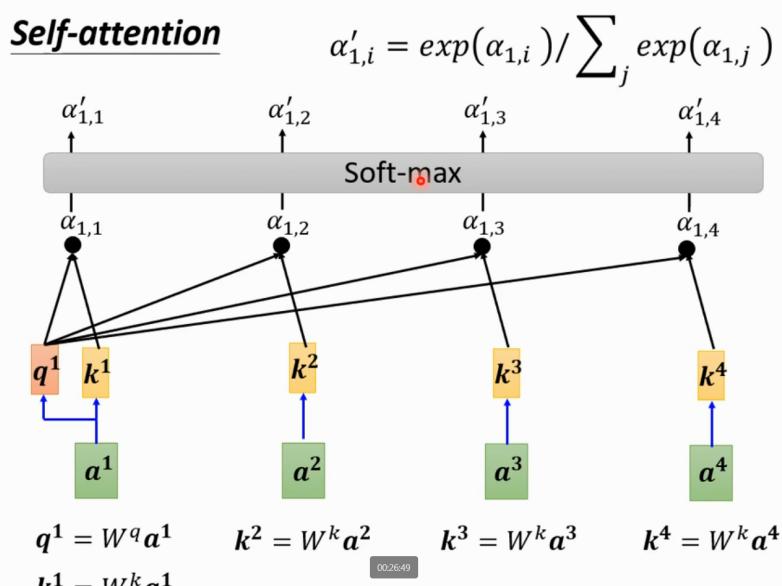
一般情况, $a_1$ 也要和自己算一个相关性

## Self-attention



17

经过softmax得到 $\alpha'_{1,2}$

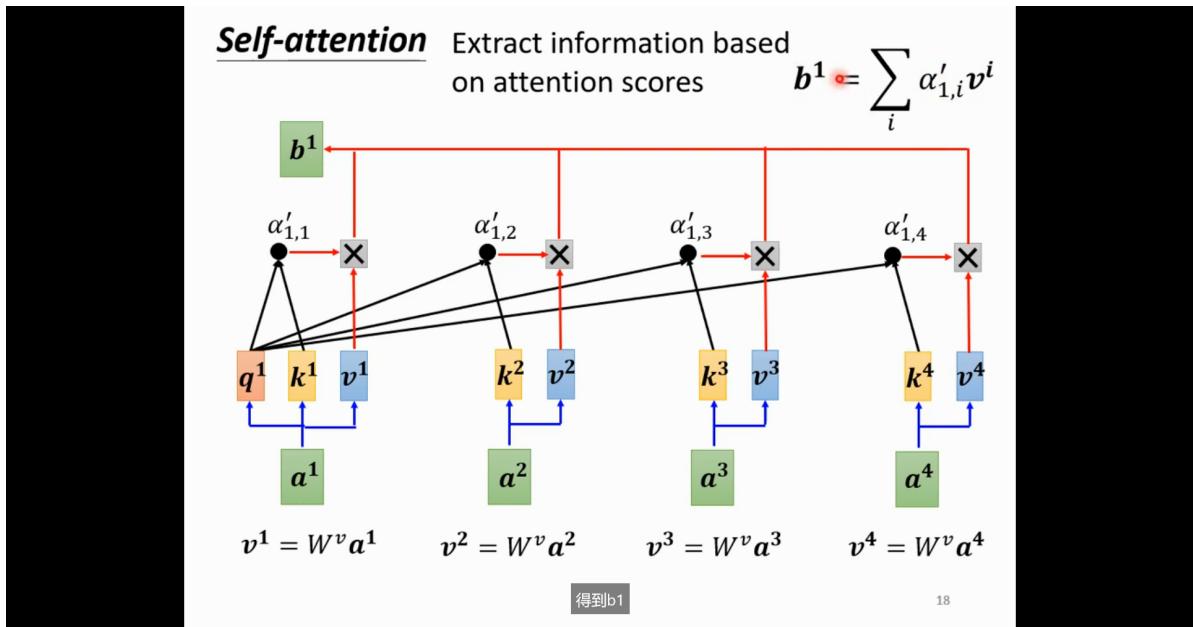


17

得到 $\alpha'_{1,2}$ 后会根据权重抽取信息

把输入乘上 $W^v$ 后得到 $v$

$v$ 再乘上 $\alpha'$ 后,所有的相加即为b1



$b_1$  到  $b_4$  是同时被计算出来的

## 矩阵乘法

可以把  $a_1, a_2, a_3, a_4$  拼起来(分别为列)变成一个矩阵  $W$  相乘

$$q^i = W^q a^i \quad \begin{matrix} q^1 & q^2 & q^3 & q^4 \end{matrix} = \begin{matrix} W^q \\ Q \end{matrix} \begin{matrix} a^1 & a^2 & a^3 & a^4 \end{matrix} \quad I$$

$$k^i = W^k a^i \quad \begin{matrix} k^1 & k^2 & k^3 & k^4 \end{matrix} = \begin{matrix} W^k \\ K \end{matrix} \begin{matrix} a^1 & a^2 & a^3 & a^4 \end{matrix} \quad I$$

$$v^i = W^v a^i \quad \begin{matrix} v^1 & v^2 & v^3 & v^4 \end{matrix} = \begin{matrix} W^v \\ V \end{matrix} \begin{matrix} a^1 & a^2 & a^3 & a^4 \end{matrix} \quad I$$

得到的  $k$  都变成行向量和  $q$  相乘分别得到  $\alpha$

|                 |                 |                 |                 |                 |                 |                 |                 |                 |                 |                 |                 |                 |                 |                 |                 |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| $\alpha'_{1,1}$ | $\alpha'_{2,1}$ | $\alpha'_{3,1}$ | $\alpha'_{4,1}$ | $\alpha'_{1,2}$ | $\alpha'_{2,2}$ | $\alpha'_{3,2}$ | $\alpha'_{4,2}$ | $\alpha'_{1,3}$ | $\alpha'_{2,3}$ | $\alpha'_{3,3}$ | $\alpha'_{4,3}$ | $\alpha'_{1,4}$ | $\alpha'_{2,4}$ | $\alpha'_{3,4}$ | $\alpha'_{4,4}$ |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|

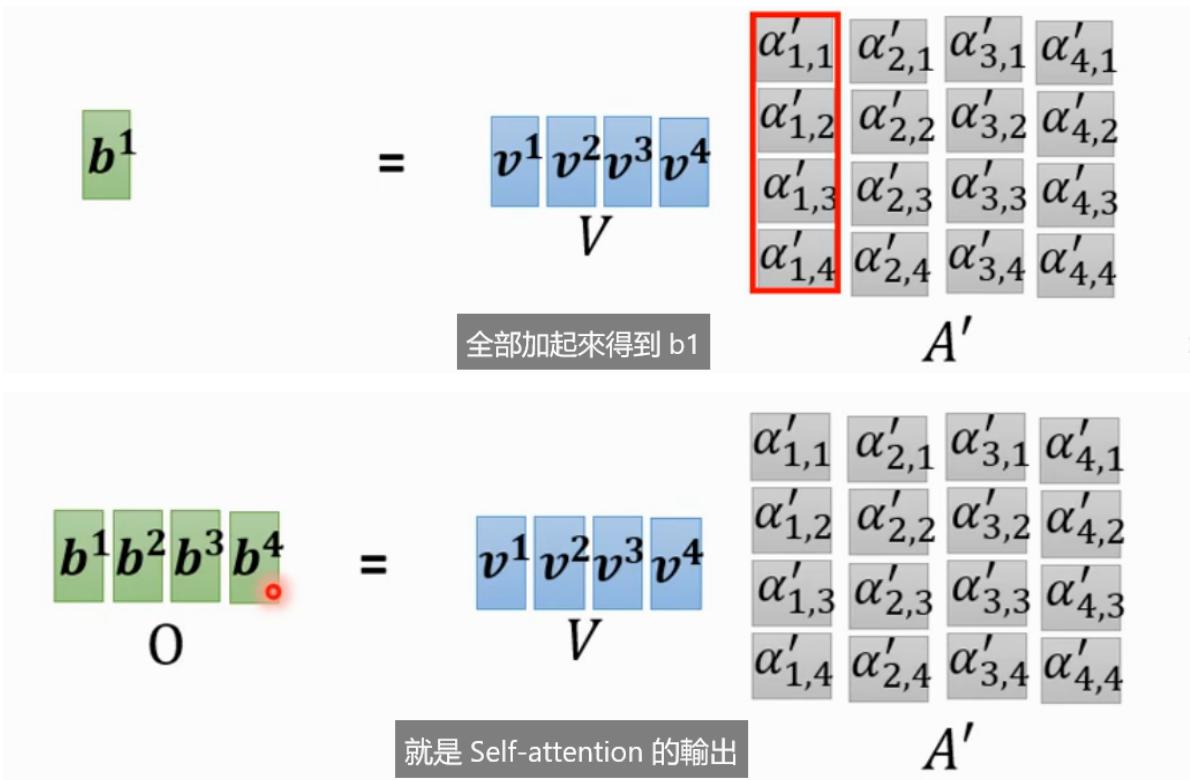
$A'$

softmax 你會對這邊的每一個 column

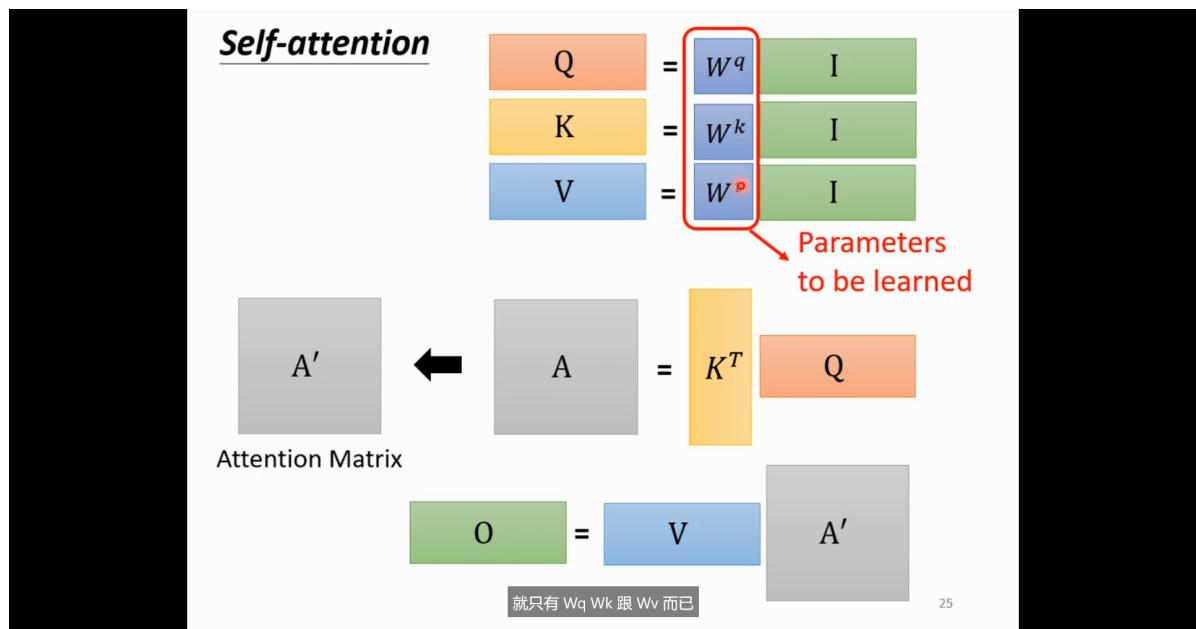
$= \begin{matrix} k^1 \\ k^2 \\ k^3 \\ k^4 \end{matrix} \begin{matrix} q^1 & q^2 & q^3 & q^4 \end{matrix} \quad Q \quad K^T$

每列做 SoftMax

再乘上  $v$



综合一下,只要学习  $W$



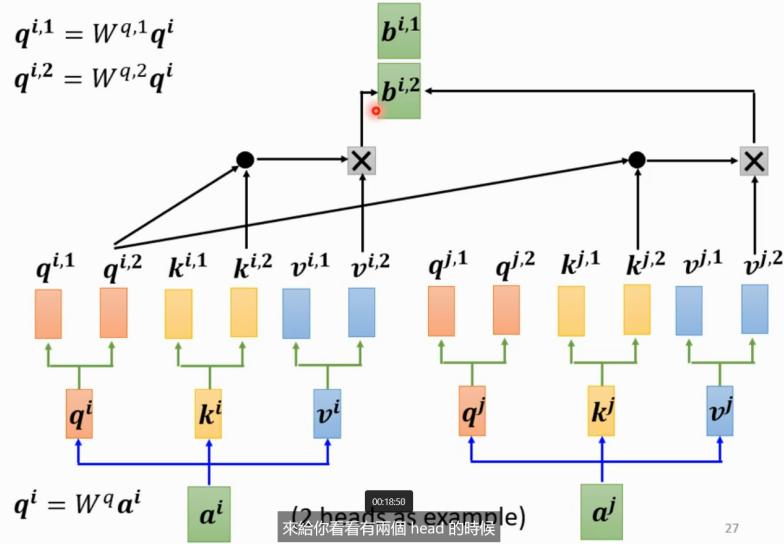
## Multi-head Self-attention

1的那一类自己做attention

2的那一类自己做attention

下图为2head,就是一个q,k,v分别乘两个举证变成了两类q,k,v

## Multi-head Self-attention Different types of relevance



## Position Encoder

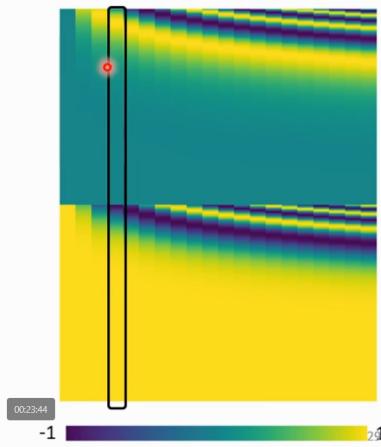
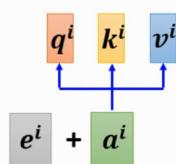
Self-attention是缺少位置信息,每个都乘了其余的信息

为每个输入都设置一个positional vector  $e^i$  来代表位置信息,并加到  $a^i$  上

## Positional Encoding

Each column represents a positional vector  $e^i$

- No position information in self-attention.
- Each position has a unique positional vector  $e^i$
- hand-crafted
- learned from data

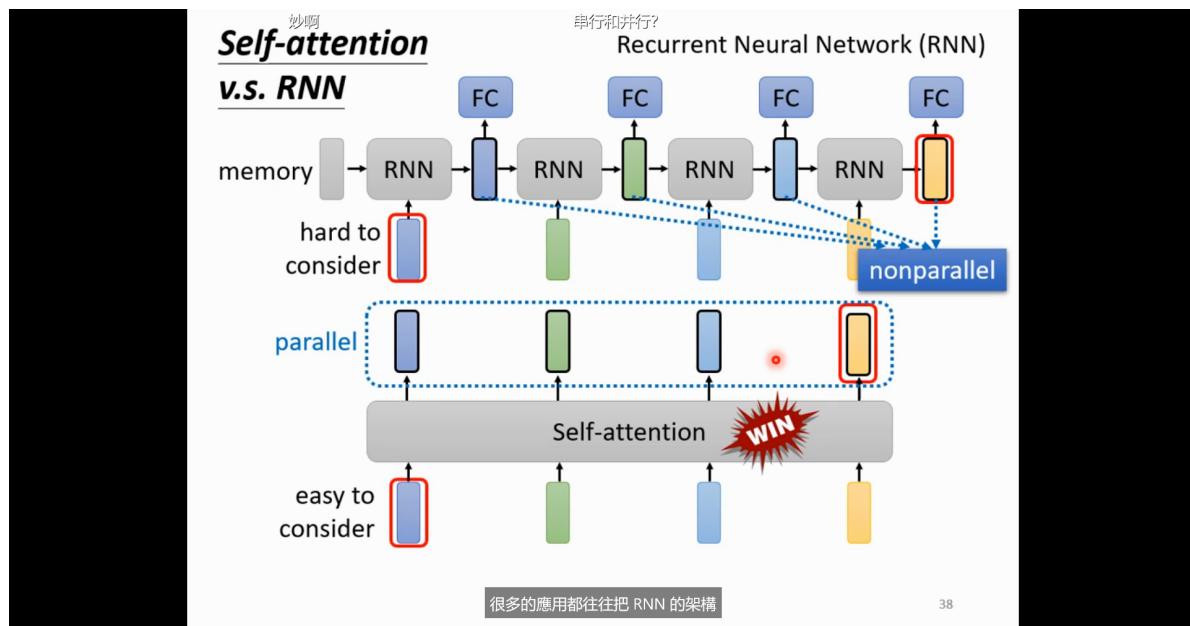


右边每一列表示每个输入的位置信息

## Tips

Self-Attention也可以不看整个句子,可能一个小的范围就可以

## RNN和Self-Attention区别



## Self-Attention与Graph

