



第三部分： **self-attention**

自注意力机制

常见的输入。



李哥考研



图片

从上海到北京，
买家没有卖家精。

文字



声音

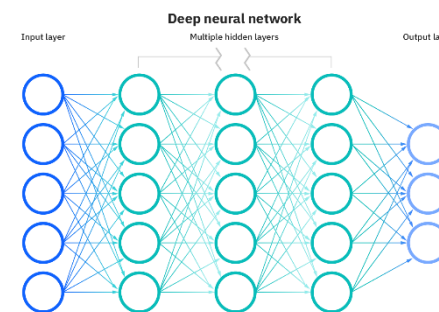
怎么用数据表示文字。



李哥考研

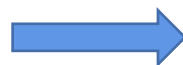


$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1j} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2j} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3j} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & a_{i3} & \dots & a_{ij} & \dots & a_{in} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mj} & \dots & a_{mn} \end{bmatrix}$$

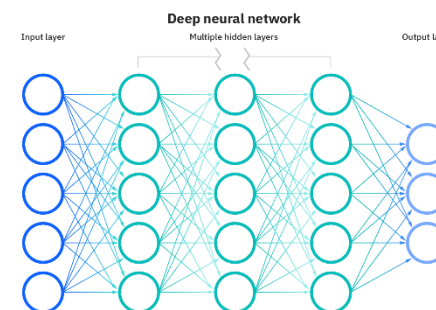
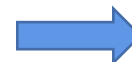


知道图片的样子

“我”



?



知道输入的是“我”字

编码为向量

One-hot Encoding

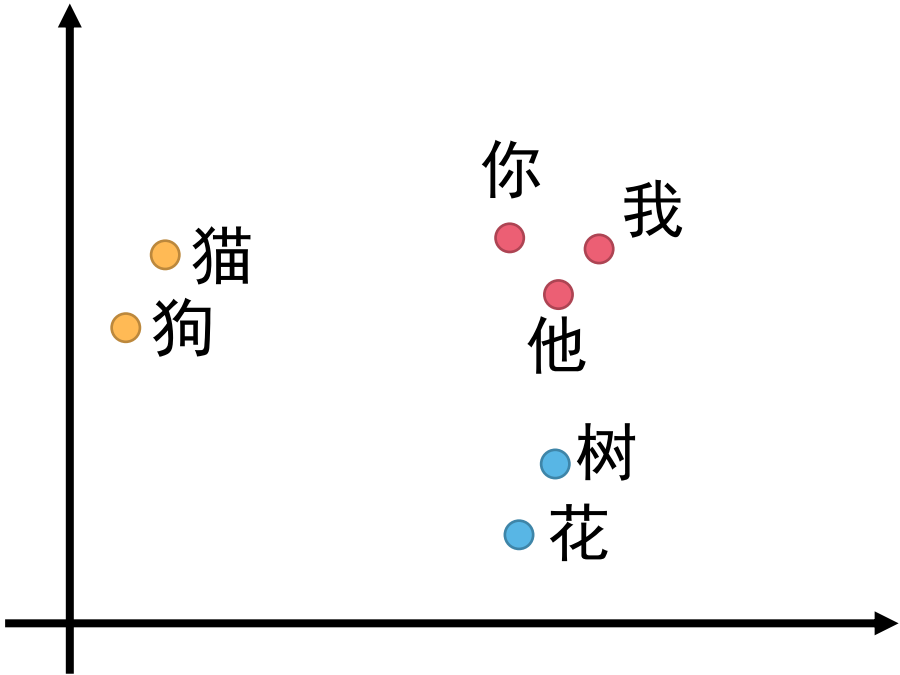
- 1, 维度太长。
- 2, 体现不出关系

汉字个数

我 = [1 0 0 0 0 0]
你 = [0 1 0 0 0 0]
他 = [0 0 1 0 0 0]
猫 = [0 0 0 1 0]
狗 = [0 0 0 0 1]
⋮

我 要 上 岸
768

Word Embedding



让模型自己学



李哥考研

One-hot Encoding

汉字个数

我 = [1 0 0 0 0]

你 = [0 1 0 0 0]

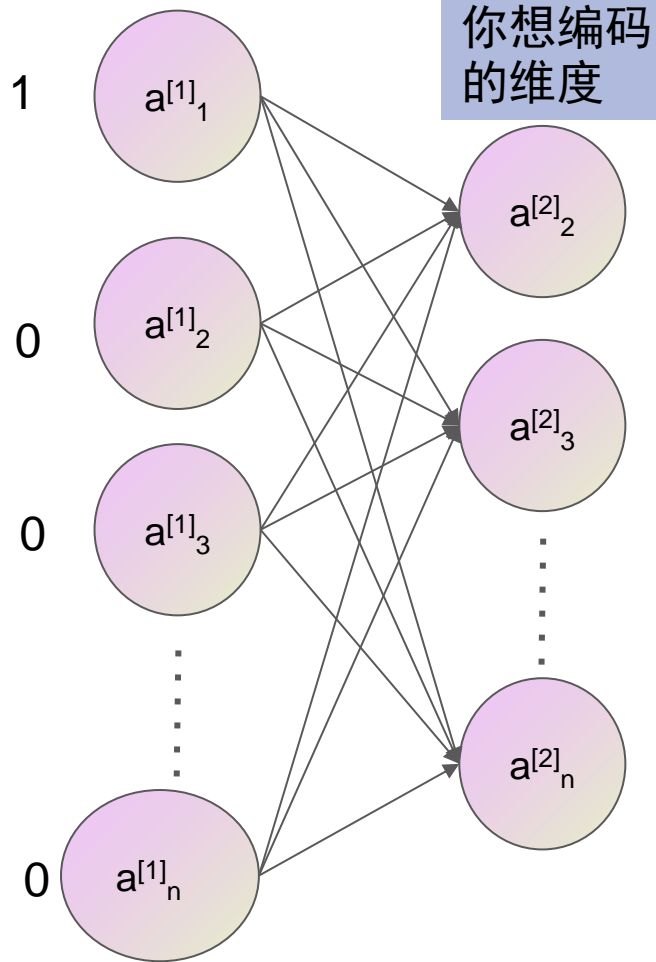
他 = [0 0 1 0 0]

猫 = [0 0 0 1 0]

狗 = [0 0 0 0 1]

⋮

汉字个数



词embedding

我们可以
得到文字
对应的输入

Linear (21128, 768)

常见的输入。



李哥考研

- 每个词对应一个编码。

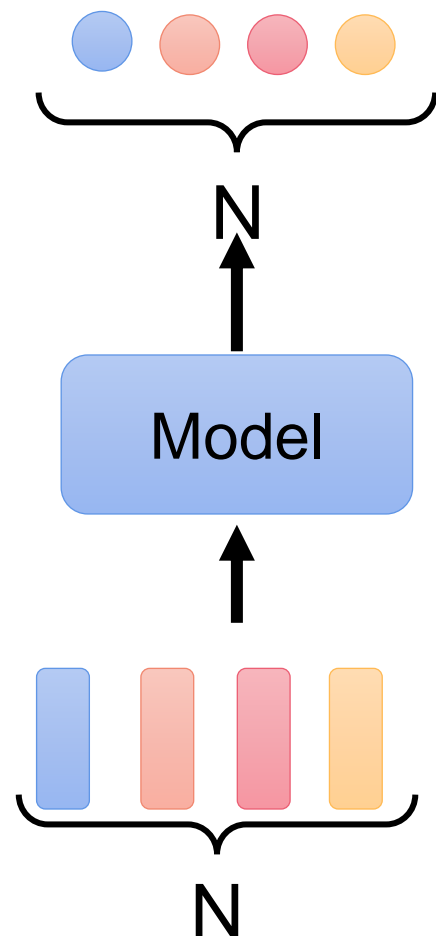
我今天要上天天向上。 ➡ 21111, 223, 2324, 231, 312, 2324, 2324, 531, 12234, 10

常见的输出。



李哥考研

- 每个词都有输出一个值.



Example Applications

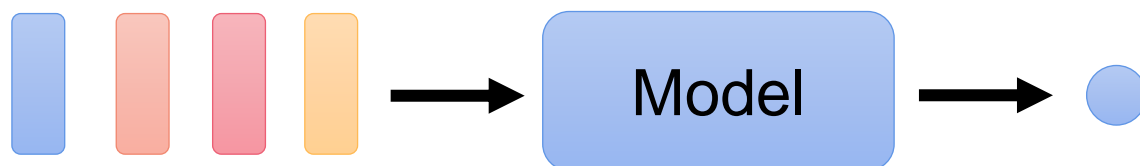
我	爱	你
↓	↓	↓
代词	动词	代词

常见的输出。



李哥考研

- 所有词输出一个值.



Example Applications

你做的不错

情感分类

正向

唉，又又又emo了

情感分类

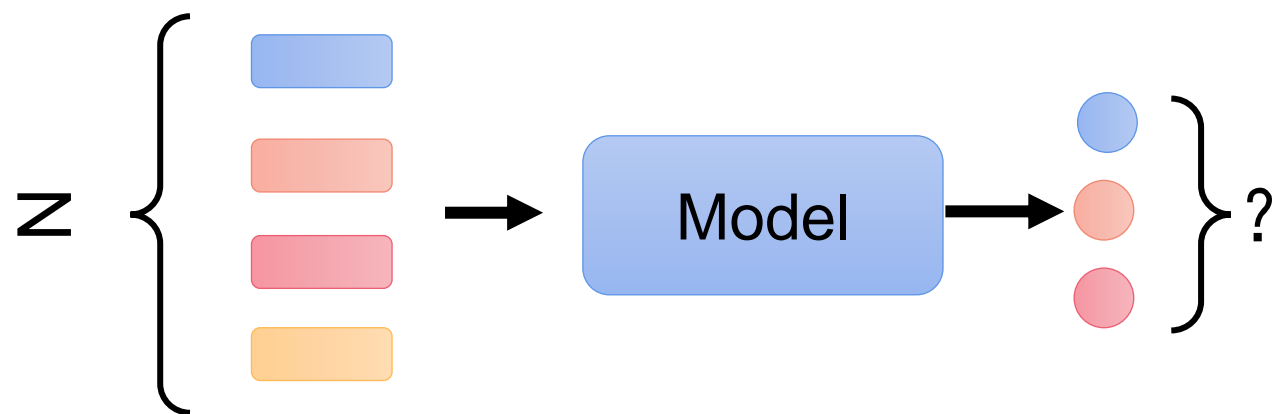
负向

常见输出。



李哥考研

- 输入输出长度不对应。



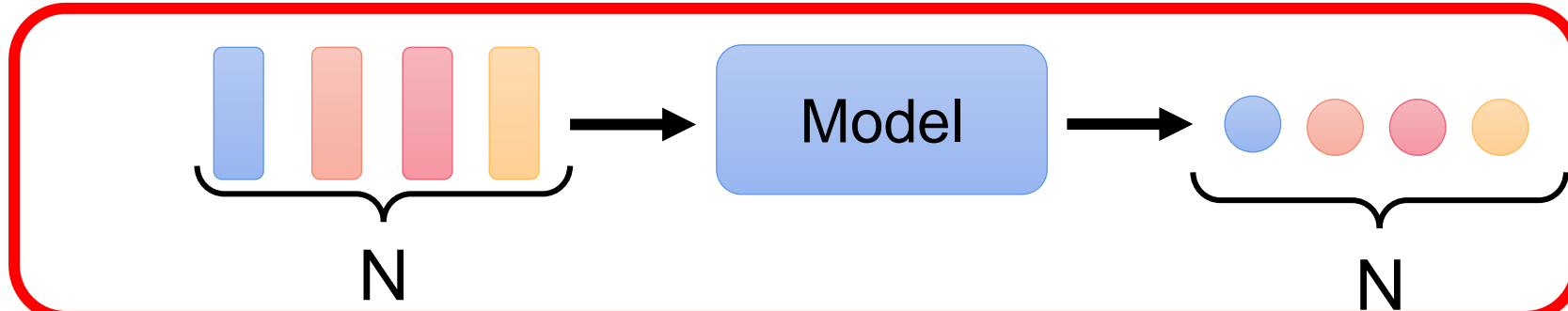
我爱中国 翻译 I love China

What is the output?



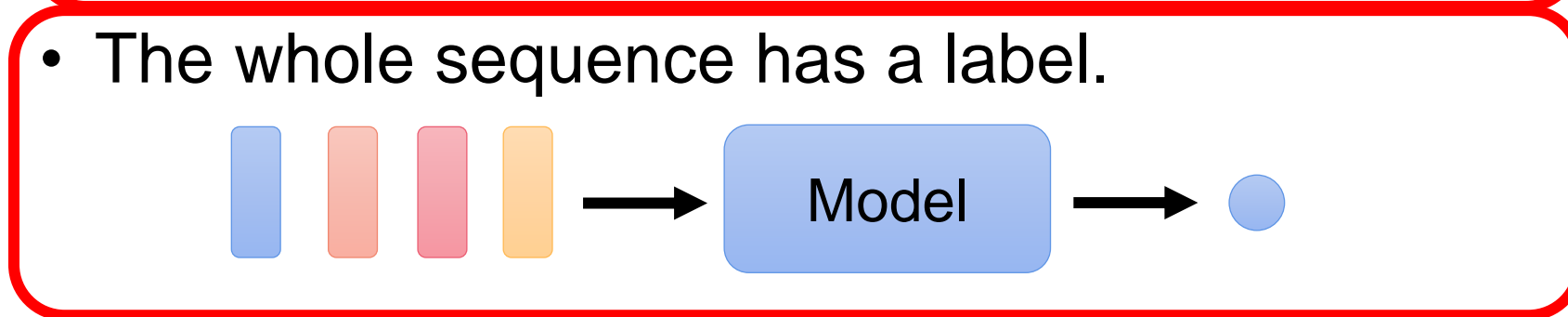
李哥 考研

- Each vector has a label.

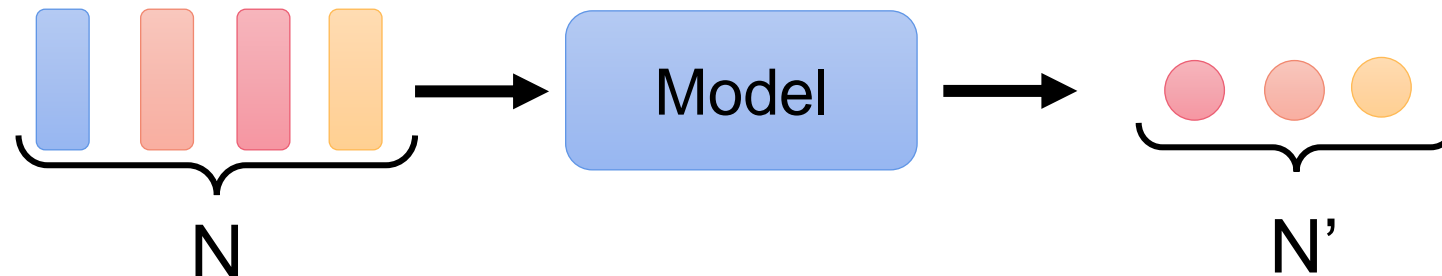


这节课

- The whole sequence has a label.



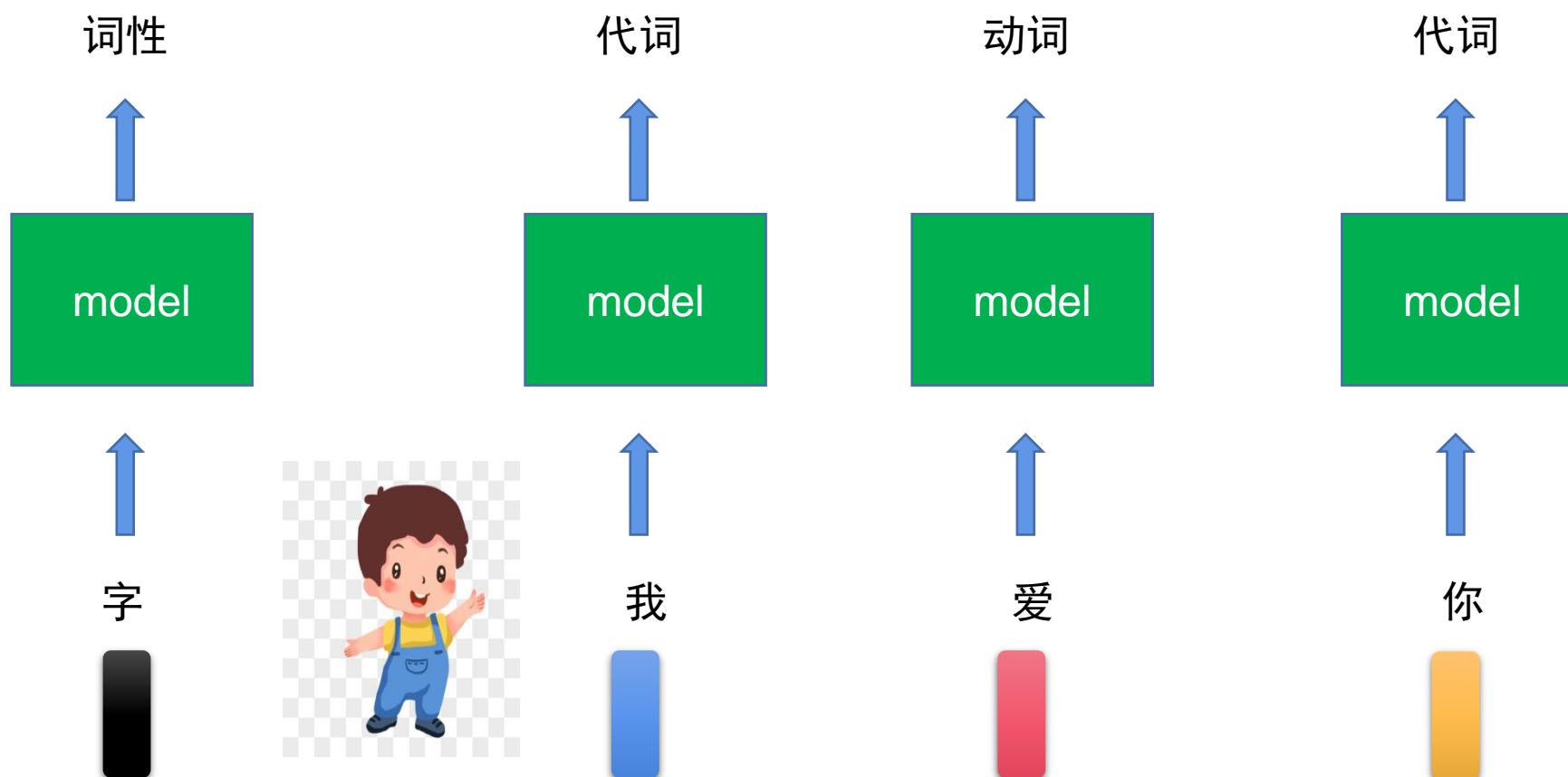
- Model decides the number of labels itself.



假如有个词性识别任务。



李哥考研



问题是。



李哥考研



名词

动词?

model

model

model

model

model

model

model

这

爱

爱

得

够

深

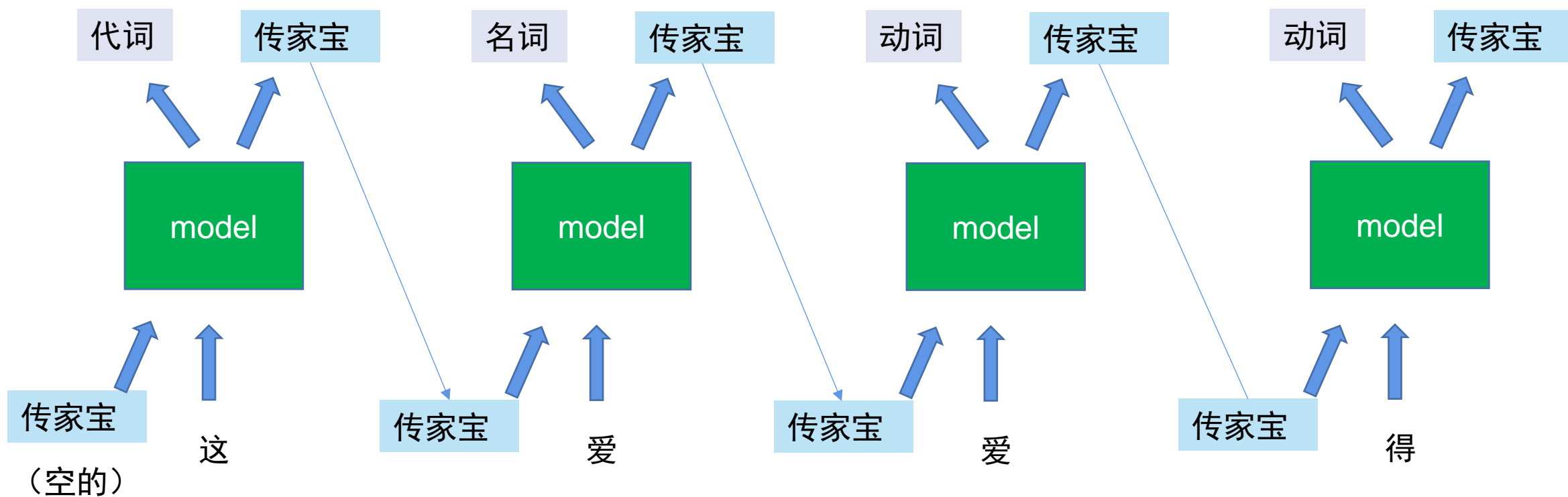
?

Can you can a can as a cancer can a can?

所以要考虑前后关系。



李哥考研



传家宝： 向量, 记忆单元

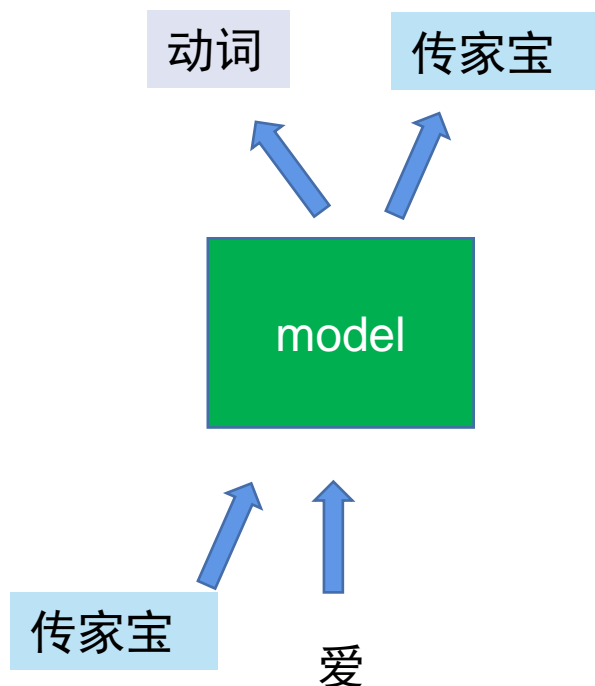
RNN: [Recurrent neural network](#)

太远了，够不到。



李哥考研

我漫步来到赛汗塔拉城中草原，站在草原深处，牧草及膝，草香花香阵阵袭击鼻腔，眼睛对大自然的饥渴就这么突然被填满，这里的草原也恍惚就是那时的草原，虽青春已不再，可草原仍在，那些青春的记忆也仿佛一下子堆满了脑海——二十年前的今天，青春飞扬、意气风发，军校毕业来到包头，从此便与赛汗塔拉结下了不解之缘，闲暇之余总会来到这里，放飞心情，驱散迷茫，感受草原胸怀。



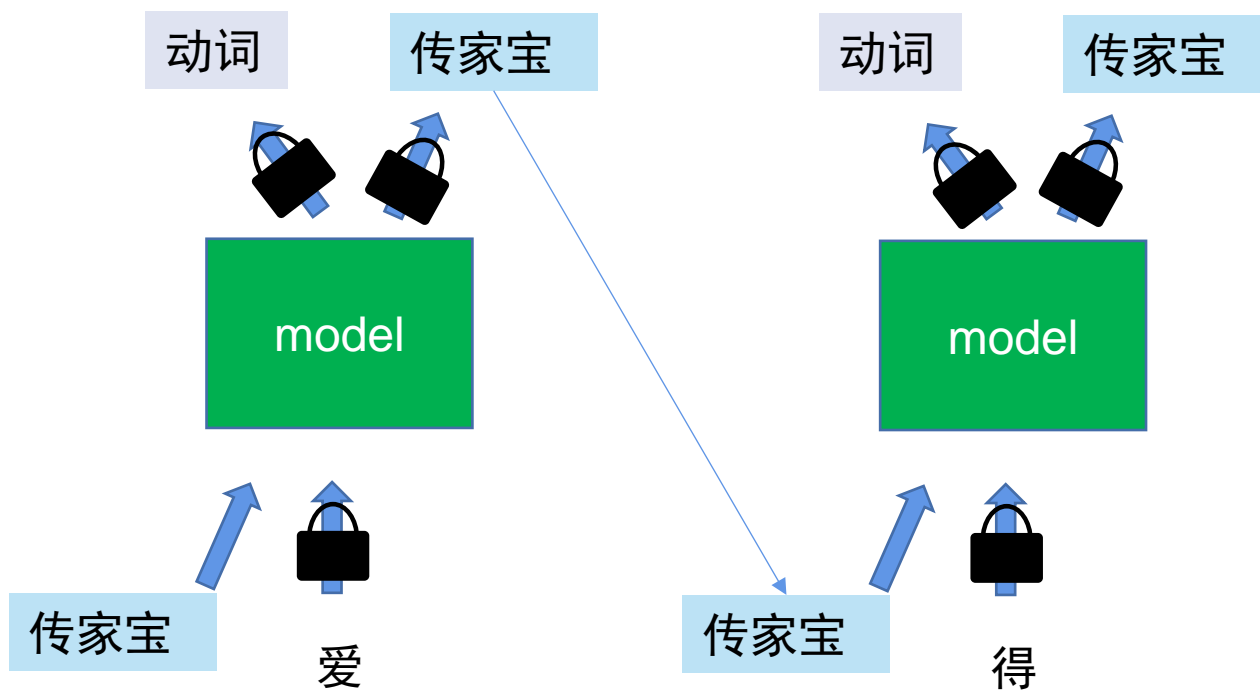
问：谁与赛罕塔拉结下了不解之缘？

败家子孙太多





长短期记忆(Long short-term memory, **LSTM**)



输入门

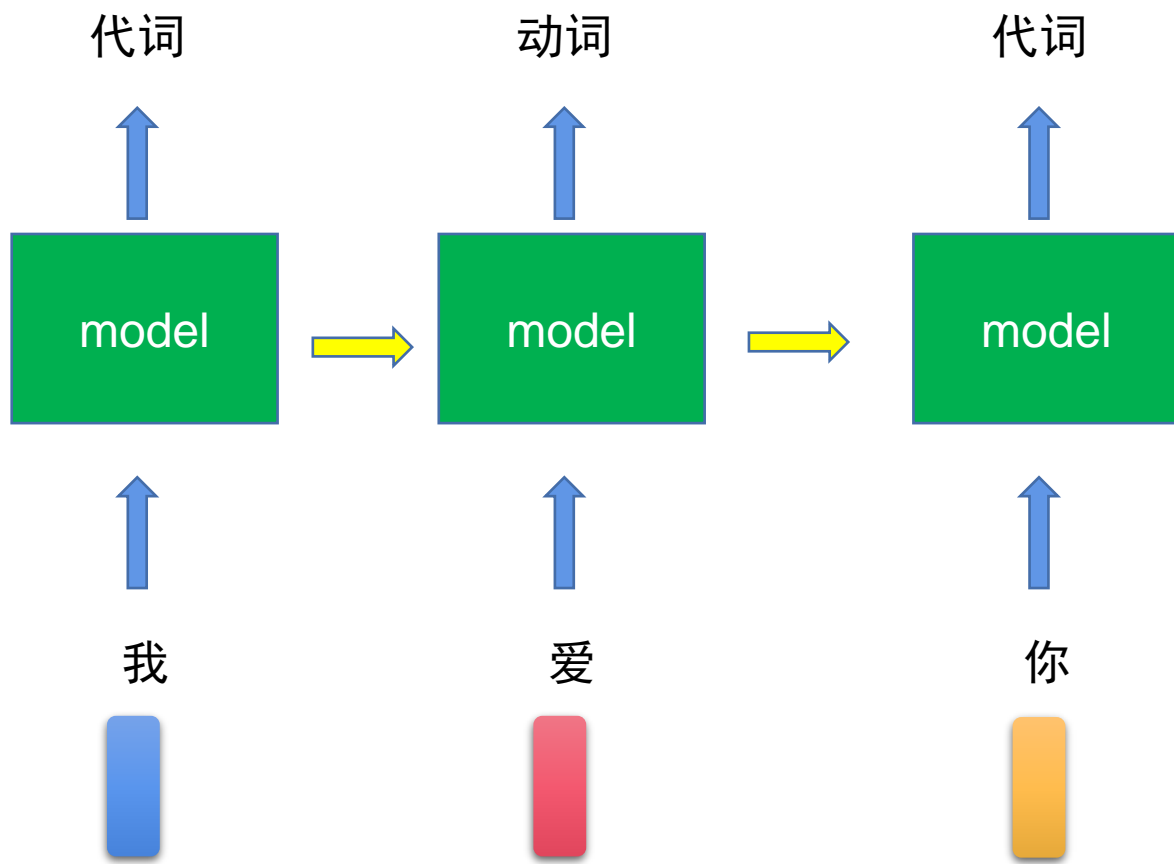
遗忘门

记忆门

RNN和LSTM太慢了。



李哥 考研



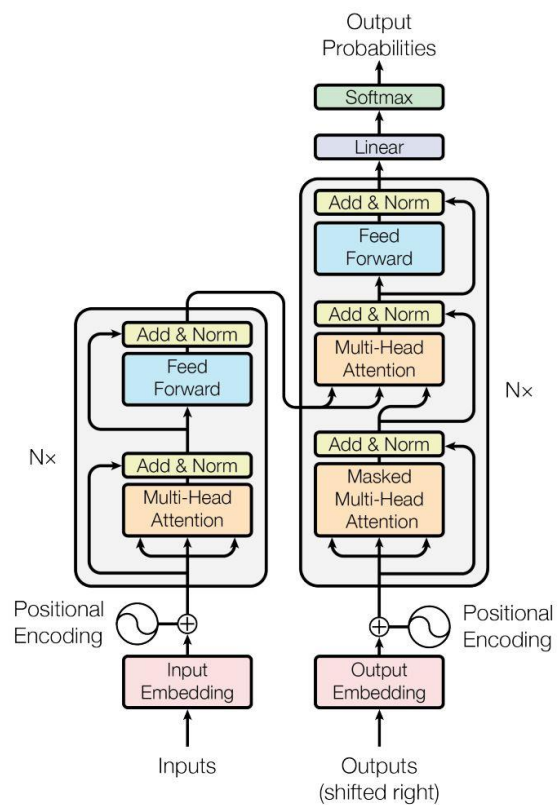
只能一个一个来，
一代接一代，

我们有没有办法，
一下子看完整篇文章，
输出结果。

Self-attention: 自注意力机制



李哥考研



《Attention is All You Need》



《Money is All You Need》

Self-attention

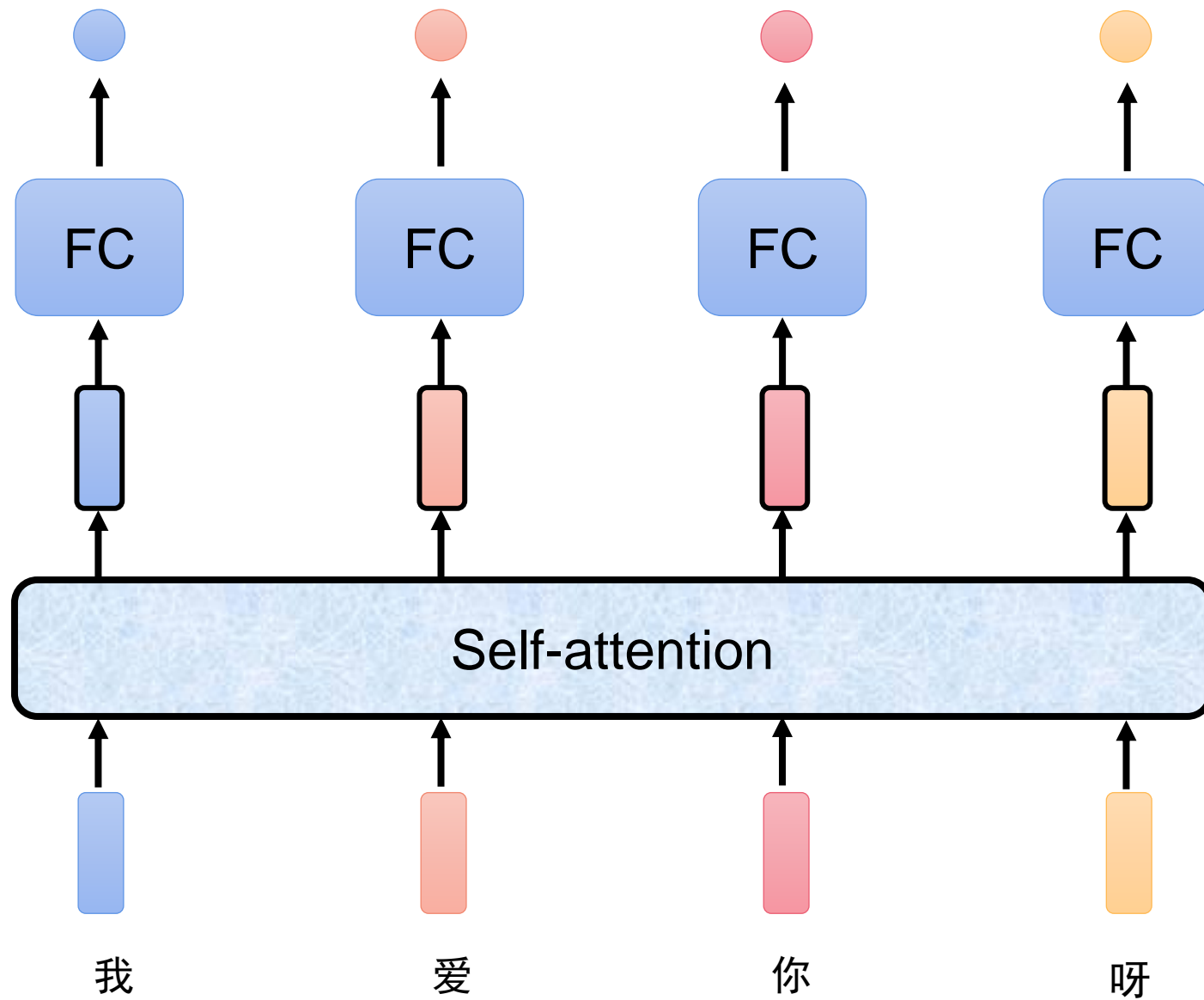


李哥考研

考虑了整个
句子后的特征

Self-attention

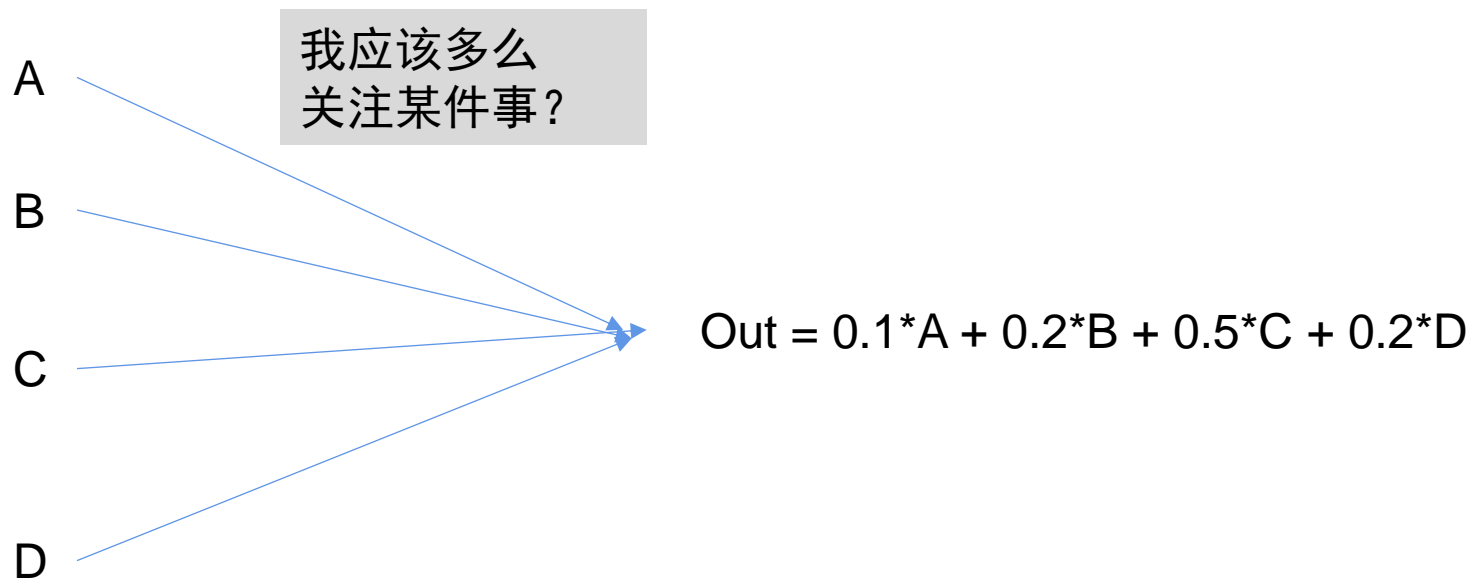
是一个特征转
换器



什么是注意力？



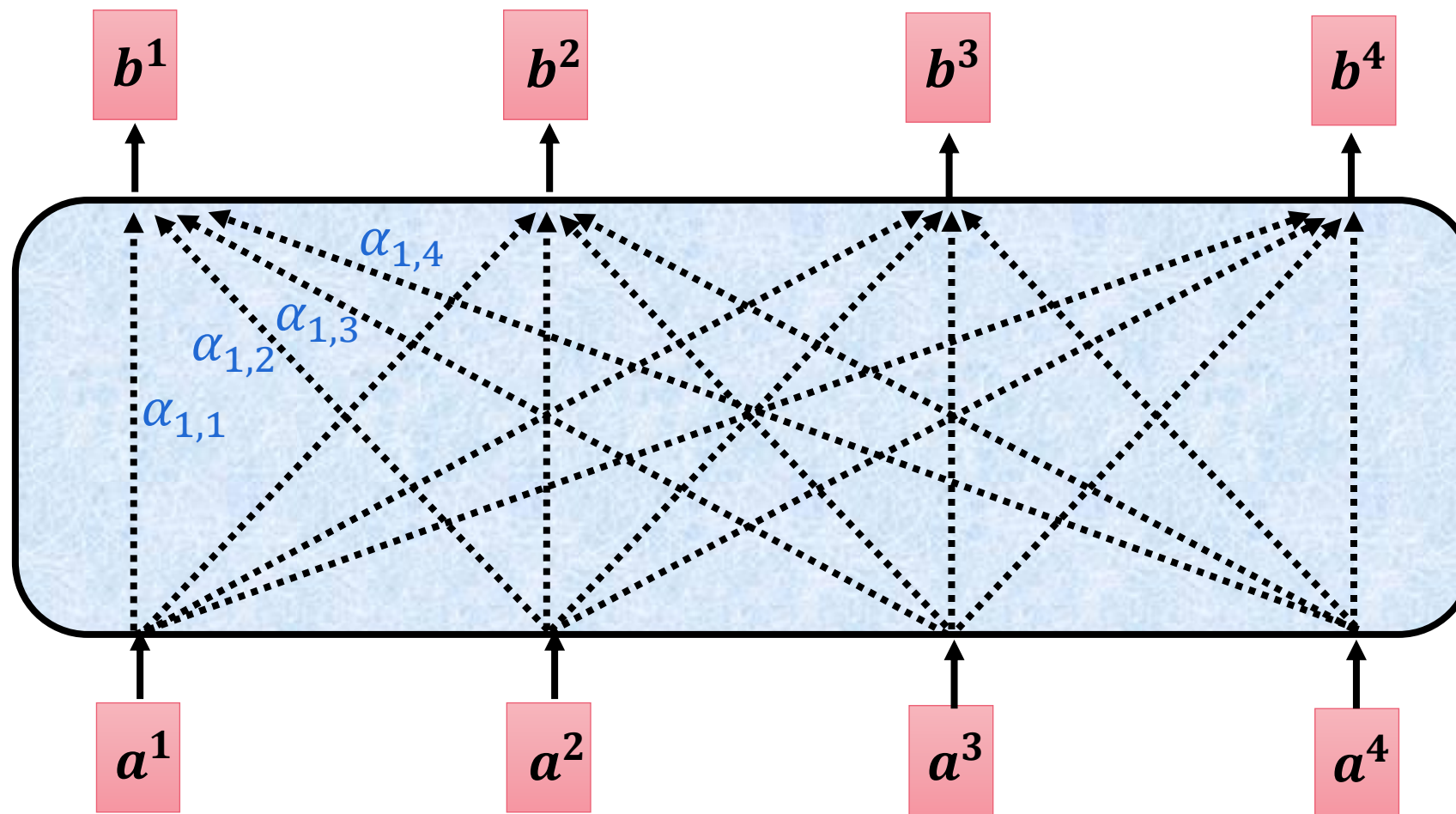
李哥考研



Self-attention



李哥考研



$\alpha_{1,1}, \alpha_{1,2}, \alpha_{1,3}, \alpha_{1,4} = (0.1, 0.3, 0.2, 0.4)$ 表示了, a^1 对其他 输入向量的注意力。其他一样

如何计算注意力。

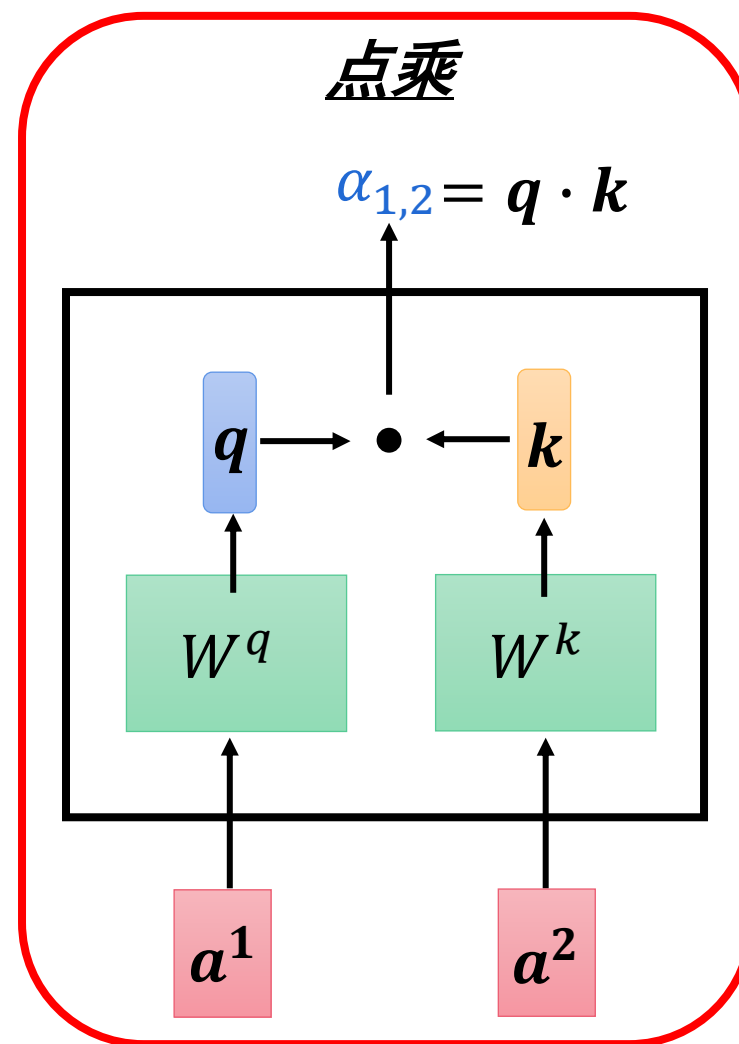
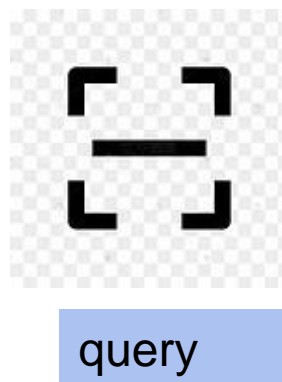


李哥考研

第一种想法

$$\alpha_{1,2} = a^1 a^2$$

这样固定的两个字乘出来永远是一个值，不好。

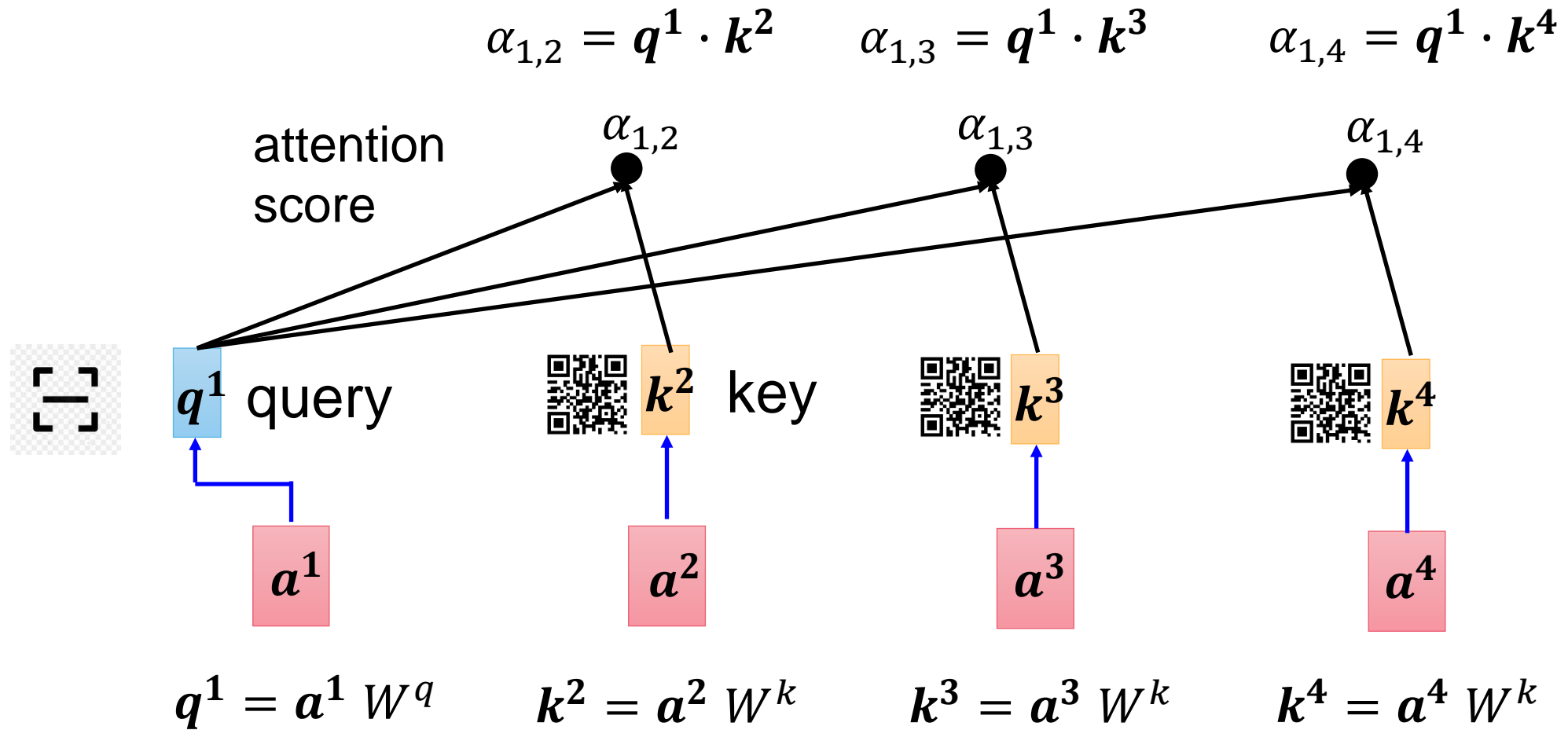


key

Self-attention



李哥考研

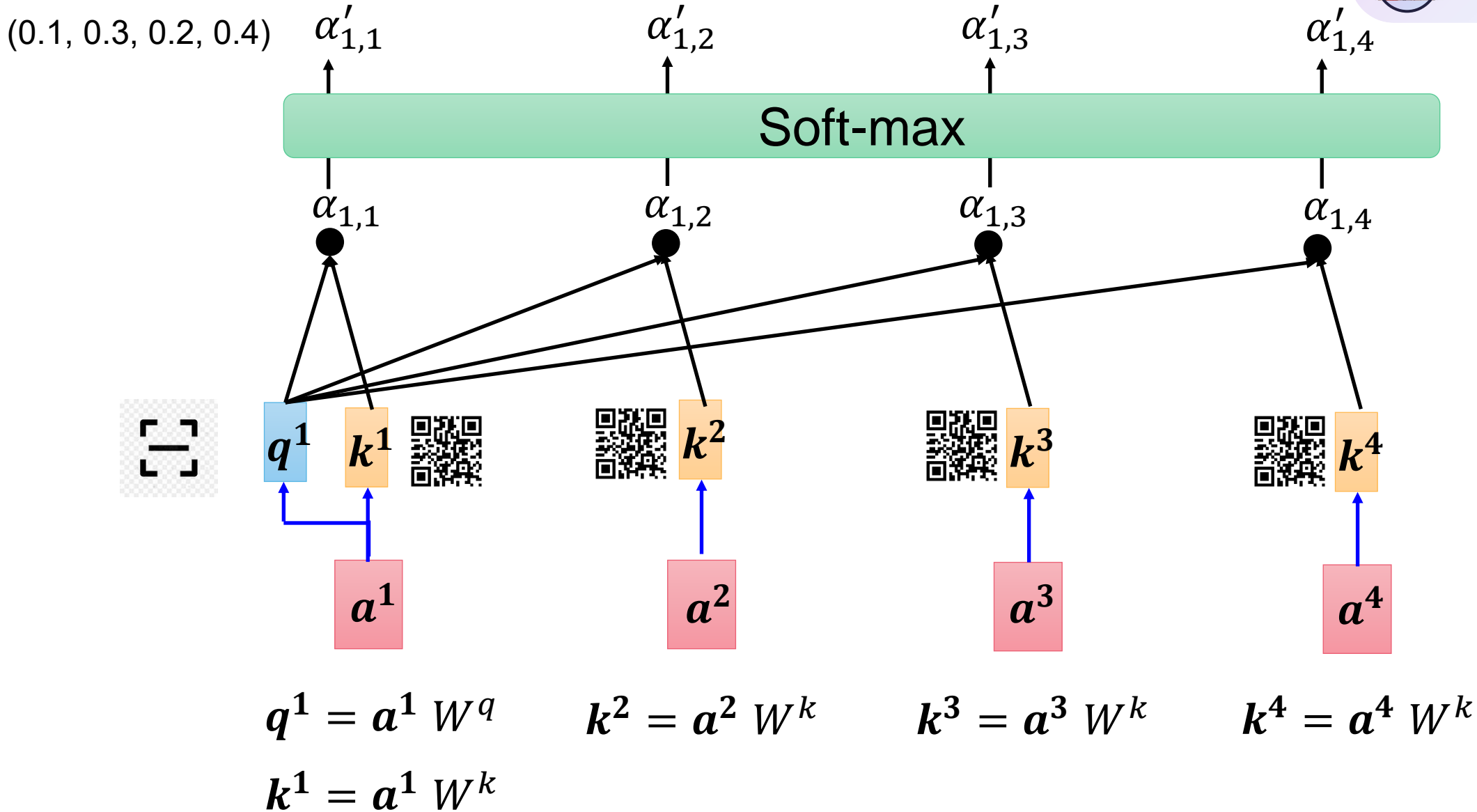


Self-attention

$$\alpha'_{1,i} = \exp(\alpha_{1,i}) / \sum_j \exp(\alpha_{1,j})$$



李哥考研



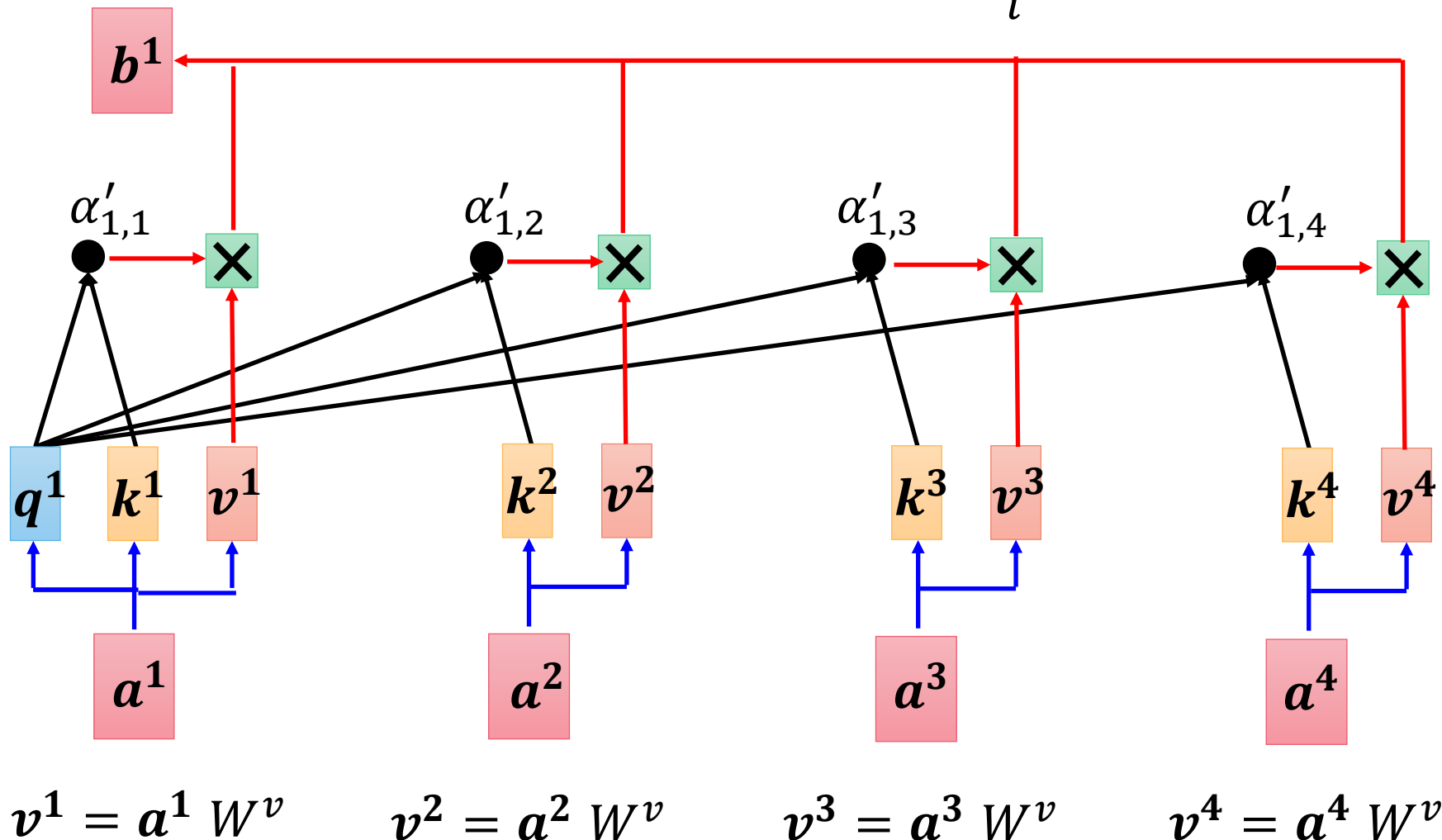
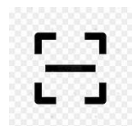
Self-attention

$$b^1 = \sum_i \alpha'_{1,i} v^i$$



李哥考研

我们凭注意力去把谁加起来？

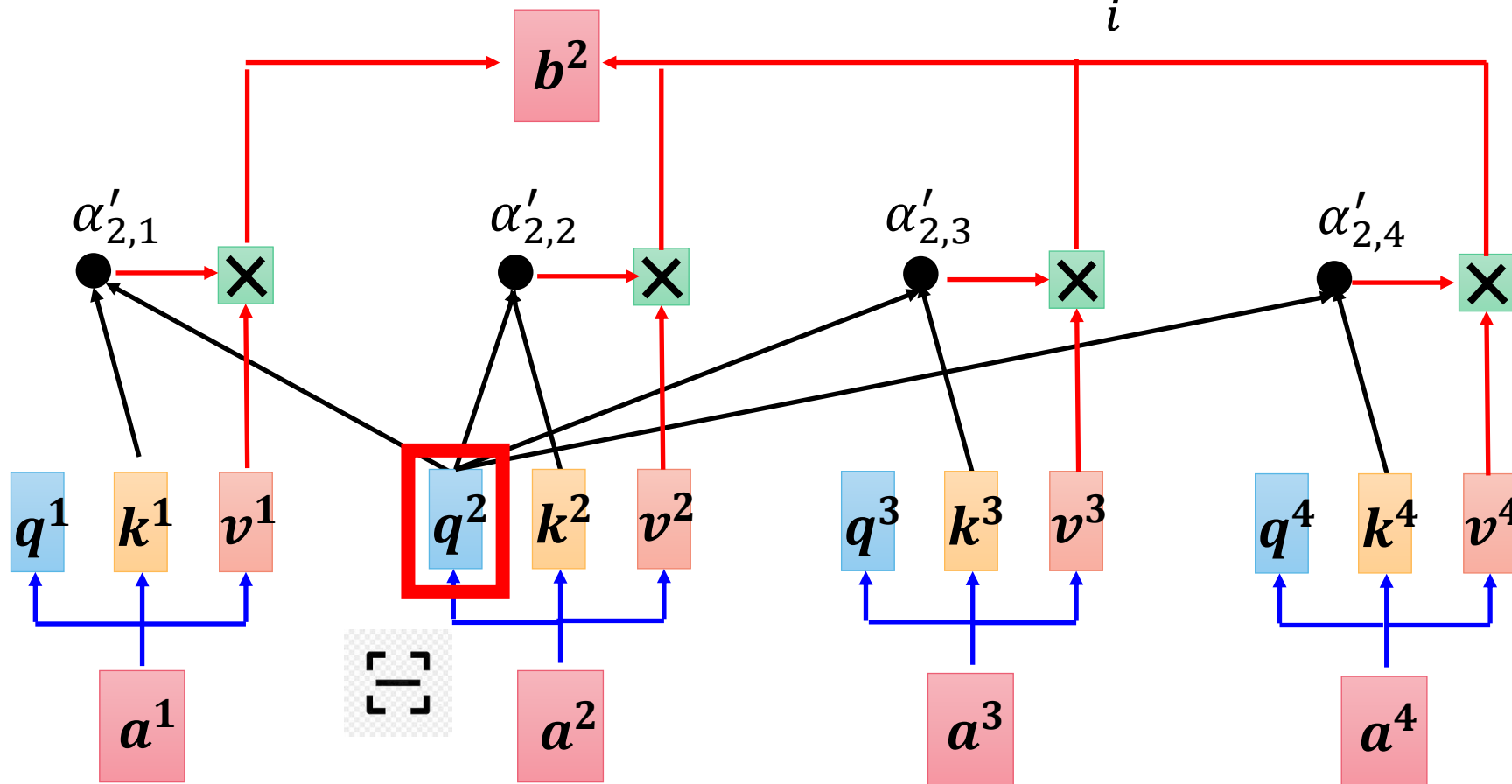


Self-attention

$$b^2 = \sum_i \alpha'_{2,i} v^i$$



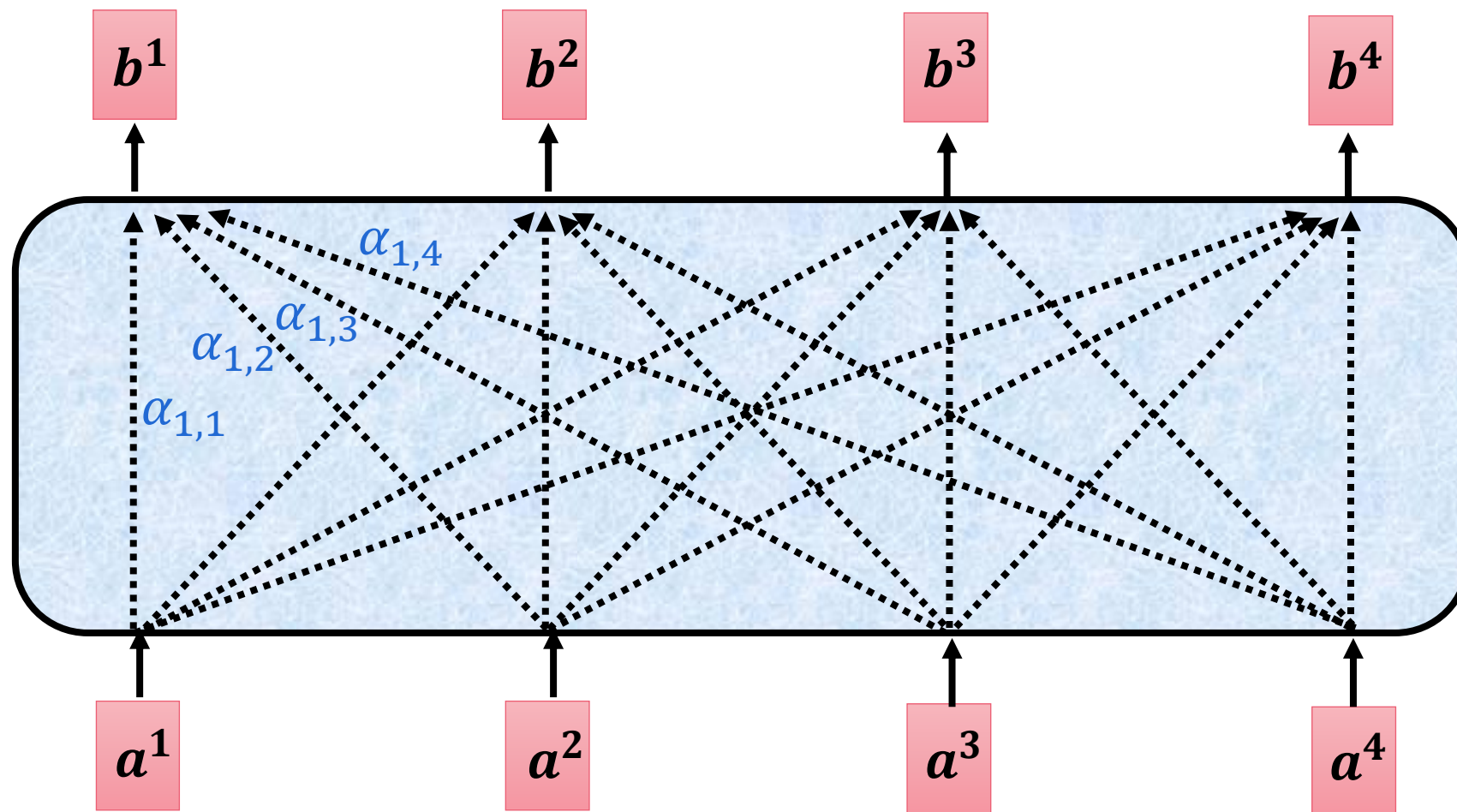
李哥考研



Self-attention



李哥考研



每一个字的特征提取，都可以同时进行。

因为它不需要等待前一个完成。

完全并行！

$$\alpha_{1,1}, \alpha_{1,2}, \alpha_{1,3}, \alpha_{1,4} = (0.1, 0.3, 0.2, 0.4)$$

Self-attention



李哥 考研

$$q^i = a^i W^q$$

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

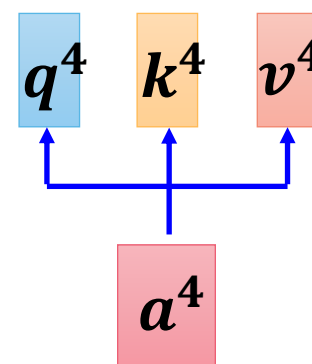
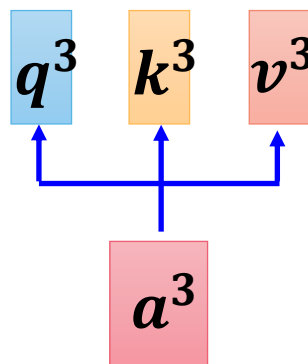
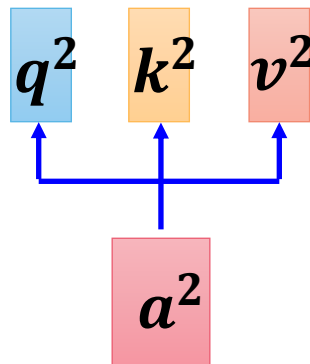
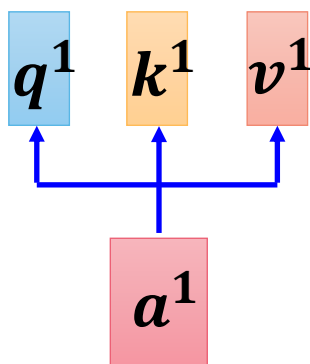
$$k^i = a^i W^k$$

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

$$v^i = a^i W^v$$

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

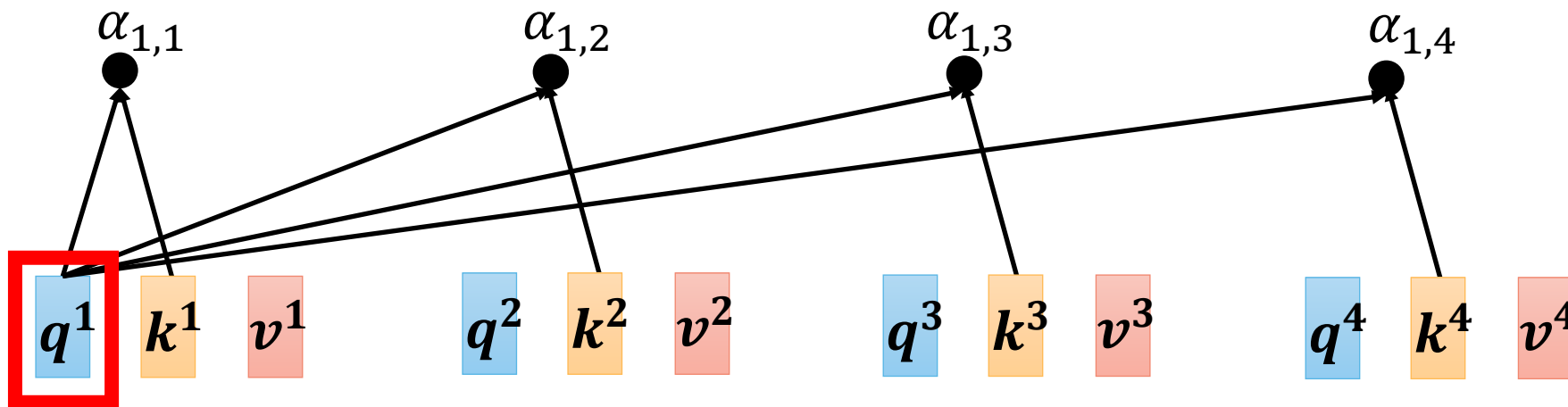
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$





Self-attention

$$\begin{aligned}\alpha_{1,1} &= q^1 k^1 & \alpha_{1,2} &= q^1 k^2 \\ \alpha_{1,3} &= q^1 k^3 & \alpha_{1,4} &= q^1 k^4\end{aligned}\quad \begin{matrix} \alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} & \alpha_{1,4} \\ & & & \\ & & & \\ & & & \end{matrix} = q^1 \begin{matrix} k^1 \\ k^2 \\ k^3 \\ k^4 \end{matrix}$$

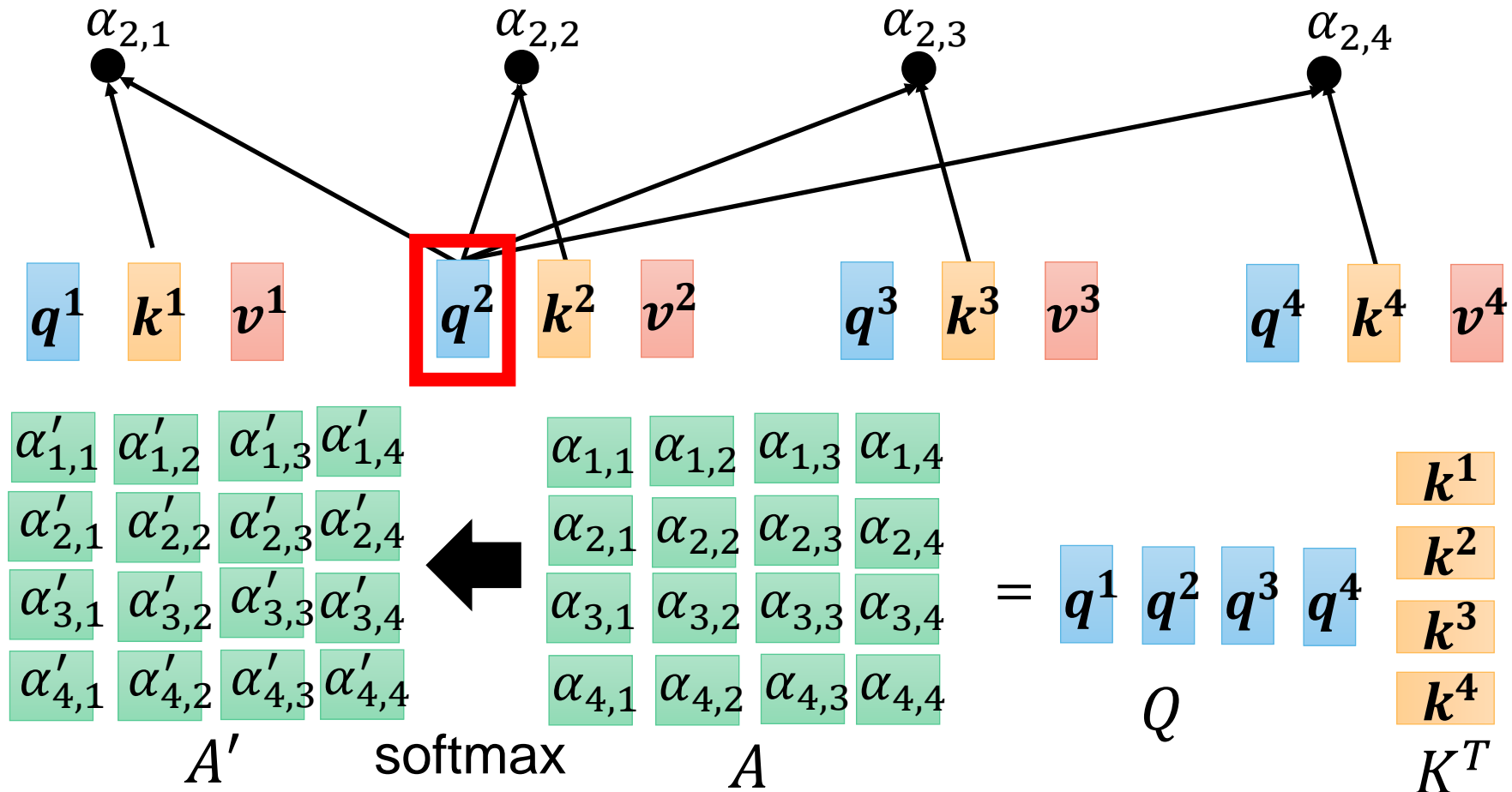


Self-attention

$$\begin{aligned}
 \alpha_{1,1} &= q^1 k^1 & \alpha_{1,2} &= q^1 k^2 & \alpha_{1,3} &= q^1 k^3 & \alpha_{1,4} &= q^1 k^4 \\
 \alpha_{1,3} &= q^1 k^3 & \alpha_{1,4} &= q^1 k^4
 \end{aligned}
 \quad
 \begin{matrix}
 \alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} & \alpha_{1,4} \\
 = & q^1 & k^1 & k^2 & k^3 & k^4
 \end{matrix}$$



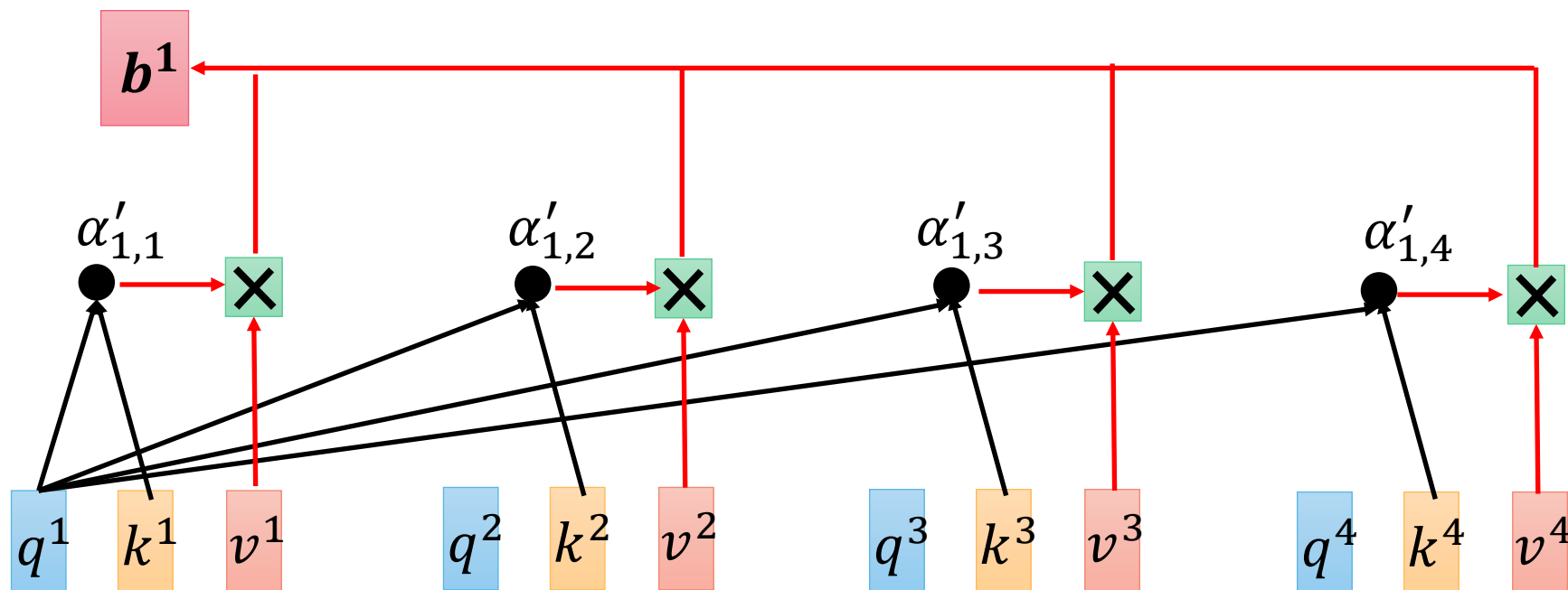
李哥 考研



Self-attention



李哥考研

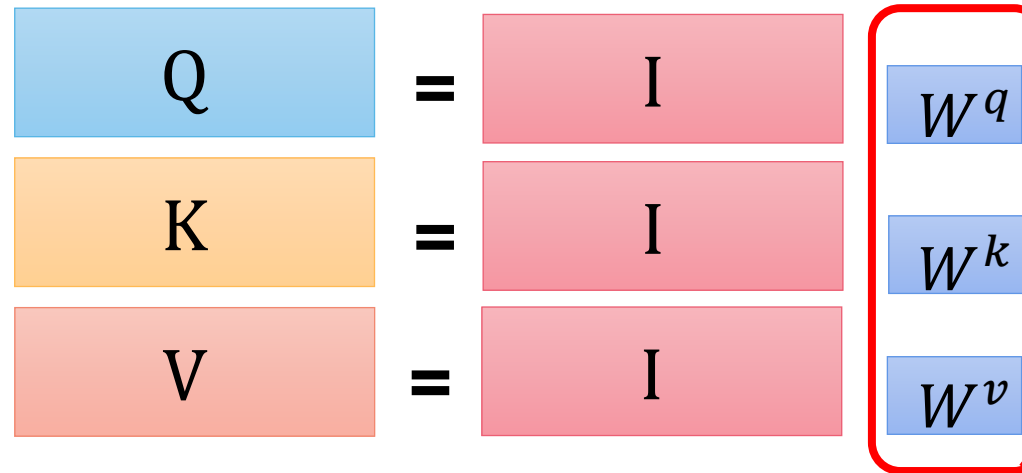


$$\begin{matrix} b^1 & b^2 & b^3 & b^4 \\ 0 \end{matrix} = \begin{matrix} \boxed{\alpha'_{1,1} & \alpha'_{1,2} & \alpha'_{1,3} & \alpha'_{1,4}} \\ \alpha'_{2,1} & \alpha'_{2,2} & \alpha'_{2,3} & \alpha'_{2,4} \\ \alpha'_{3,1} & \alpha'_{3,2} & \alpha'_{3,3} & \alpha'_{3,4} \\ \alpha'_{4,1} & \alpha'_{4,2} & \alpha'_{4,3} & \alpha'_{4,4} \end{matrix} \begin{matrix} v^1 & v^2 & v^3 & v^4 \\ V \end{matrix}$$

A'

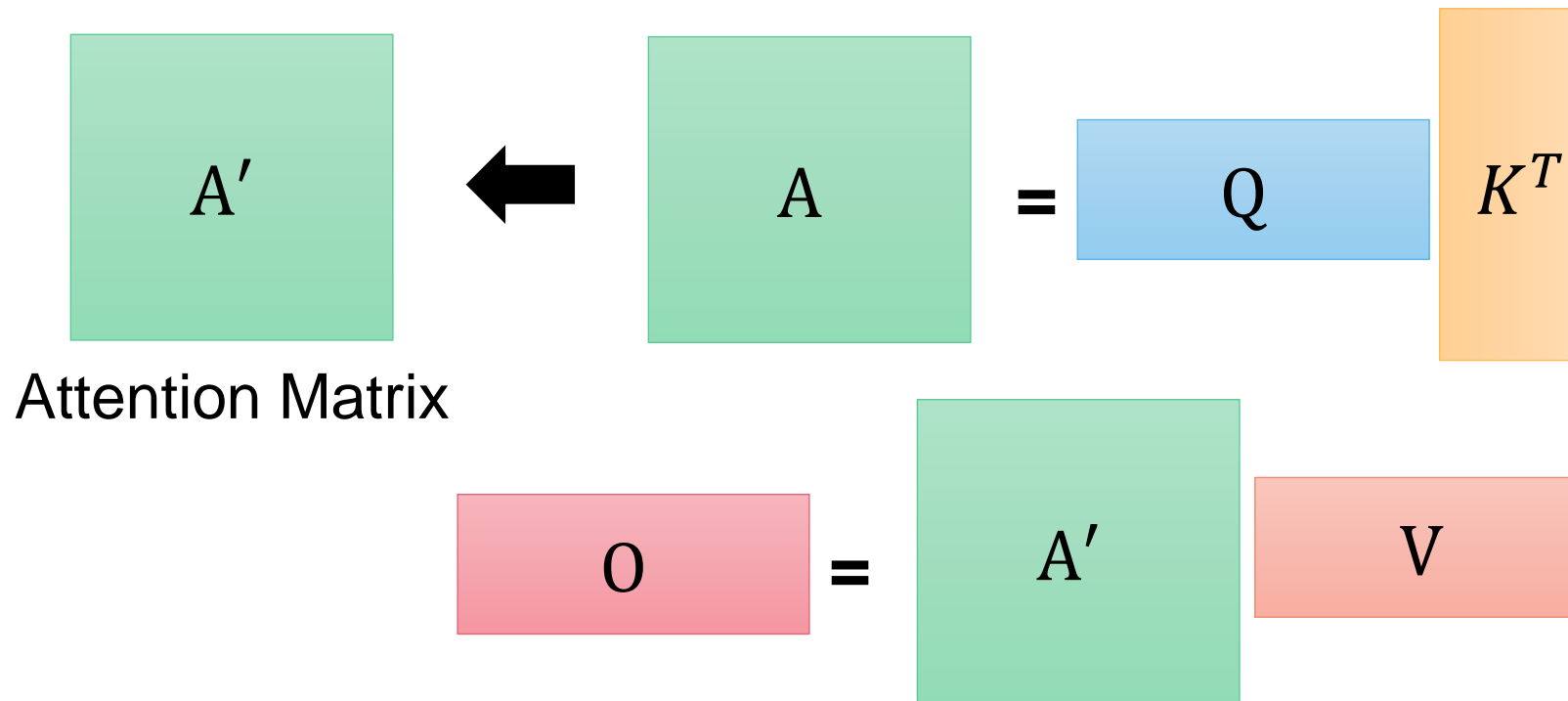
Self-attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



李哥 考研

只有这些是要学习的



回头过一遍维度。

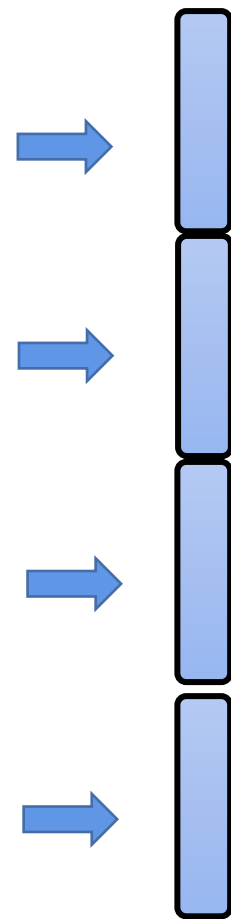
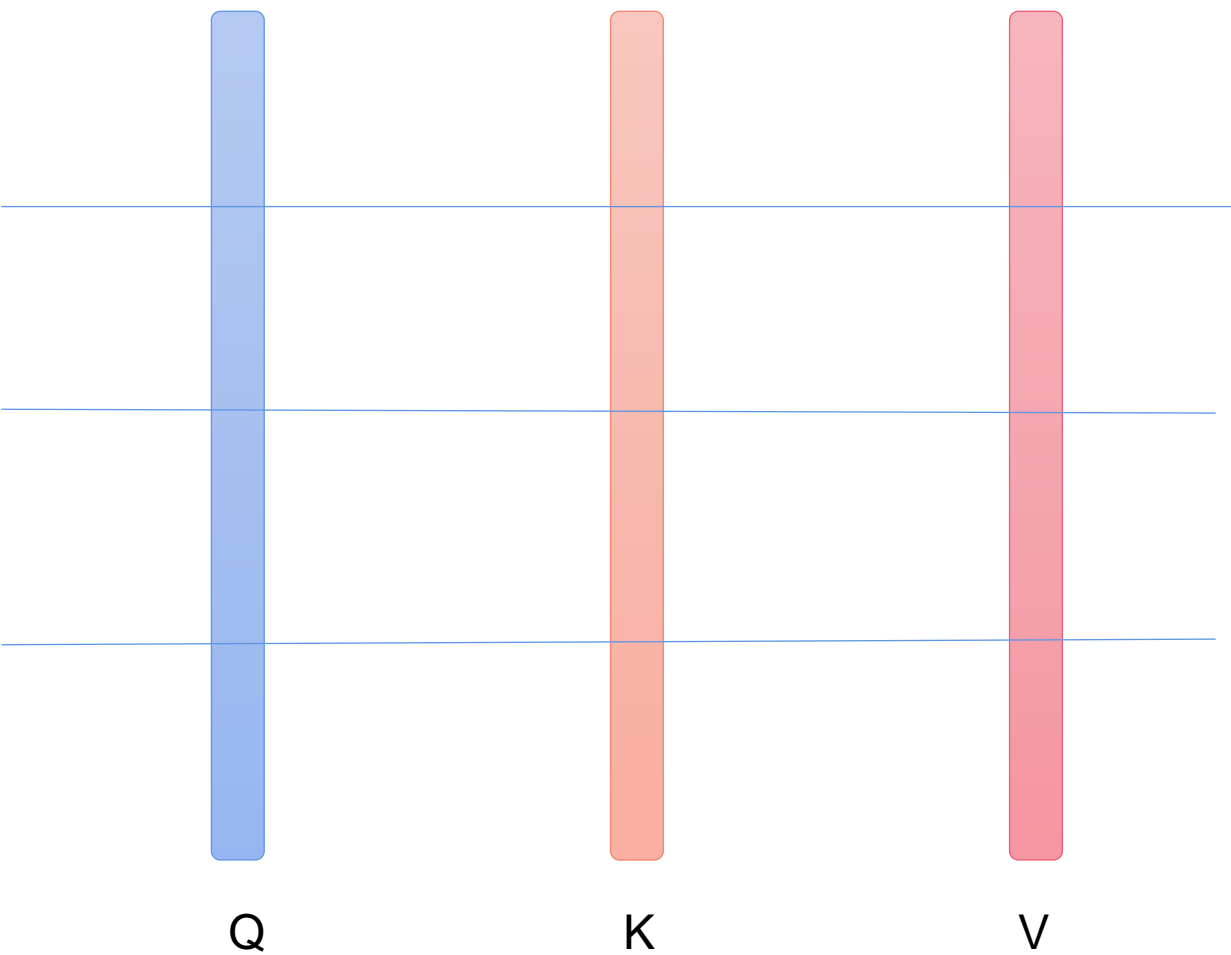


李哥考研

- 一般一个self-attention 模型， 一个字的编码维度是768， 字的个数可以按照128， 或者512计算。

多头自注意力机制

4个头

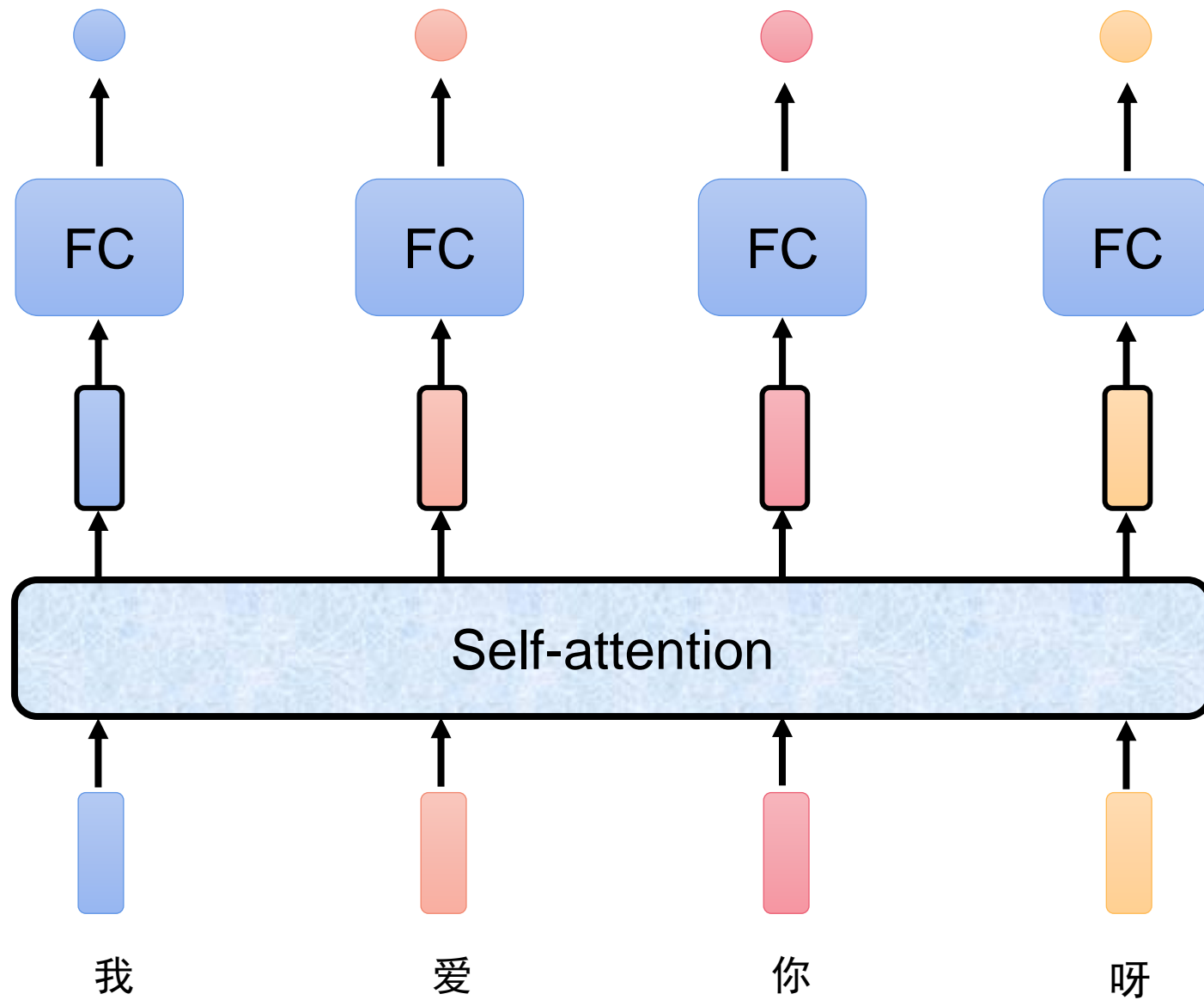


李哥考研

Self-attention

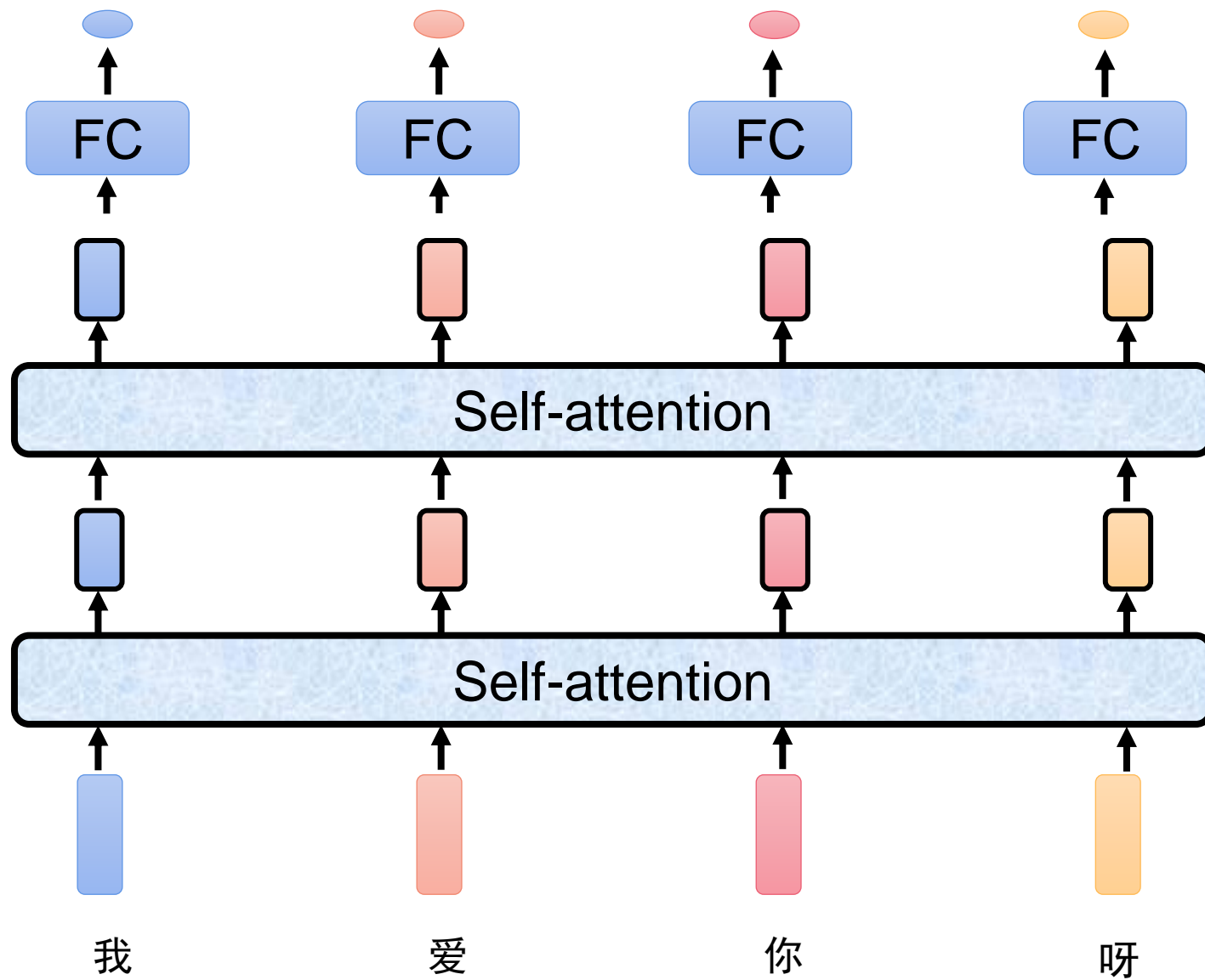


李哥考研





想擦几层擦几层

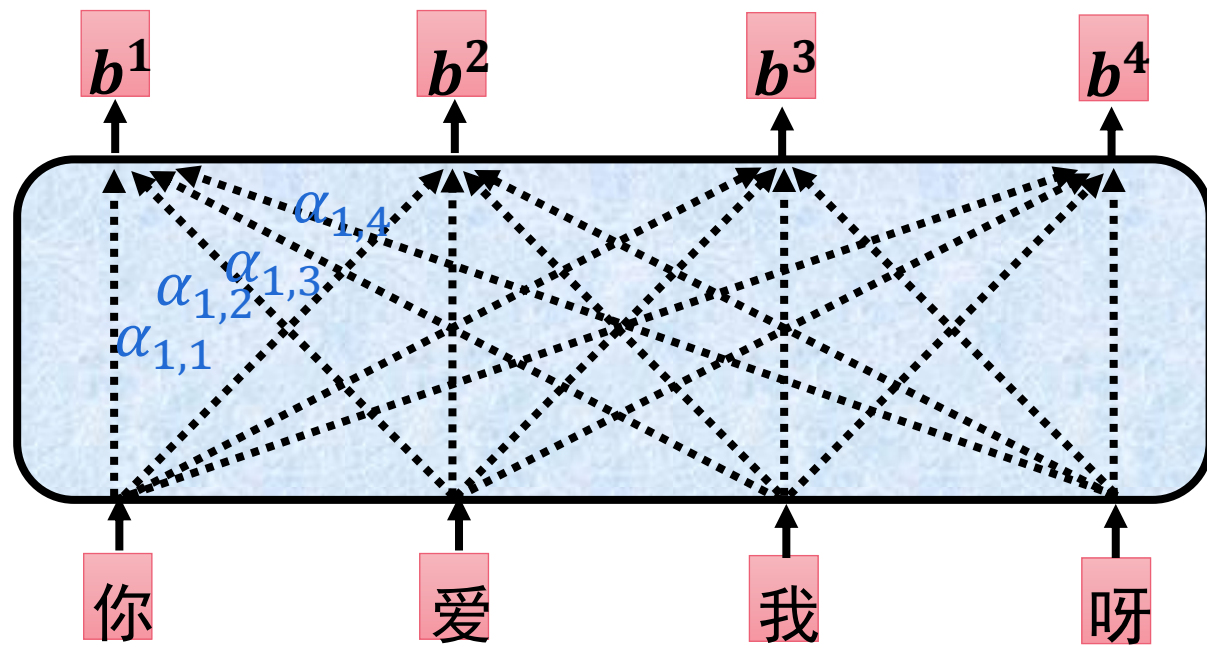
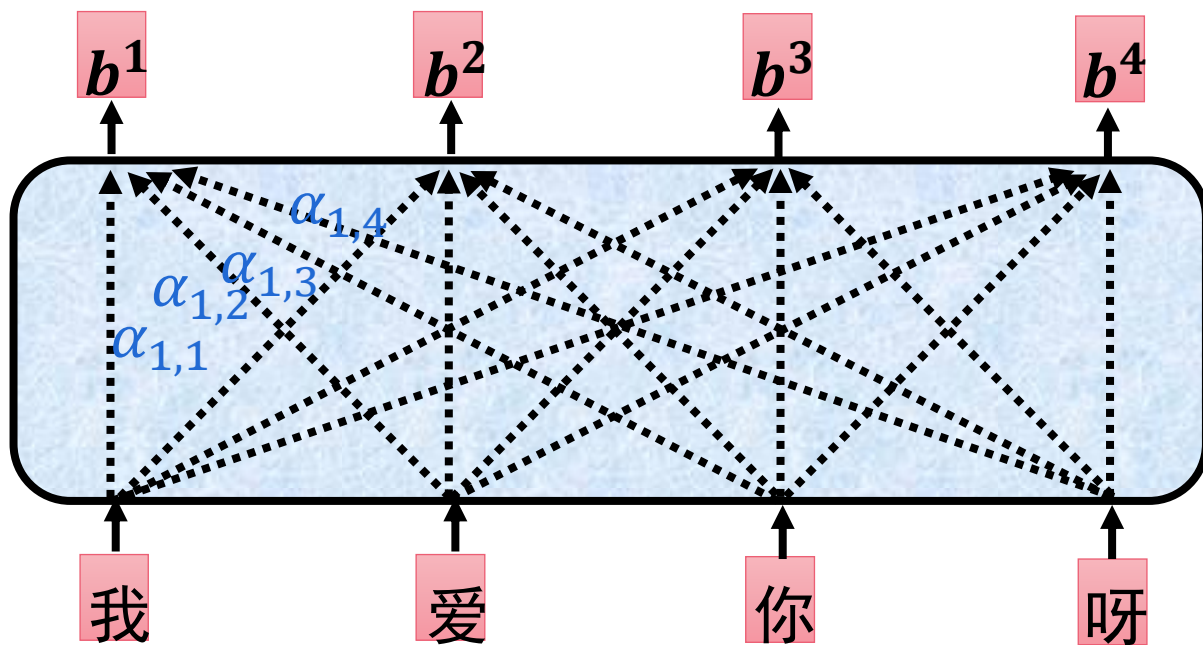


位置信息。

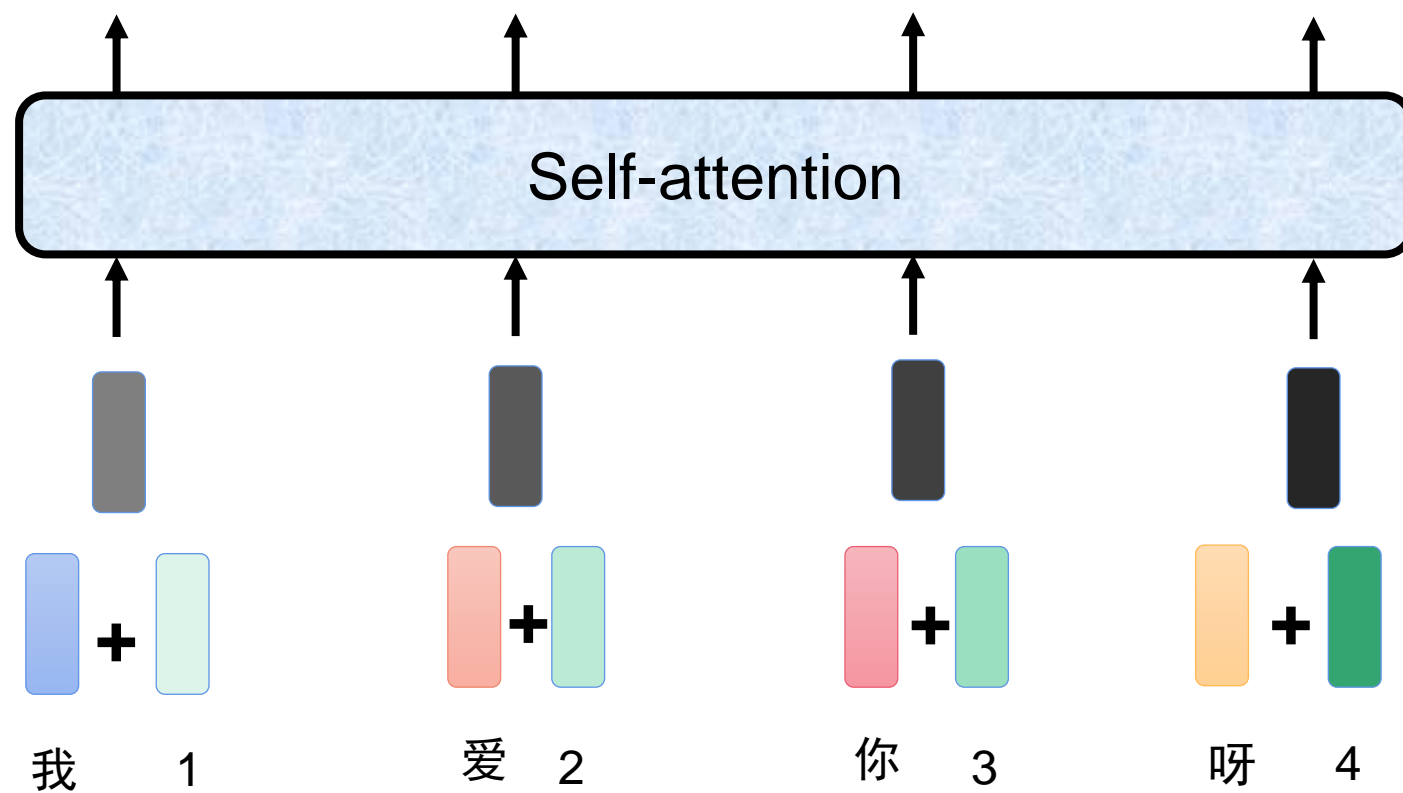


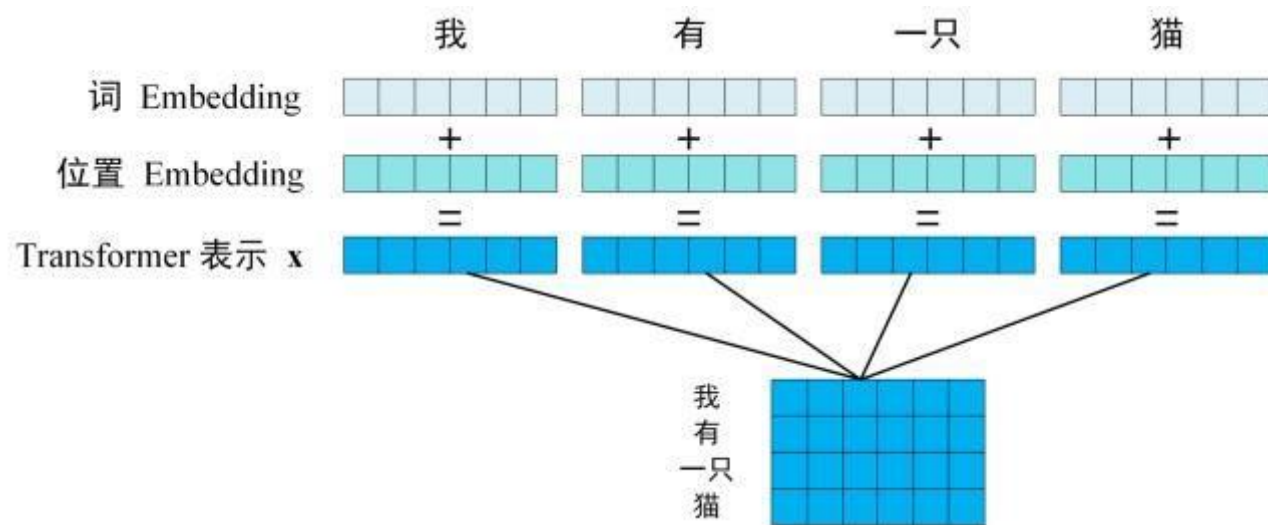
李哥考研

当分类时，可能把所有的特征横着加起来。作为输出特征。



模型无法区分这两种情况！





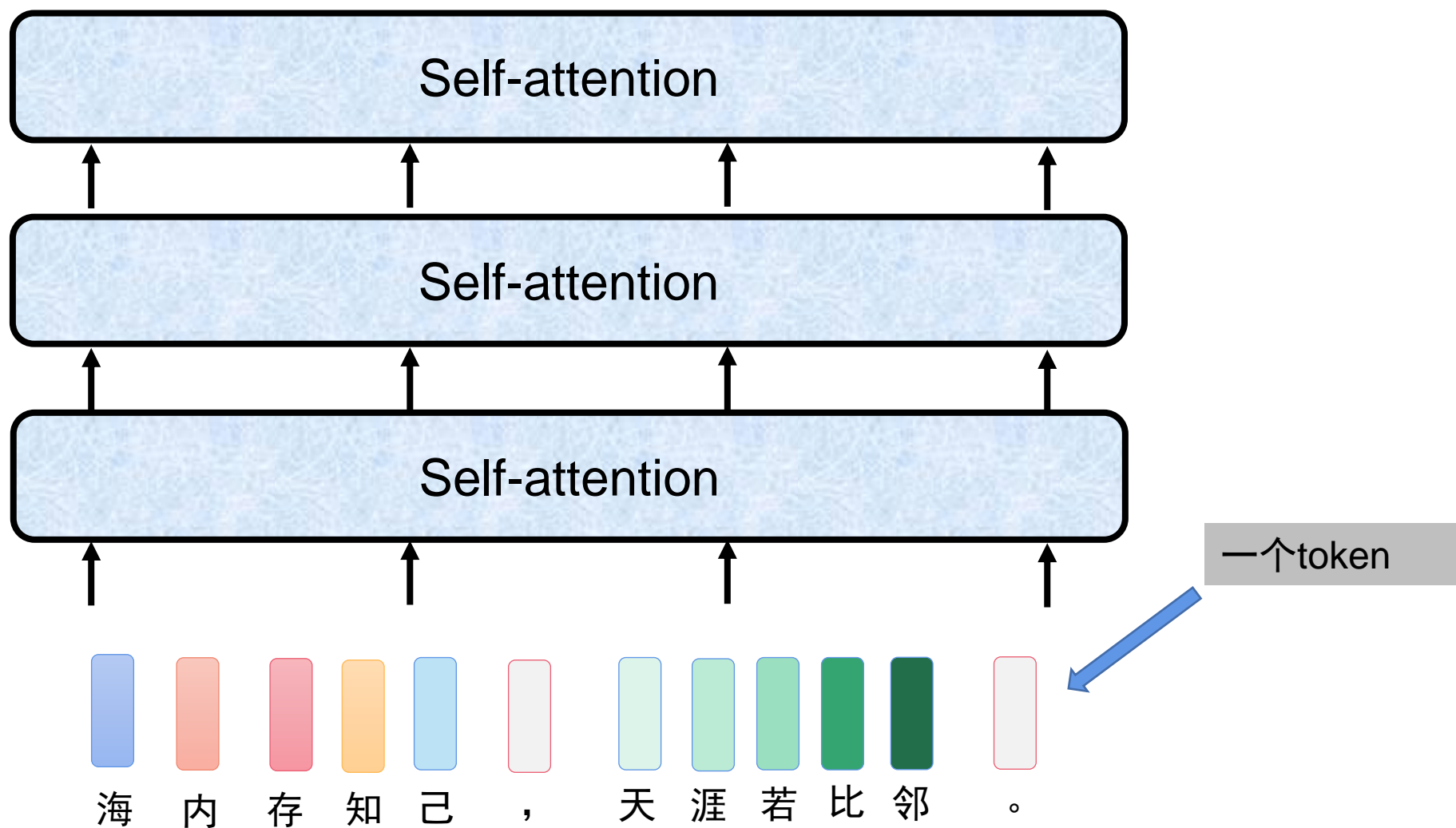
Linear (21128, 768)

Linear (512, 768)

这就是self-attention。

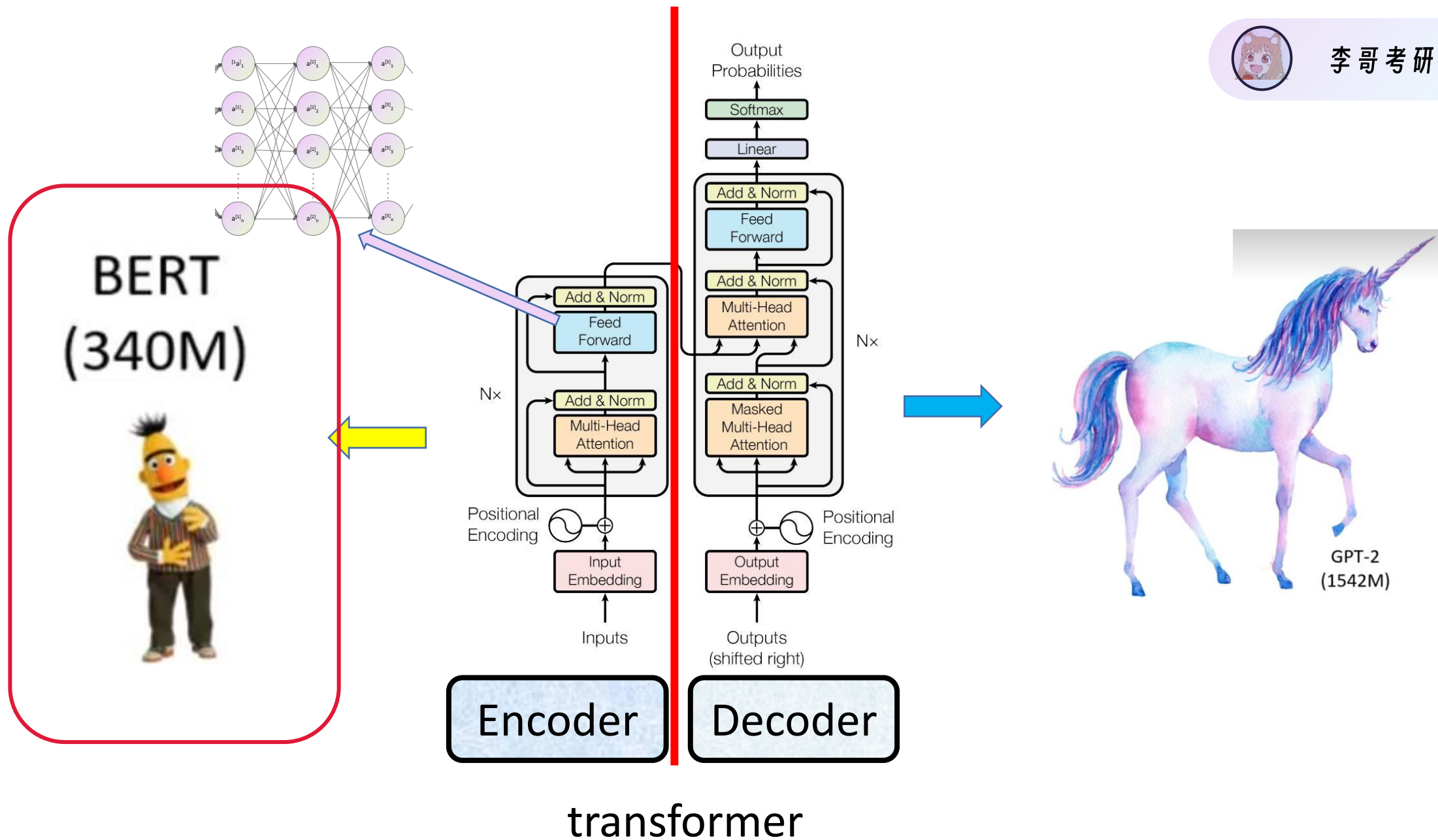


李哥考研

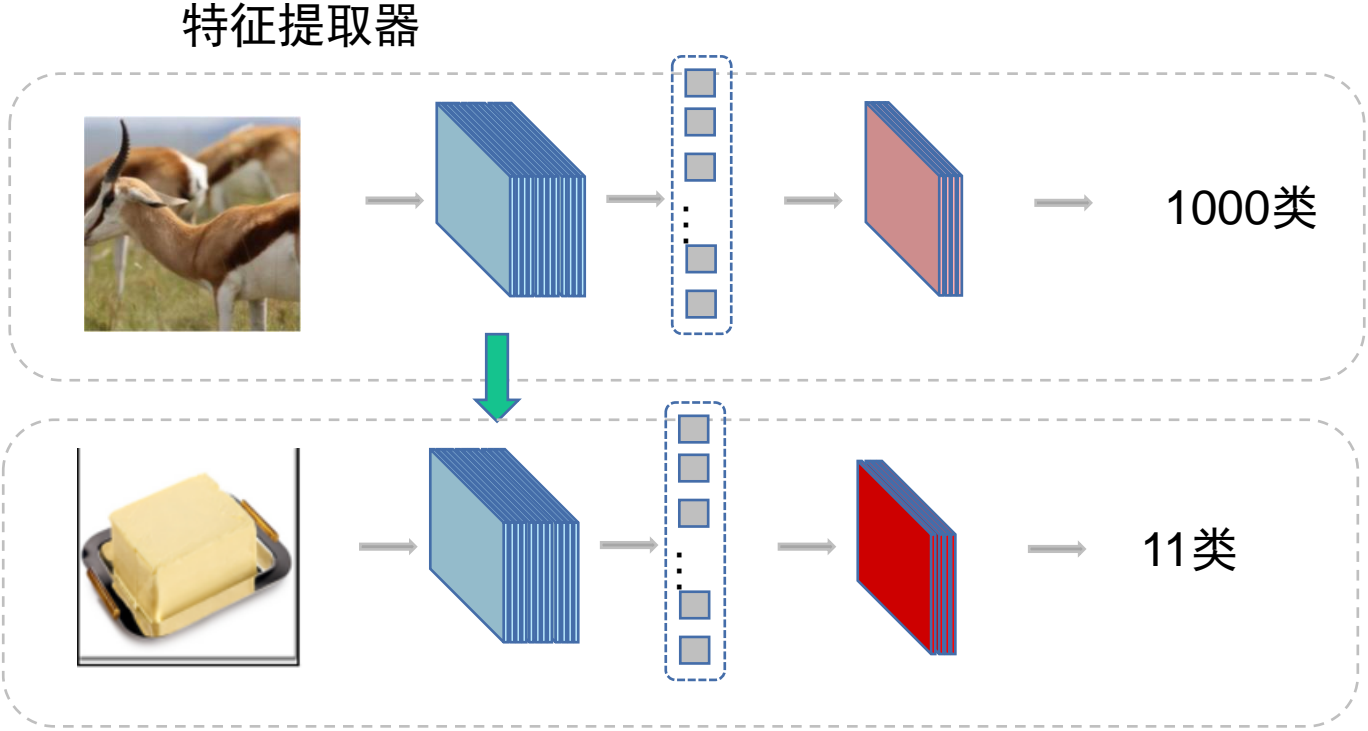





李哥考研



迁移学习

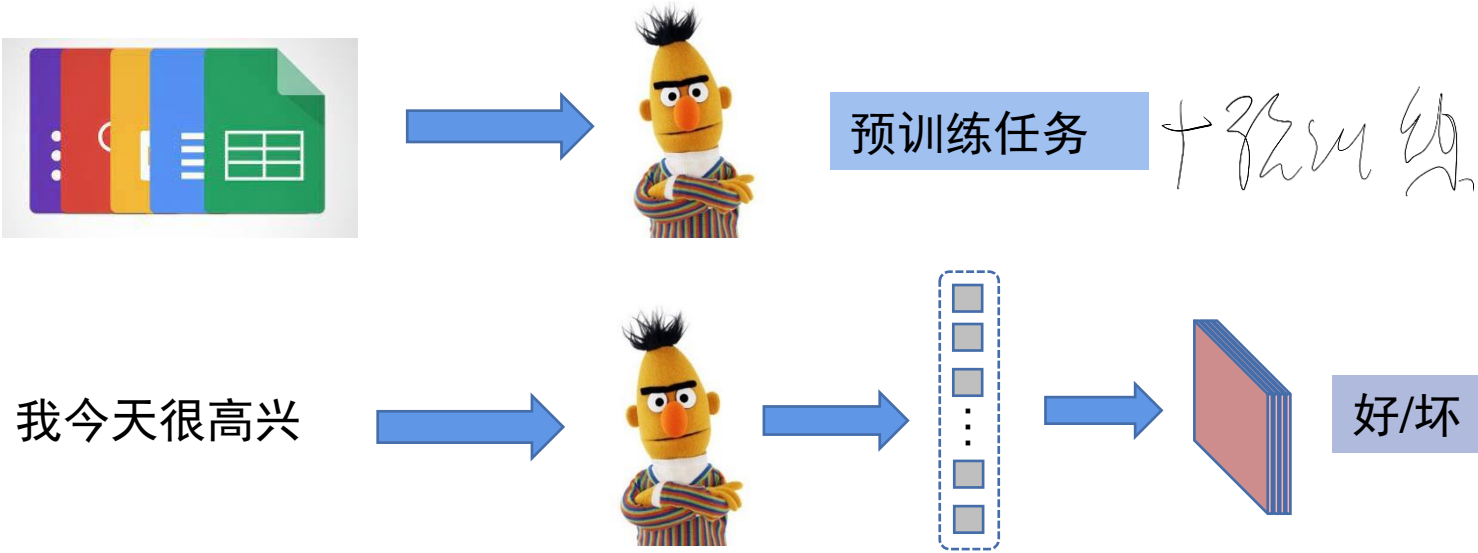




李哥 考研

预训练

微调



Bert就是一个特征提取器

预训练任务

微调

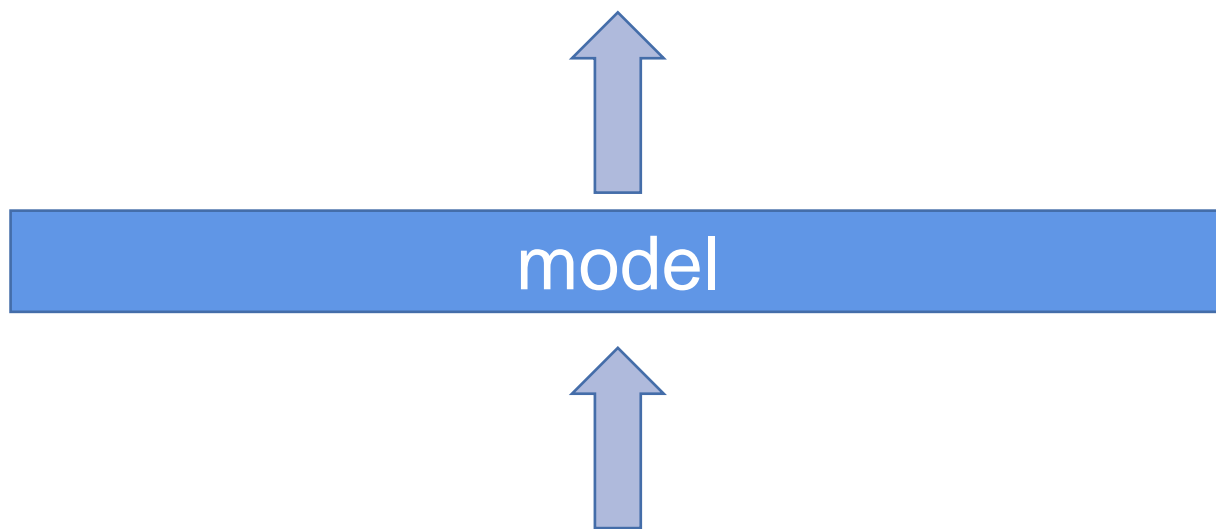
好/坏

自监督预训练



李哥考研

去的尽管去了，来的尽管来着；去来的中间，又怎样地匆匆呢？早上我起来的时候，小屋里射进两三方斜斜的太阳。太阳他有脚啊，轻轻悄悄地挪移了；我也茫茫然跟着旋转。



没错，就是我

去的尽M去了，来的尽管来M；去来的中间，又怎样M匆匆呢？早上我起M的时候，小屋里射进两三方M斜的太阳。太M他有脚啊，轻轻悄悄地挪移了；我也茫茫然跟着旋转。

Bert-Pre-Training



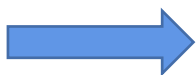
李哥 考研

Masked Language Model

80% : my dog is hairy → my dog is [MASK]

10% : my dog is hairy → my dog is apple

10% : my dog is hairy → my dog is hairy.



预训练任务

Next Sentence Prediction

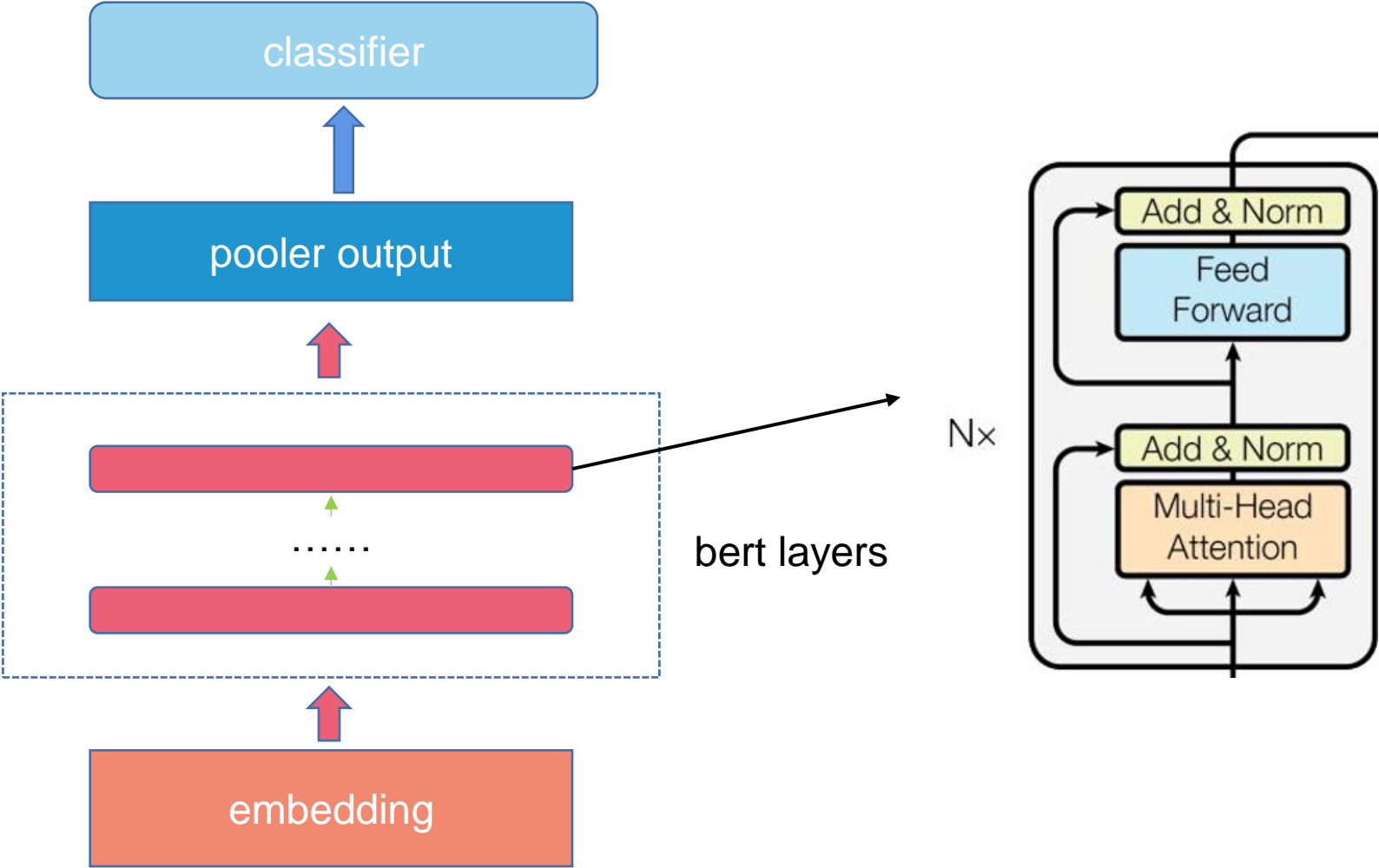
Input = [CLS] the man went to [MASK] store [SEP]
he bought a gallon [MASK] milk [SEP]

Label = IsNext

Input = [CLS] the man [MASK] to the store [SEP]
penguin [MASK] are flight ##less birds [SEP]

Label = NotNext

Bert结构。



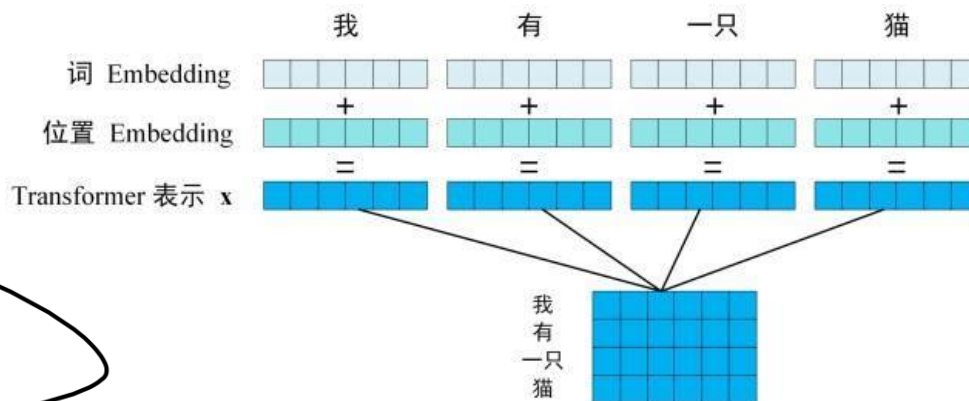


Bert输入embedding

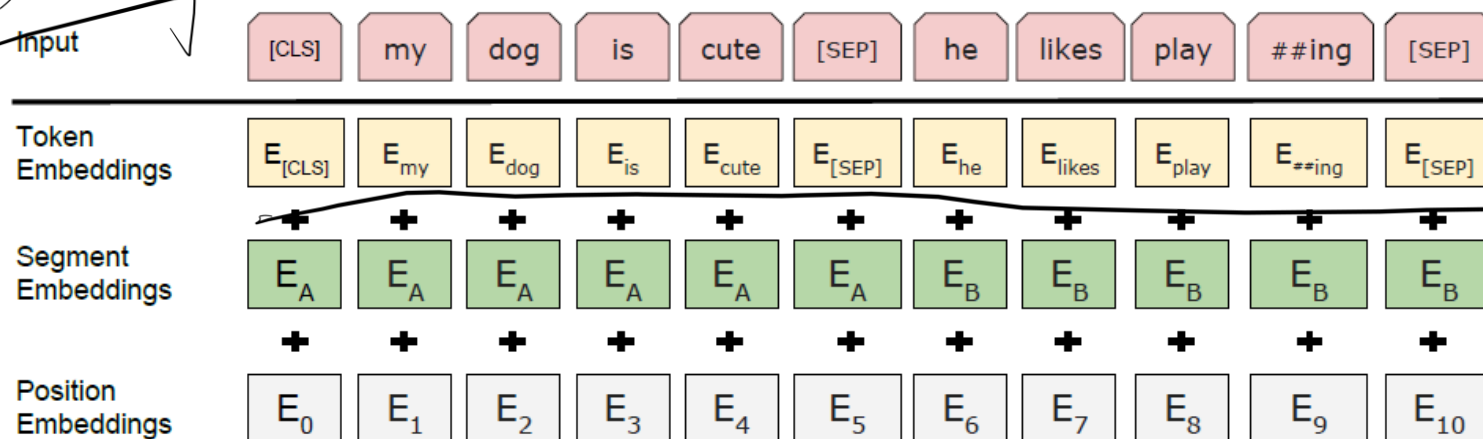


李哥考研

Transformer:



Bert:



可以看到一些特殊的token。



Bert输出pooler



李哥考研

classifier

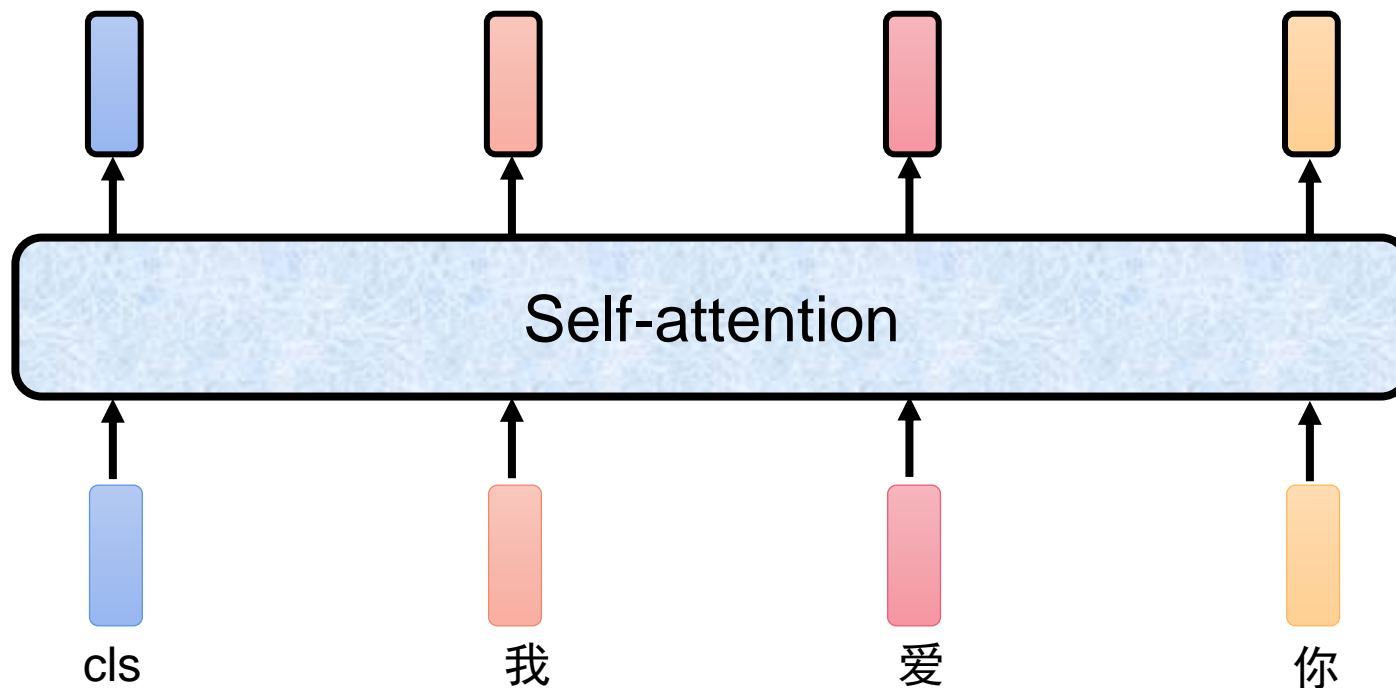
: 我只需要一个输入

1, 只输出CLS

2, 平均池化

3, 最大池化

4, 其他



本节回顾



李哥考研

- 这节课内容非常多。

一： 字的表示。

二： 为了知道上下文， 引入了RNN和LSTM。

三： 上面的模型， 不能并行， 速度慢， 只能单向（也有双向模型）。所以引入Self-attention

四： Self-attention， 每一个字加上位置都embedding， 相加为token。之后， 通过Q,K,V来交互， Q是query， K是key， V是value。 Q和K算出注意力， V按照注意力相加为输出。

五： bert， 是一个编码器， 目的是把一句话编码为特征。 它采用自监督预训练获得特征提取能力。 之后在下游任务可以提取特征后， 让特征去做分类任务。

六： bert的结构， 三部分。 一， embedding层。二， Self-attention层让特征交互。
三， pooler输出。

下节预告。



李哥考研

带大家计算Bert的参数量。

用Bert进行情感分类。



作业：

回顾今天的内容，

计算self-attention 维度变化。

搜索了解Bert的架构和预训练任务。

在CSDN写下你所学到的内容。



答疑和结束

THANKS


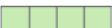












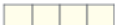
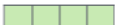








bert



李哥 考研

What is the best contextualized embedding for “Help” in that context?
For named-entity recognition task CoNLL-2003 NER

		Dev F1 Score
12 	First Layer Embedding 	91.0
• • •	Last Hidden Layer 12 	94.9
7 	12 	95.5
6 	+ ... +	
5 	2 	
4 	+ 1 	
3 	= 	95.6
2 	Second-to-Last Hidden Layer 11 	
1 	Sum Last Four Hidden	95.9
	12 	
Help	+ 11 	
	+ 10 	
	+ 9 	96.1
	= 	
	Concat Last Four Hidden 9 10 11 12 	https://blog.csdn.net/iterate7

