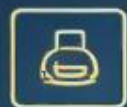


第09章 注意力机制

欧 新 宇

自注意力机制

Self-Attention Mechanism

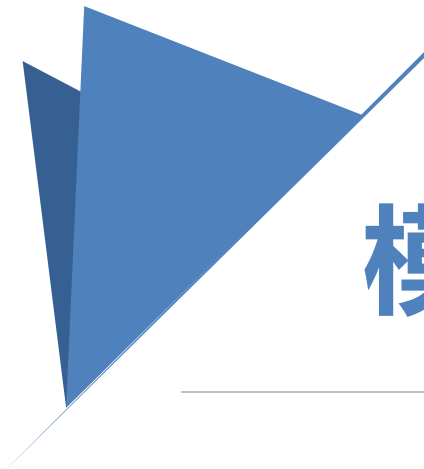


$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$



- 模型的输入和输出
- 自注意力
- 位置编码



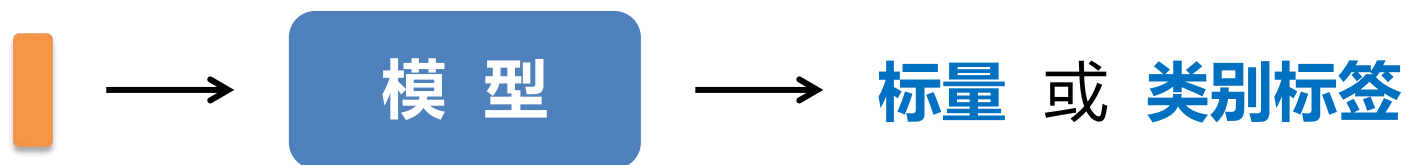


模型的输入和输出

自注意力机制

模型的输入

- 输入是一个向量



- 输入是一组向量



自注意力机制

基于向量集 (Vector Set) 的文本

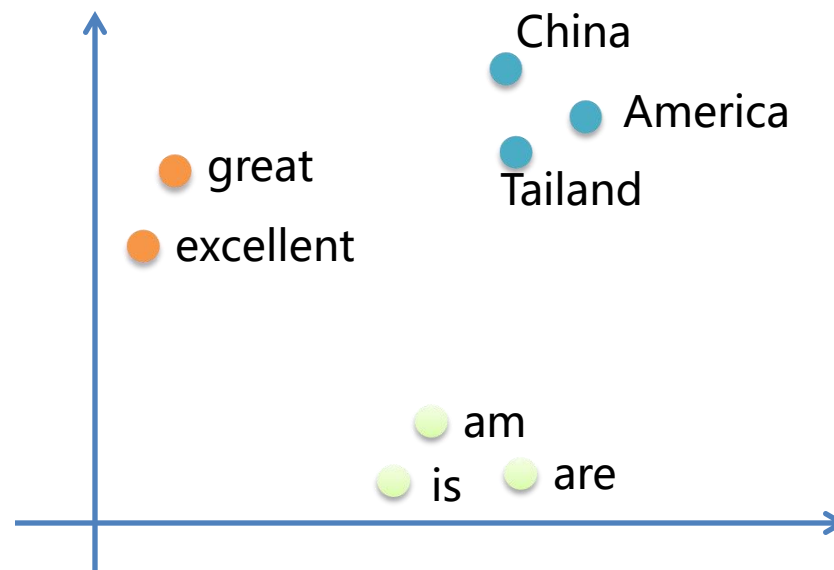
China is a great country



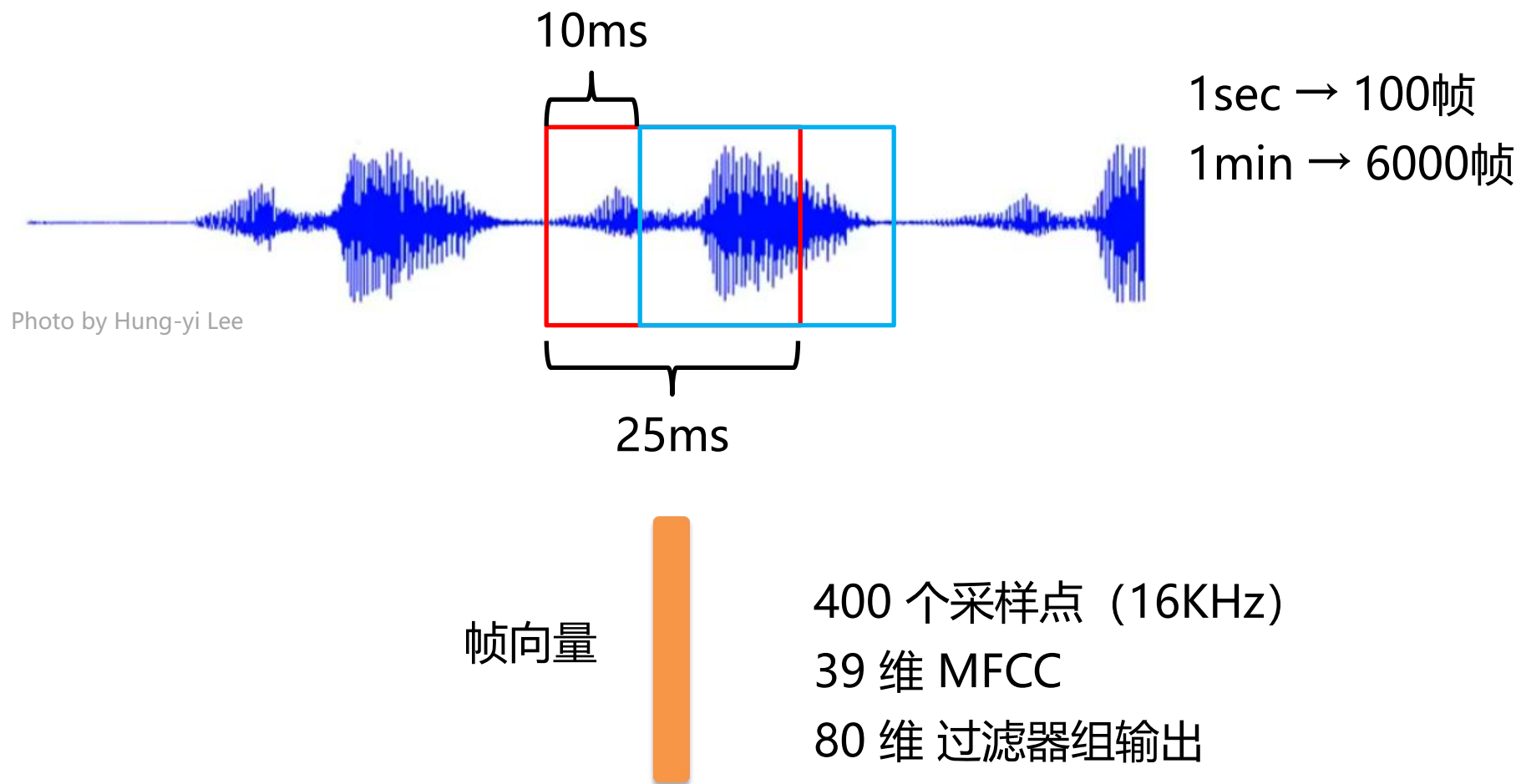
One-hot 编码

China = [1 0 0 0 0 0 0 ...]
is = [0 1 0 0 0 0 0 ...]
a = [0 0 1 0 0 0 0 ...]
great = [0 0 0 1 0 0 0 ...]
country = [0 0 0 0 1 0 0 ...]
America = [0 0 0 0 0 1 0 ...]
Tailand = [0 0 0 0 0 0 1 ...]

词嵌入 (word embedding)

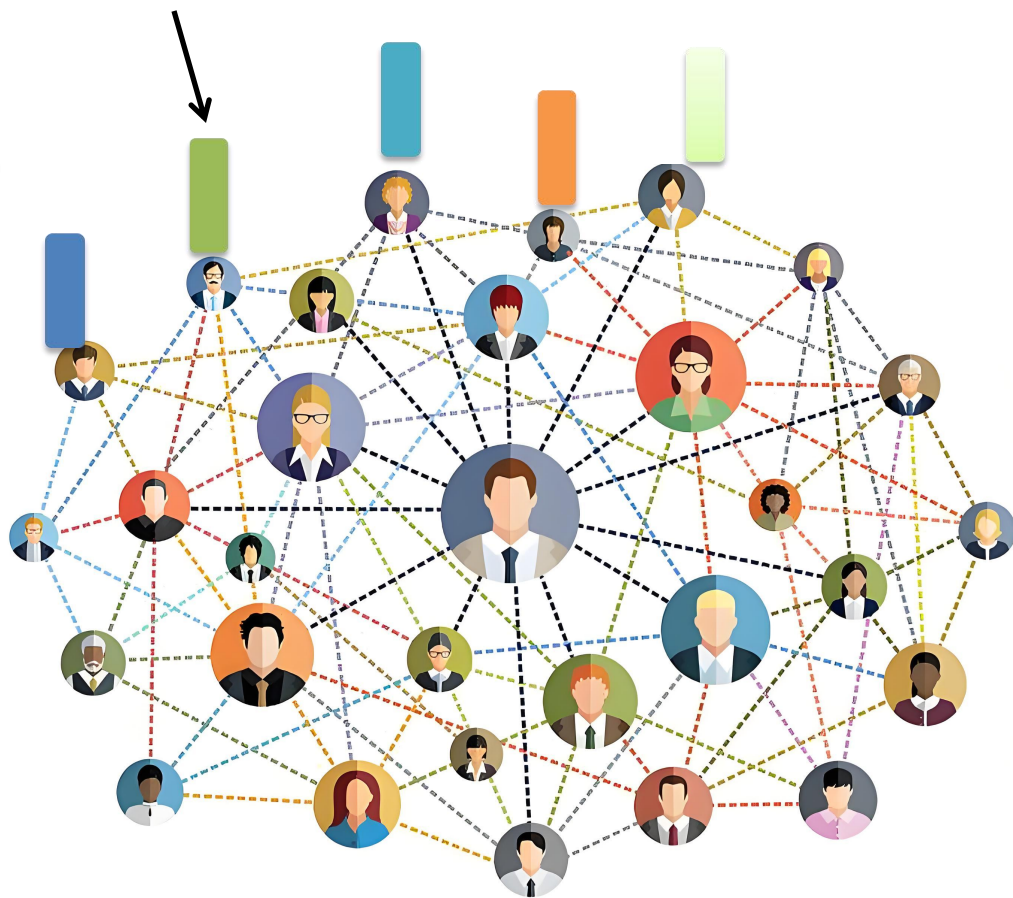


基于向量集的语音

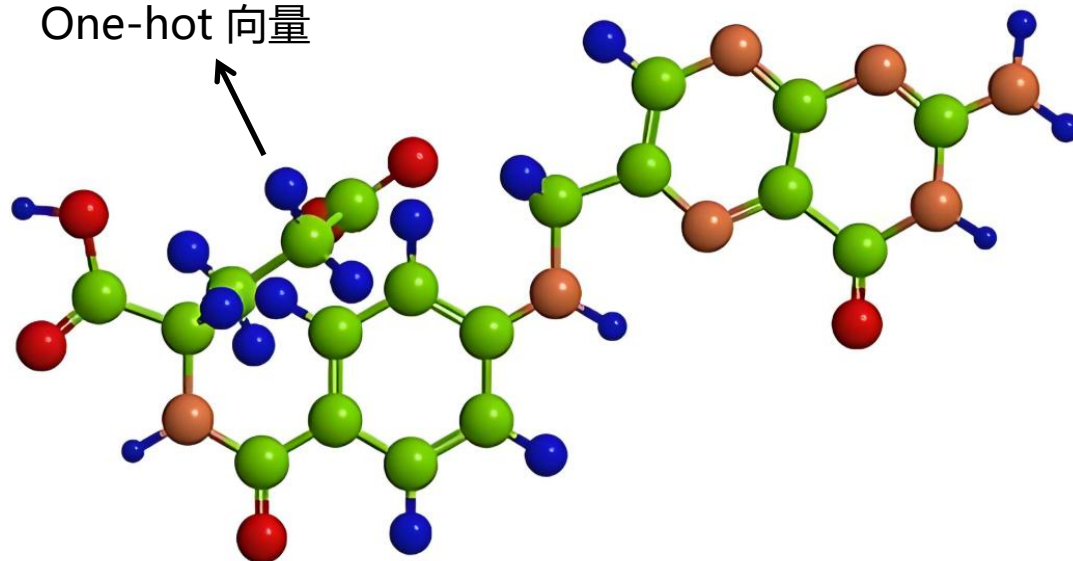


基于向量集的图网络

每个人用一个向量表示



One-hot 向量



每个原子用一个向量表示

$$\text{H} = [1 \ 0 \ 0 \ 0]$$

$$\text{C} = [0 \ 1 \ 0 \ 0]$$

$$\text{O} = [0 \ 0 \ 1 \ 0]$$

自注意力机制

多种多样的输入

- 每个向量输出一个结果 (等长输出)



Citizens have the right to right wrongs

↓
名词

↓
动词

↓
冠词

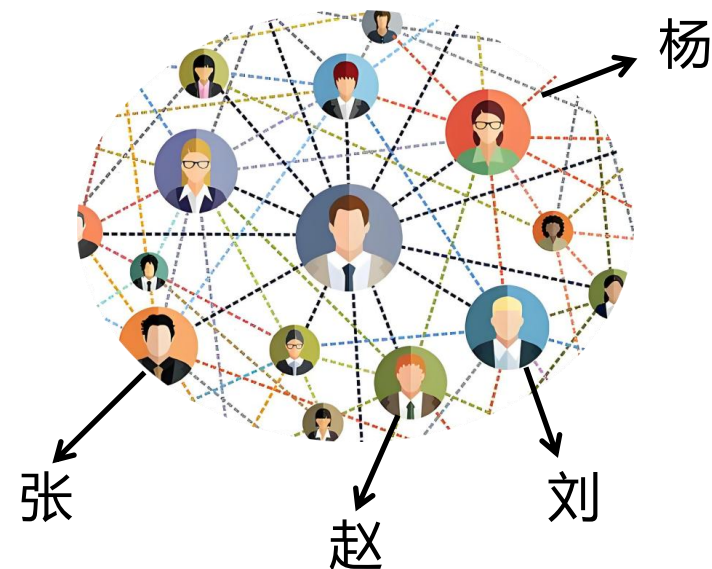
↓
名词

↓
介词

↓
动词

↓
名词

China is a great country



自注意力机制

多种多样的输入

- 整个序列输出一个结果 (全局标签预测)

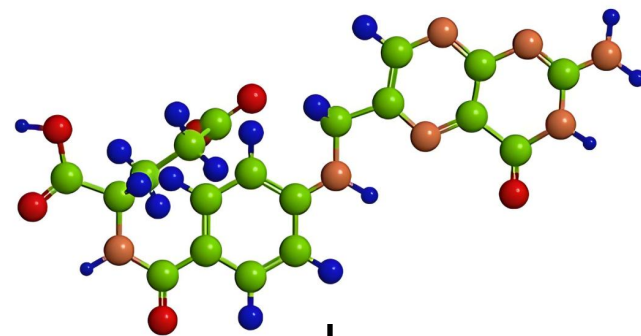


Citizens have the right to right wrongs

中性语境



三国演义

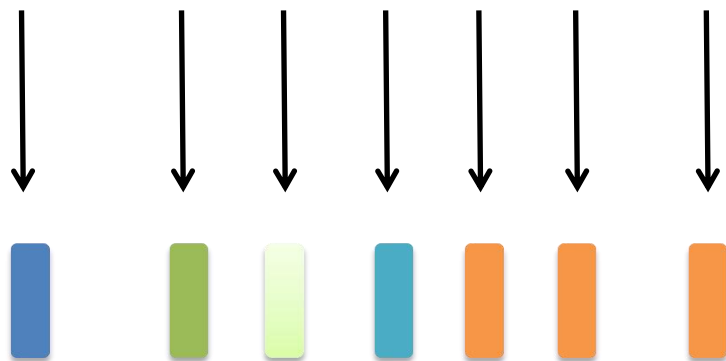


弱毒性

多种多样的输入

● 模型自己决定输出的长度 (seq2seq)

Citizens have the right to right wrongs



模 型

→ 公民享有纠正错误的权利

→ 公民有权纠正不公

→ Les citoyens ont le droit de corriger les injustices.

→ 市民は不公正を是正する権利を持っている。

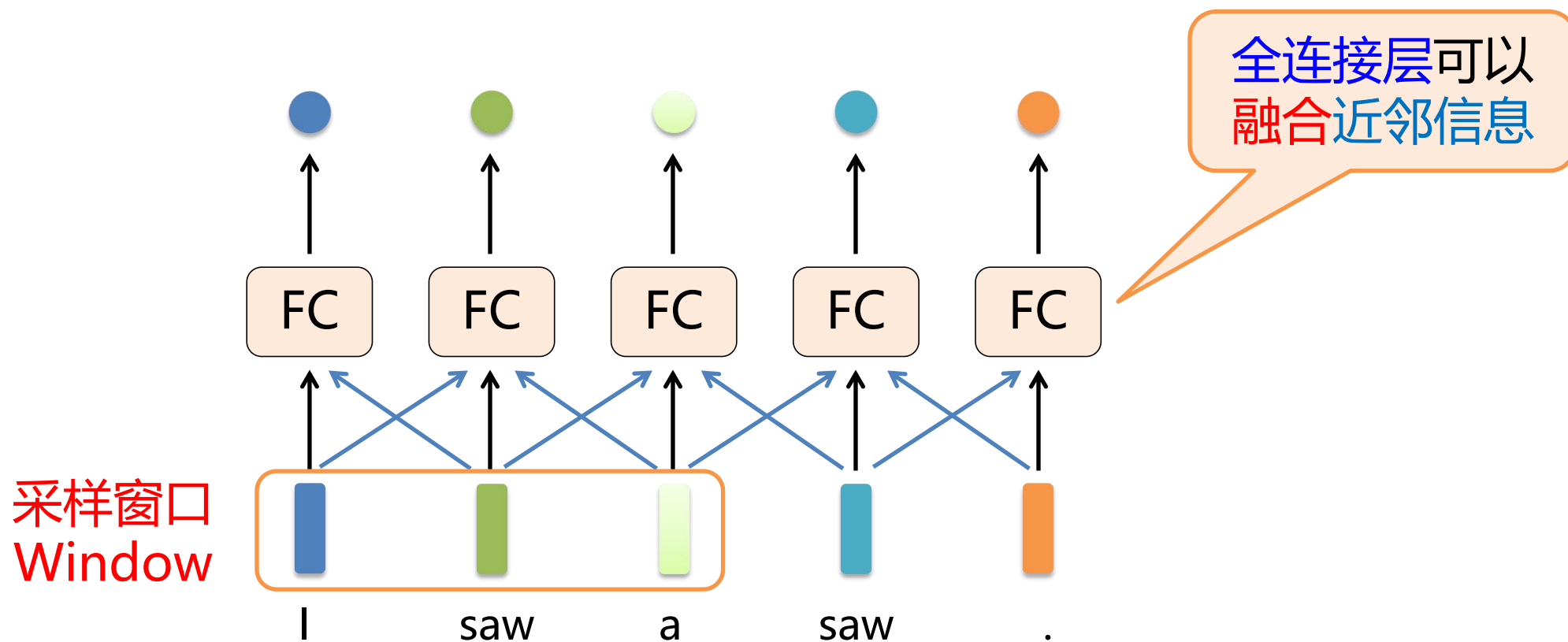


自注意力 (Self-Attention)

自注意力机制

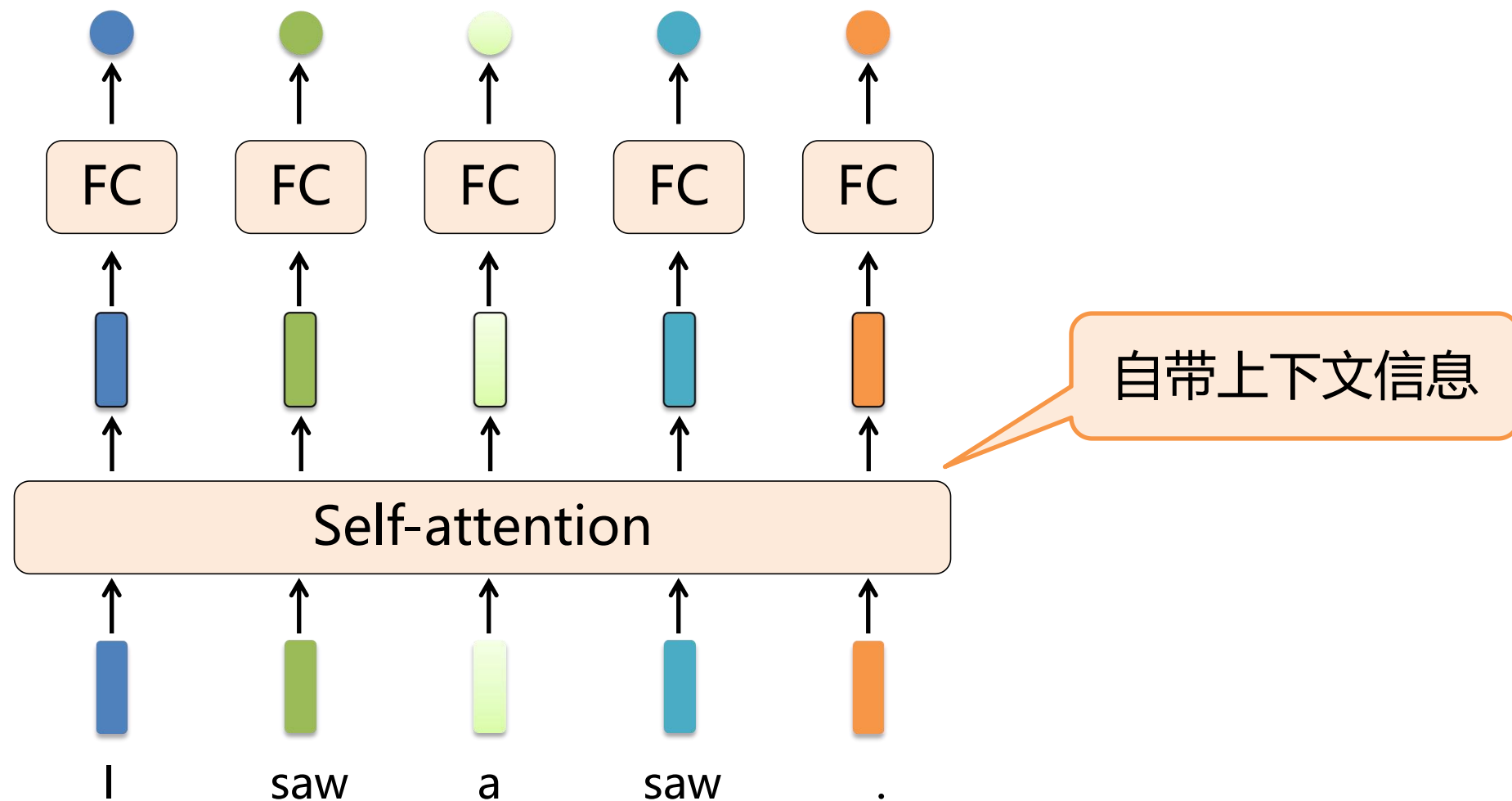
序列标注 (Sequence Labeling)

- 每个向量输出一个结果



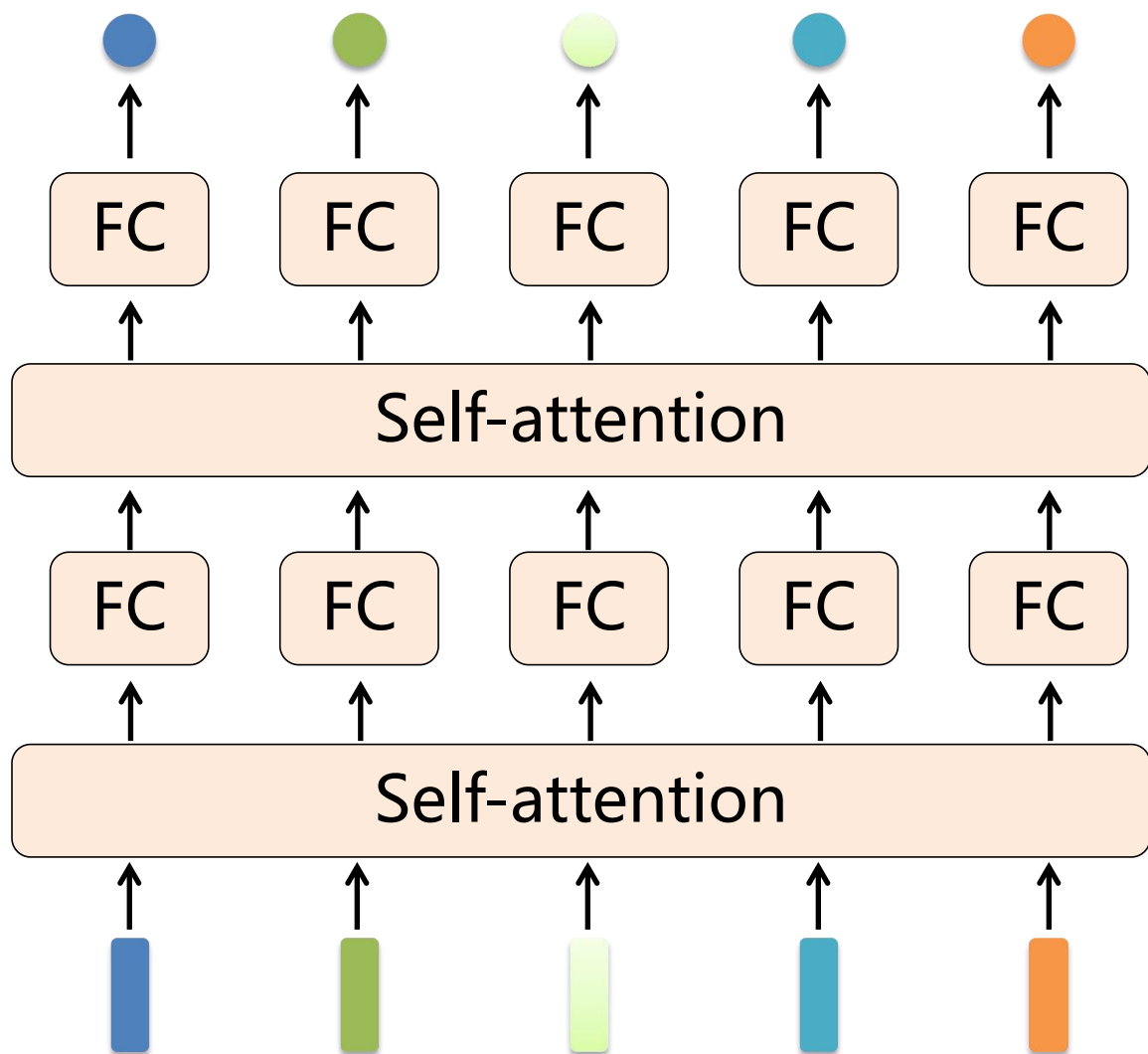
自注意力机制

从序列标注到自注意力



自注意力机制

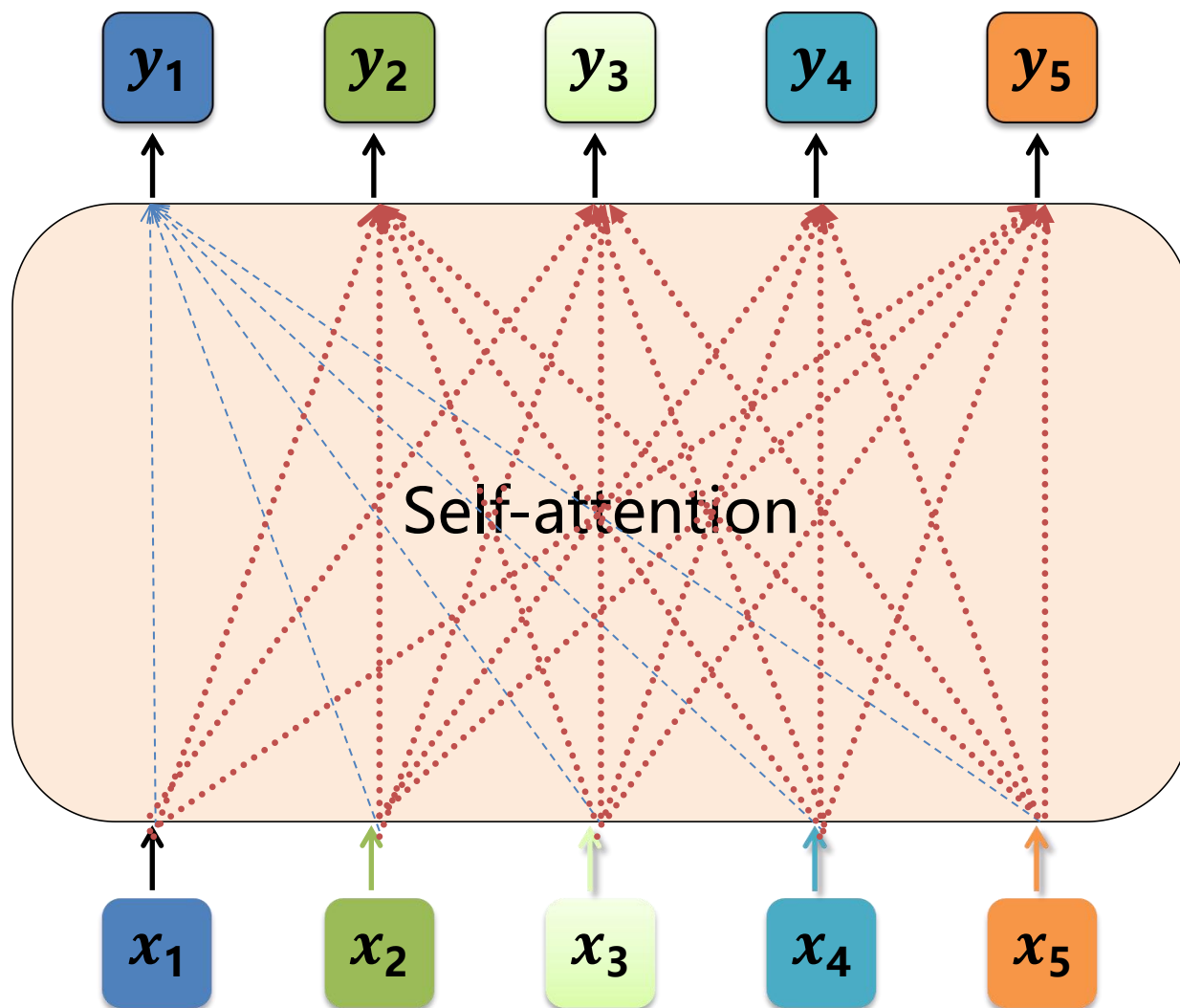
多级自注意力



Attention is all
you need

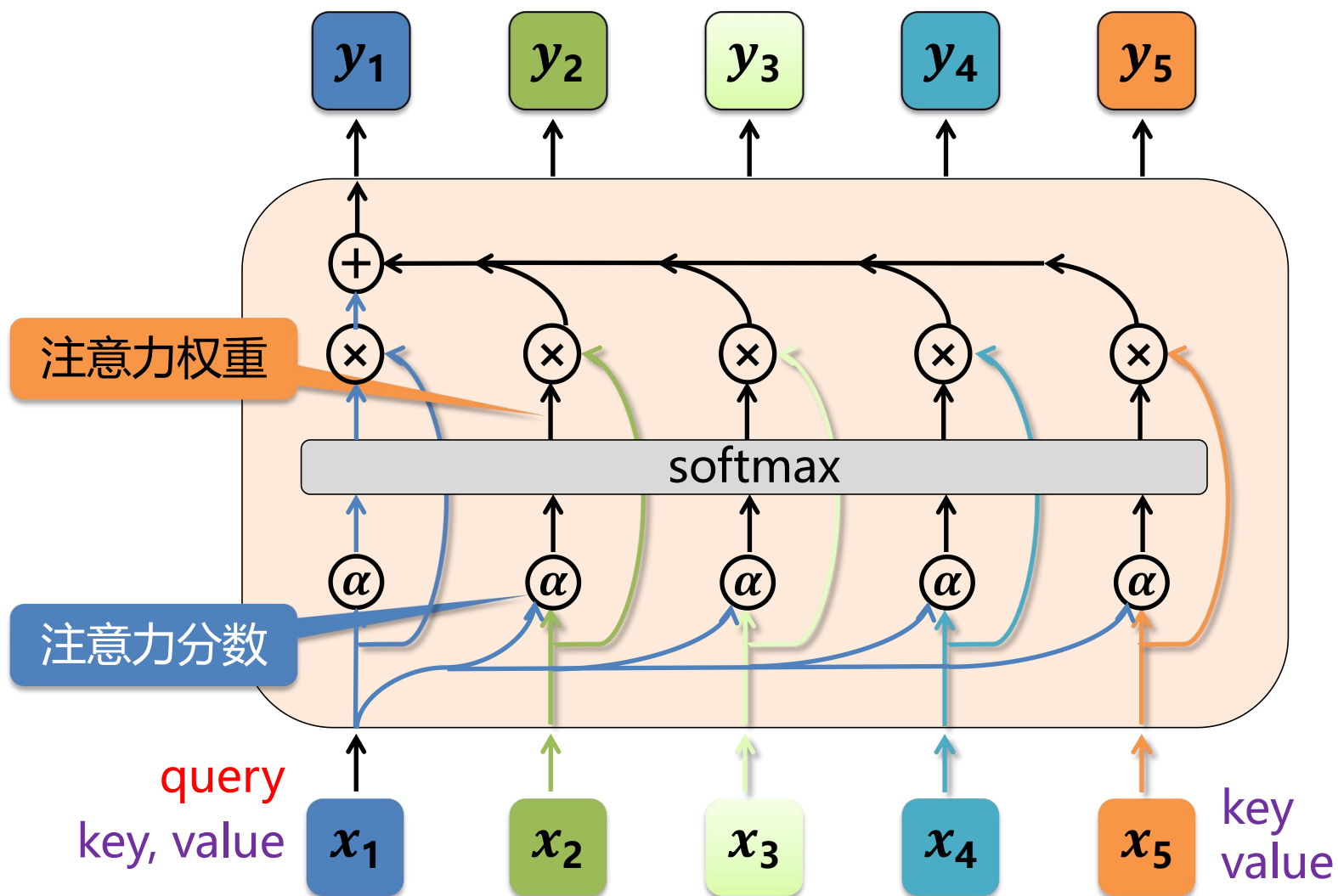
自注意力机制

自注意力的基本原理



自注意力机制

自注意力的基本原理



给定序列 $x_1, x_2, \dots, x_n, \forall x_i \in \mathbb{R}^d$

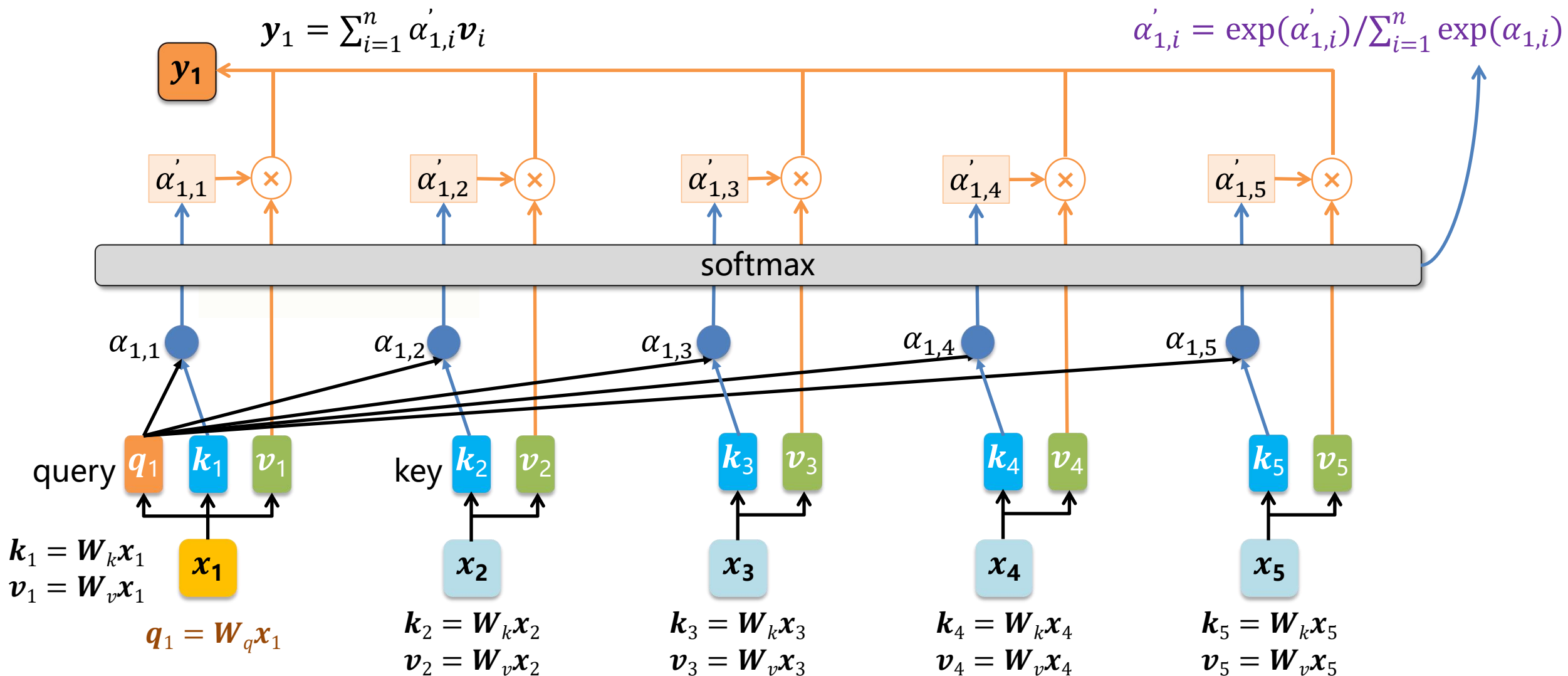
- ✓ 当 $i = 1$ 时, x_1 作为 **query** 输入到自注意力池化层
- ✓ 当 $i \neq 1$ 时, x_1 作为 **key** 和 **value** 输入到自注意力池化层

$$y_i = \text{softmax}\left(\frac{qK^T}{\sqrt{d}}\right)V$$

$$= \text{softmax}\left(\frac{x_i X^T}{\sqrt{d}}\right)X$$

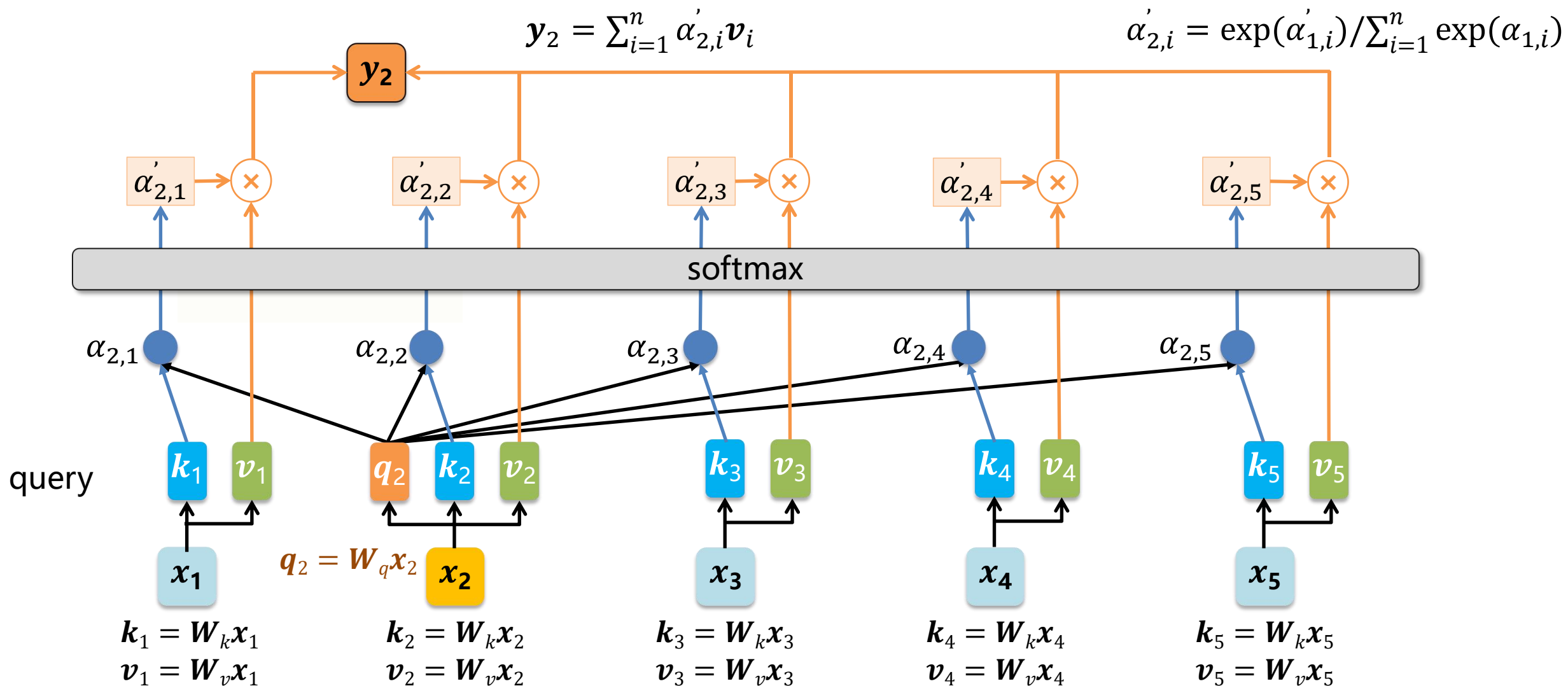
自注意力机制

自注意力的数学表达



自注意力机制

自注意力的数学表达

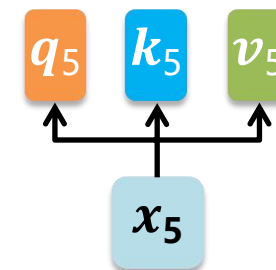
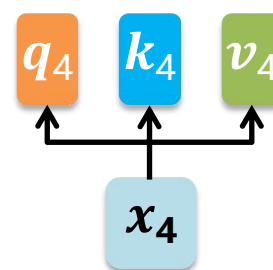
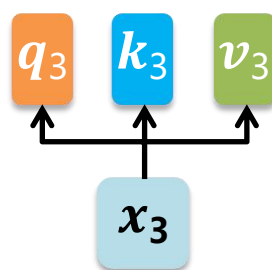
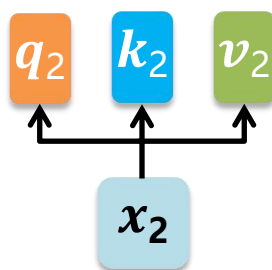
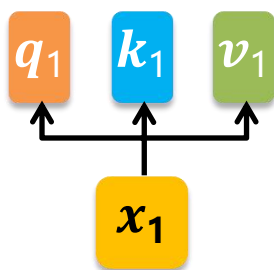


自注意力的矩阵表达

$$q_i = W_q x_i \quad \begin{matrix} W_q & x_1 & x_2 & x_3 & x_4 & x_5 \\ & \mathbf{X} & & & & \end{matrix} = \begin{matrix} q_1 & q_2 & q_3 & q_4 & q_5 \\ & \mathbf{Q} & & & \end{matrix}$$

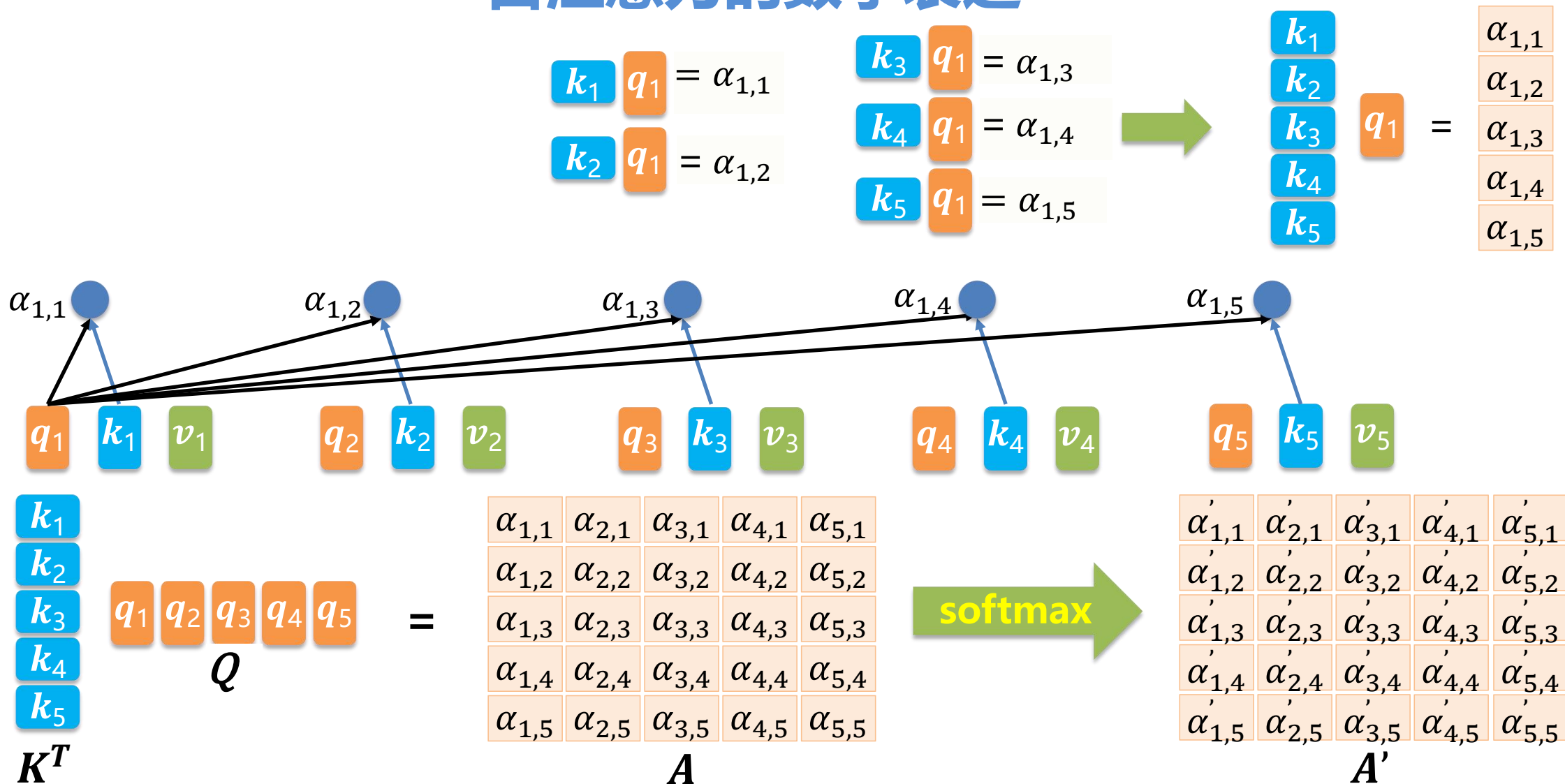
$$k_i = W_k x_i \quad \begin{matrix} W_k & x_1 & x_2 & x_3 & x_4 & x_5 \\ & \mathbf{X} & & & & \end{matrix} = \begin{matrix} k_1 & k_2 & k_3 & k_4 & k_5 \\ & \mathbf{K} & & & \end{matrix}$$

$$v_i = W_v x_i \quad \begin{matrix} W_v & x_1 & x_2 & x_3 & x_4 & x_5 \\ & \mathbf{X} & & & & \end{matrix} = \begin{matrix} v_1 & v_2 & v_3 & v_4 & v_5 \\ & \mathbf{V} & & & \end{matrix}$$



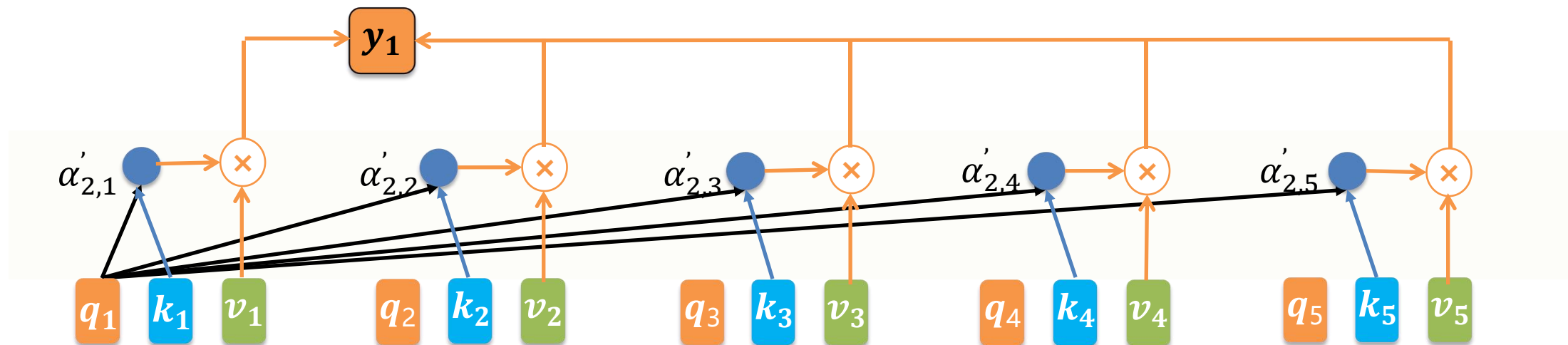
自注意力机制

自注意力的数学表达



自注意力机制

自注意力的数学表达



$$\begin{bmatrix} y_1 & y_2 & y_3 & y_4 & y_5 \end{bmatrix}$$

Y

=

$$\begin{bmatrix} v_1 & v_2 & v_3 & v_4 & v_5 \end{bmatrix}$$

V

$\alpha'_{1,1}$	$\alpha'_{2,1}$	$\alpha'_{3,1}$	$\alpha'_{4,1}$	$\alpha'_{5,1}$
$\alpha'_{1,2}$	$\alpha'_{2,2}$	$\alpha'_{3,2}$	$\alpha'_{4,2}$	$\alpha'_{5,2}$
$\alpha'_{1,3}$	$\alpha'_{2,3}$	$\alpha'_{3,3}$	$\alpha'_{4,3}$	$\alpha'_{5,3}$
$\alpha'_{1,4}$	$\alpha'_{2,4}$	$\alpha'_{3,4}$	$\alpha'_{4,4}$	$\alpha'_{5,4}$
$\alpha'_{1,5}$	$\alpha'_{2,5}$	$\alpha'_{3,5}$	$\alpha'_{4,5}$	$\alpha'_{5,5}$

A'

自注意力机制

自注意力的数学表达

$$\begin{aligned} W_q X &= Q \\ W_k X &= K \\ W_v X &= V \end{aligned}$$

$$K^T Q = A \rightarrow A'$$

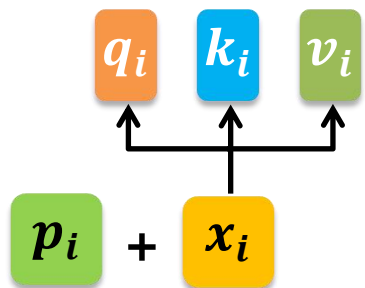
$$V A' = Y$$



位置编码 (position encoding)

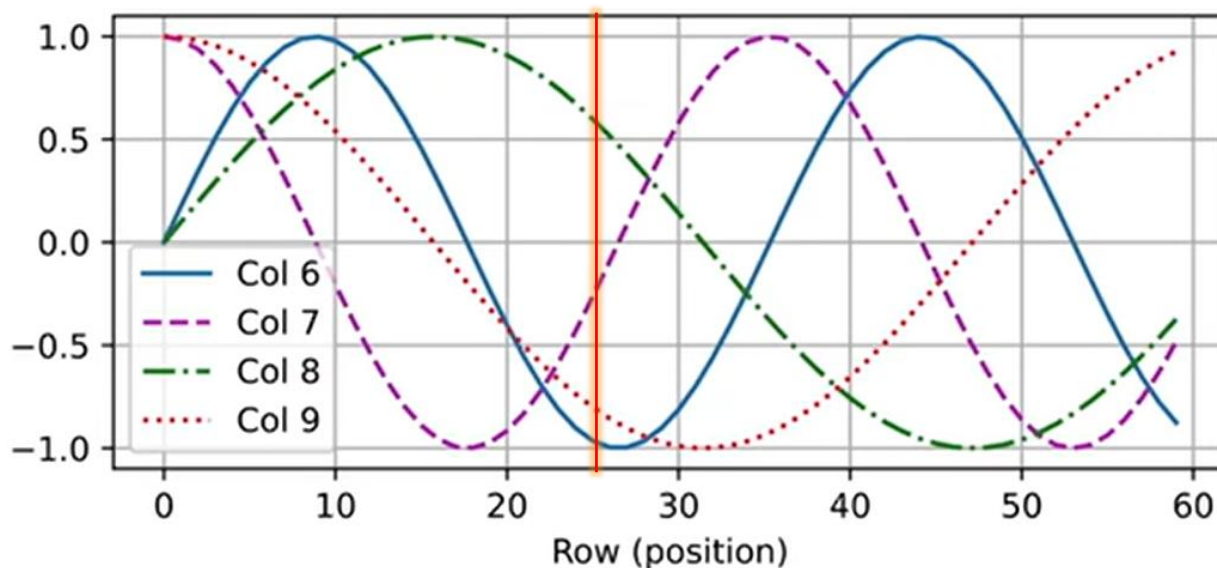
位置编码 (position encoding)

位置编码矩阵



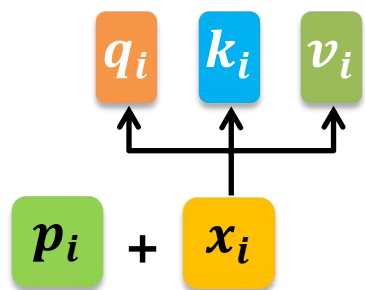
- 假设 $X \in \mathbb{R}^{n \times d}$ 是长为 n 的序列, 则可以使用位置编码矩阵 $P \in \mathbb{R}^{n \times d}$ 来将位置信息 p_i 手动注入到输入 X 中。 P 的计算公式如下:

$$p_{i,2j} = \sin\left(\frac{i}{10000^{2j/d}}\right), \quad p_{i,2j+1} = \cos\left(\frac{i}{10000^{2j/d}}\right)$$



位置编码 (position encoding)

位置编码的相对性



- 已知位置 i 处的位置编码，可以通过一个投影矩阵来获取 $i + \delta$ 处的位置编码。
- $i + \delta$ 处的位置编码与它在序列中的实际位置无关。
- 设 $\omega_j = 10000^{2j/d}$ ，则：

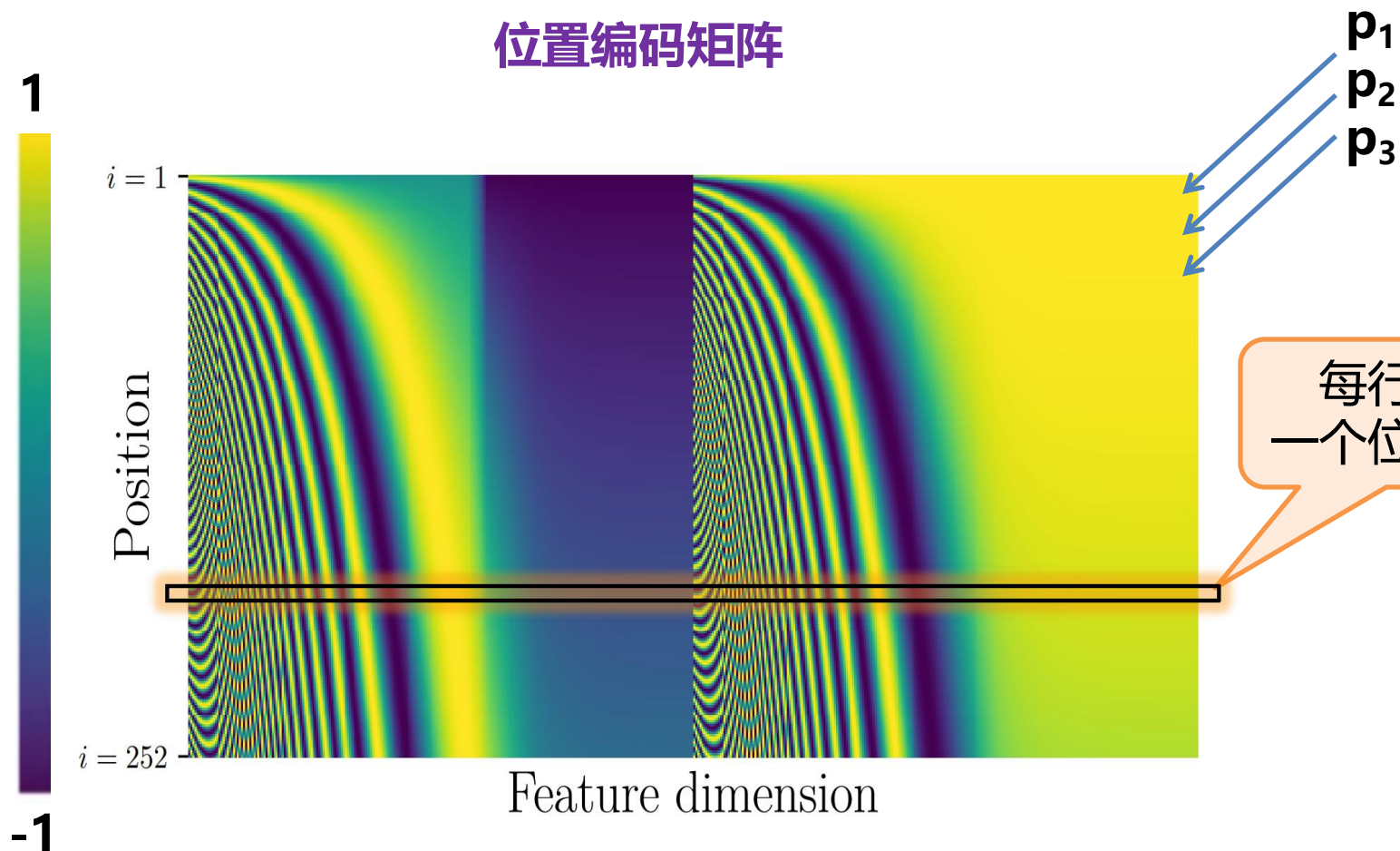
$$\begin{bmatrix} p_{i+\delta, 2j} \\ p_{i+\delta, 2j+1} \end{bmatrix} = \begin{bmatrix} \cos(\delta\omega_j) & \sin(\delta\omega_j) \\ -\sin(\delta\omega_j) & \cos(\delta\omega_j) \end{bmatrix} \begin{bmatrix} p_{i, 2j} \\ p_{i, 2j+1} \end{bmatrix}$$

投影矩阵

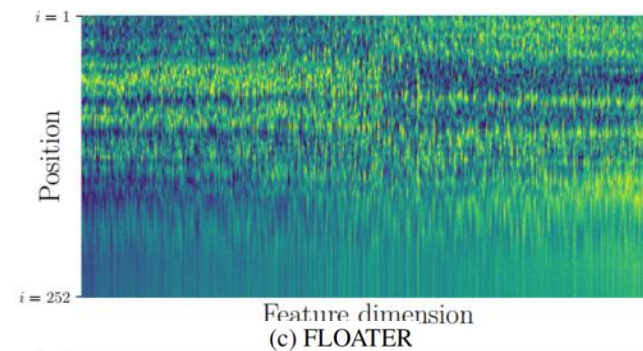
位置编码 (position encoding)

绝对位置信息

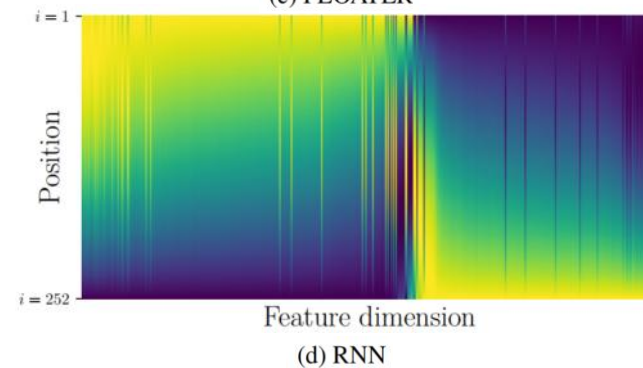
位置编码矩阵



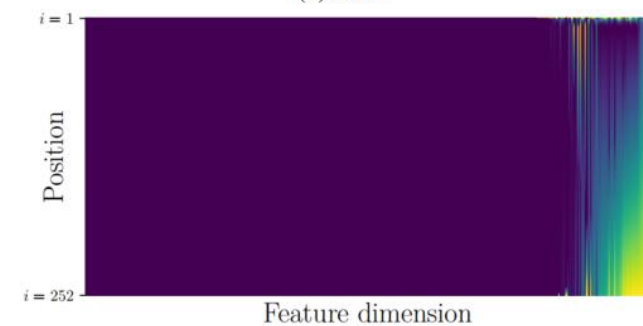
(b) Position embedding



(c) FLOATER



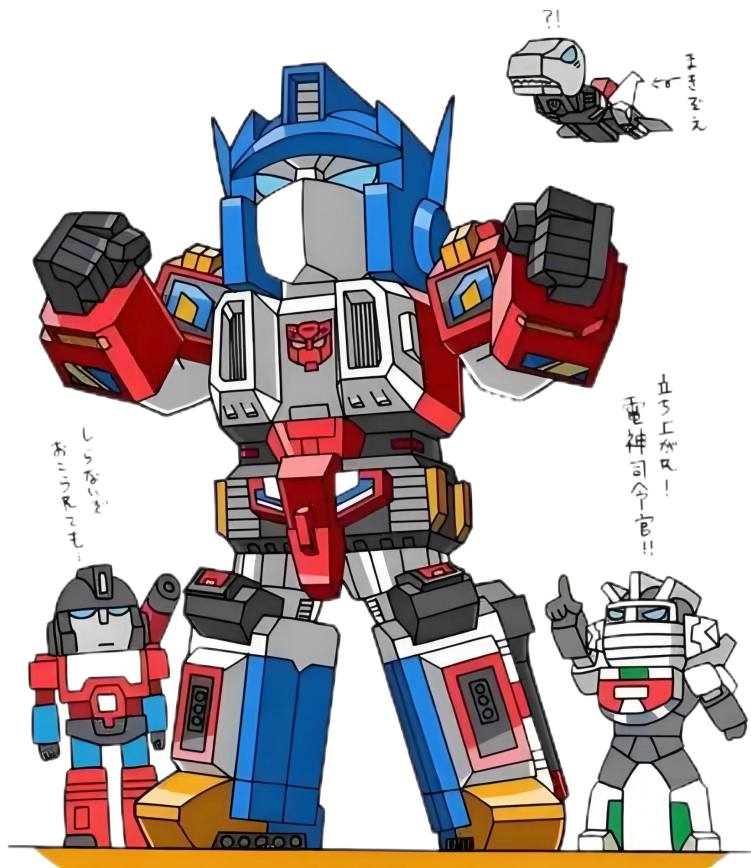
(d) RNN





自注意机制的应用

自注意机制的应用



Transformer

<http://arxiv.org/abs/1706.03762>

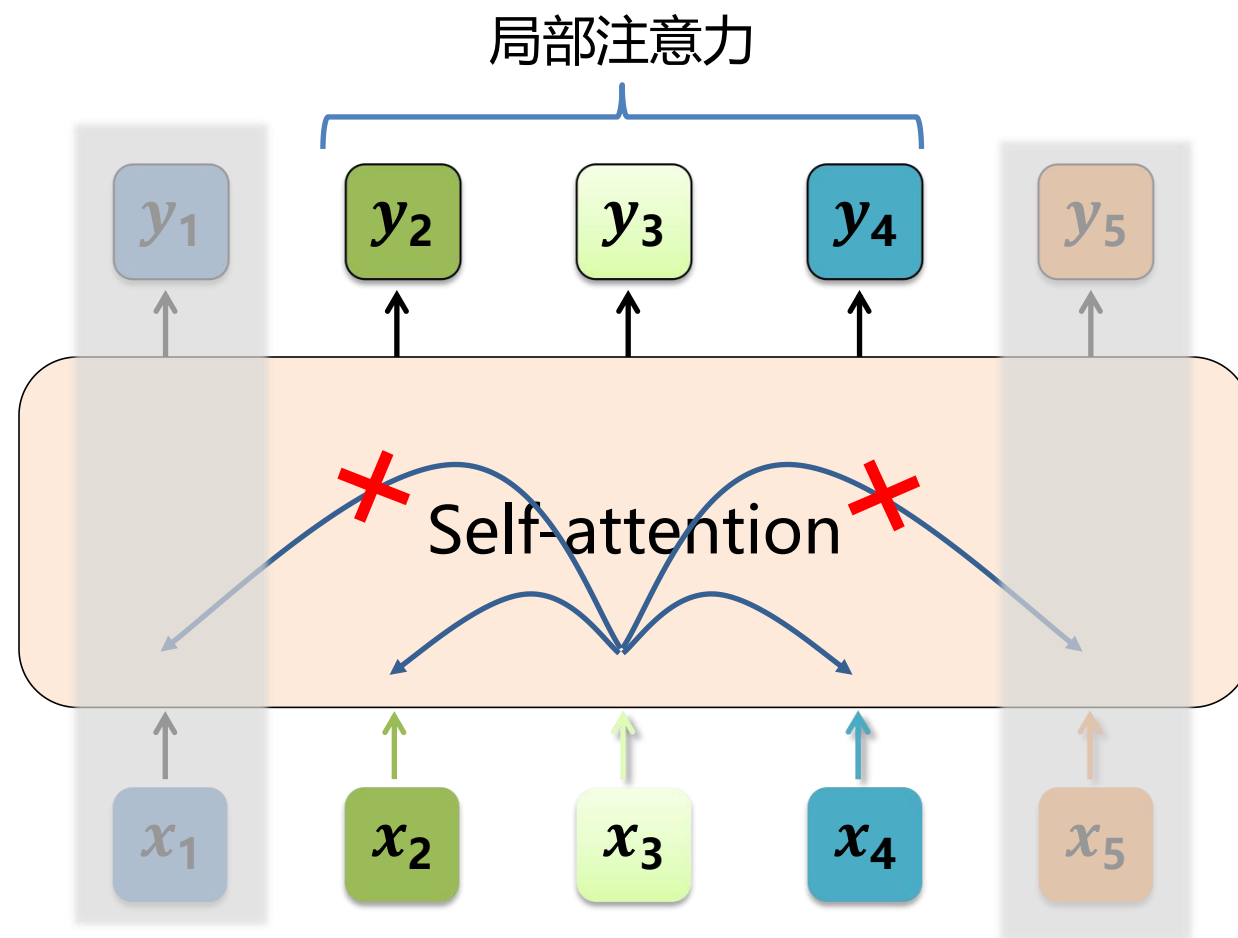
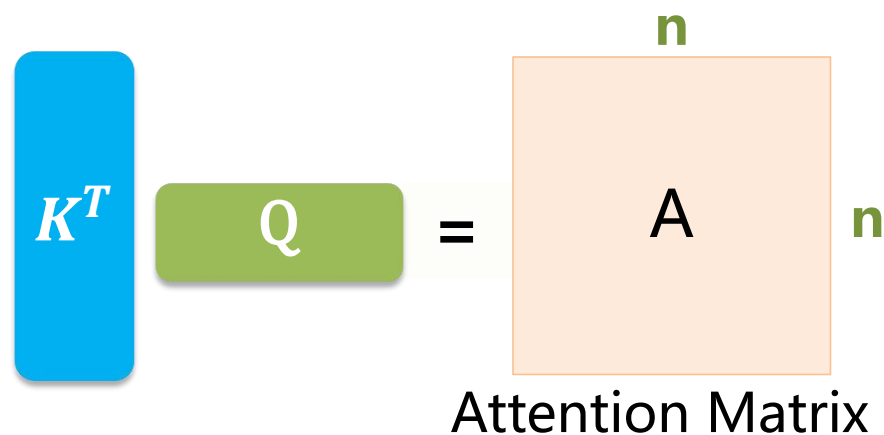
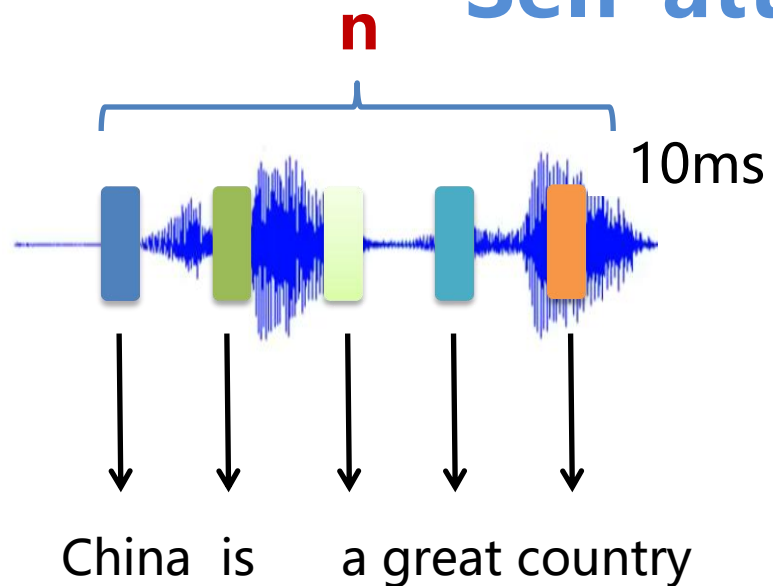


BERT

<http://arxiv.org/abs/1810.04805>

自注意机制的应用

Self-attention for Speech

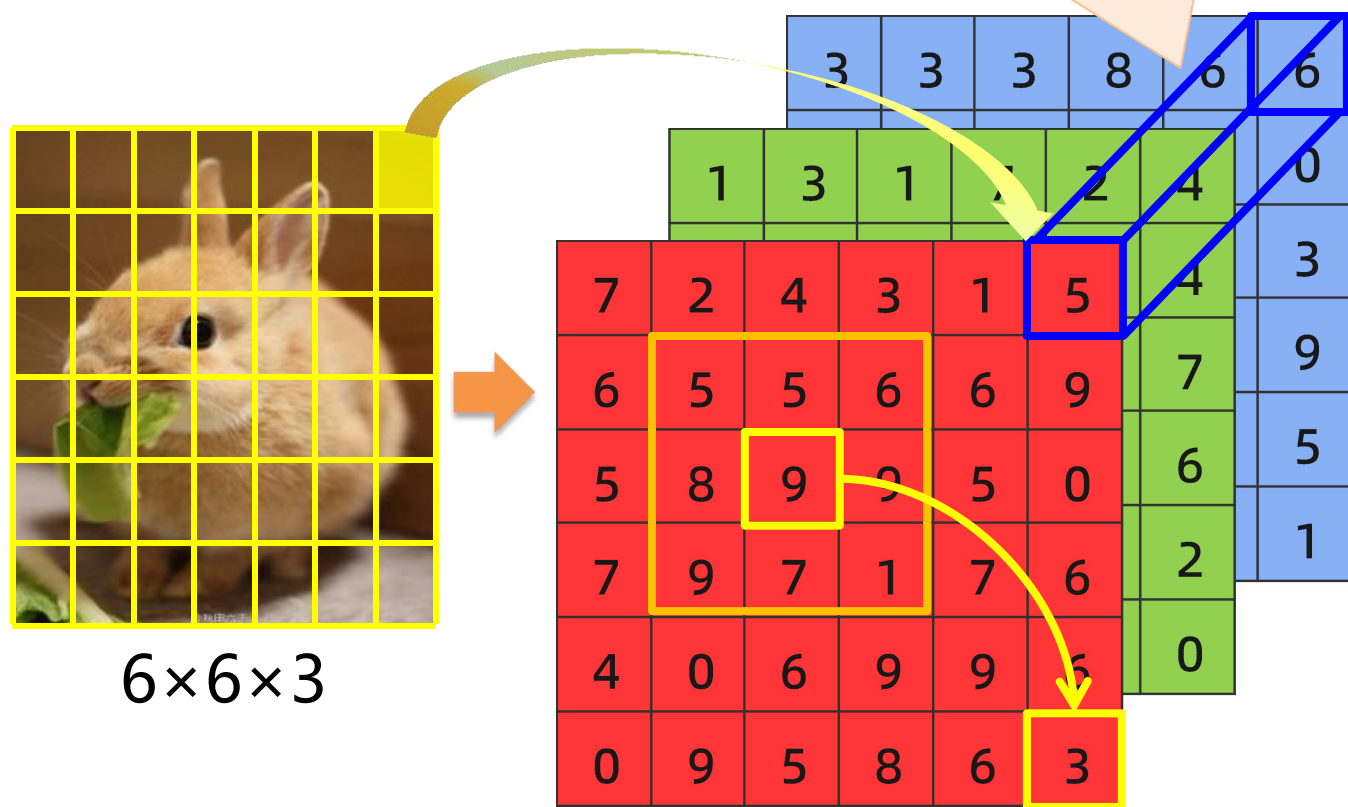


Truncated self-attention

自注意机制的应用

Self-attention for Image

一个像素位置的多个通道组成一个向量



- CNN可以被看作是一个 receptive field 中的局部 self-attention。

CNN是一个简化版的 self-attention

- Self-attention也可以被看作是一个包含可学习参数的CNN。

Self-attention是一个复杂版的CNN

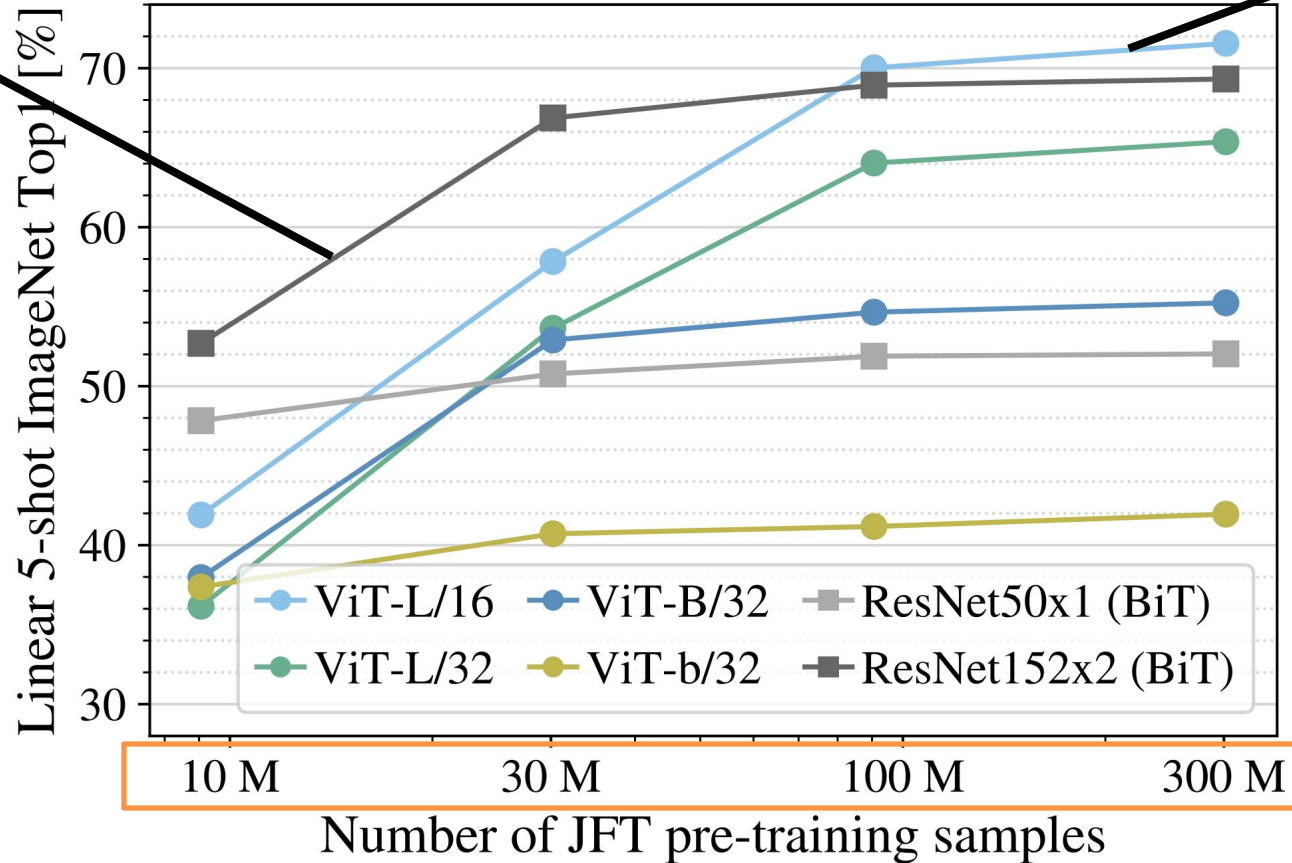


自注意机制的应用

Self-attention for Image

CNN

数据较少时也有较好的性能



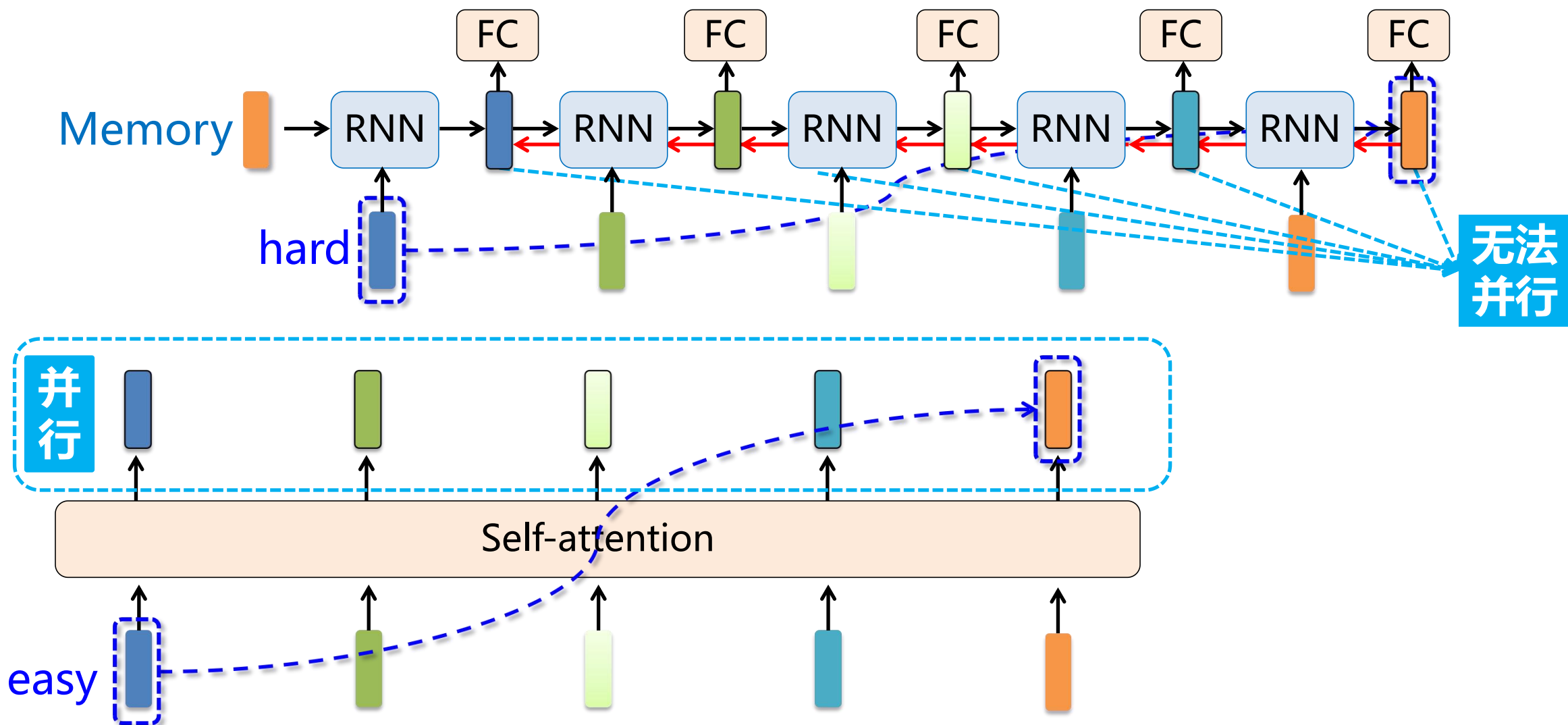
Self-attention

较多的数据能带来性能的提升

[arXiv2010] An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale

自注意机制的应用

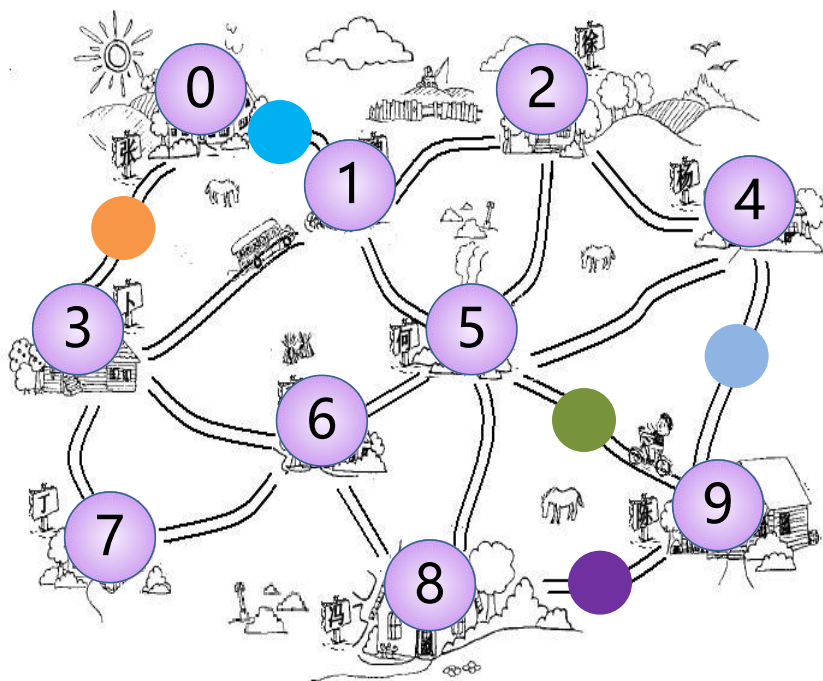
Self-attention vs RNN



自注意机制的应用

Self-attention for Graph

邻接矩阵 *Attention Matrix*



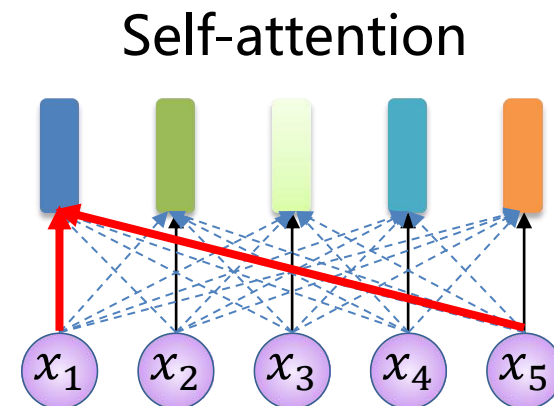
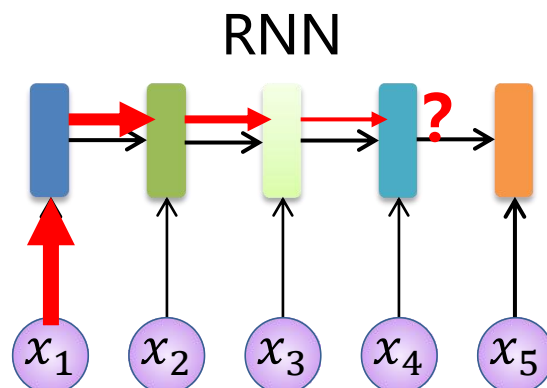
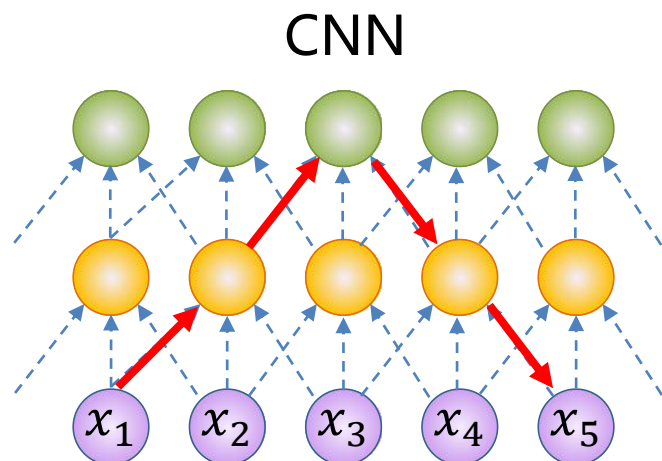
邻接结点，既包含结点，也包含边

	v_0	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
v_0	0	1	0	1	0	0	0	0	0	0
v_1	1	0	1	1	0	1	0	0	0	0
v_2	0	1	0	0	1	1	0	0	0	0
v_3	1	1	0	0	0	0	1	1	0	0
v_4	0	0	1	0	0	1	0	0	0	1
v_5	0	1	1	0	1	0	1	0	1	1
v_6	0	0	0	1	0	1	0	1	1	0
v_7	0	0	0	1	0	0	1	0	0	0
v_8	0	0	0	0	0	1	1	0	0	1
v_9	0	0	0	0	1	1	0	0	1	0

Graph Neural Network (GNN)

自注意机制的应用

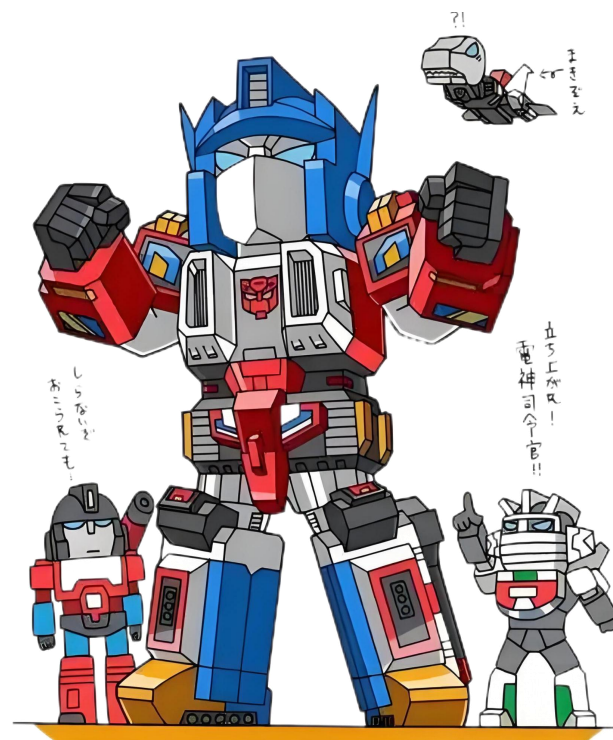
Self-attention vs CNN、RNN

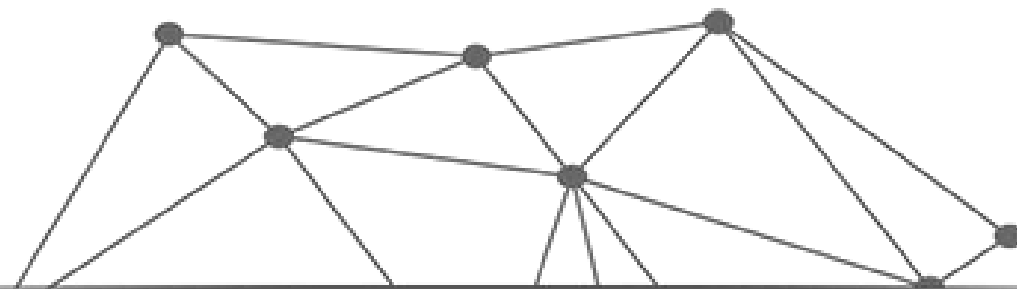


	CNN	RNN	自注意力
计算复杂度	$O(knd^2)$	$O(nd^2)$	$O(dn^2)$
并行度	$O(n)$	$O(1)$	$O(n)$
最长路径	$O(n/k)$	$O(n)$	$O(1)$

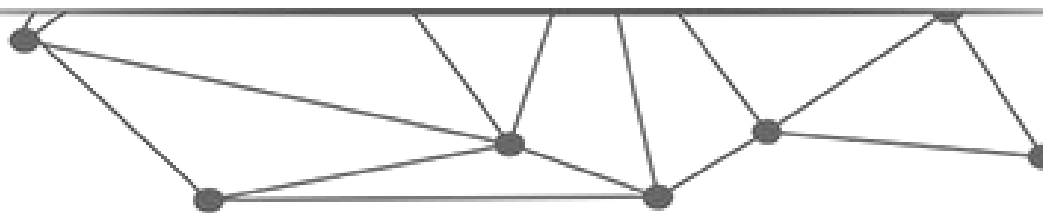
小 结

- 自注意力机制的三重角色映射
 - 输入向量 x_1 同时当作 query, key 和 value 来对序列抽取特征
 - 直接基于序列生成特征表示, 无需额外信息注入
- 全局感知与计算效率的权衡
 - 每个输出向量都具有完整序列上下文的感知能力
 - 完全并行化计算架构实现高效训练
 - 平方级计算复杂度($O(n^2)$)成为长序列处理的主要瓶颈
- 位置编码的时序注入机制
 - 将位置信息直接注入输入向量, 确保自注意力的固有排列不变性
 - Transformer 采用的 sin-cos 位置编码





课堂互动 9.2



读万卷书 行万里路 只为最好的修炼



QQ: 14777591 (宇宙骑士)

Email: ouxinyu@alumni.hust.edu.cn

Website: <http://ouxinyu.cn>

Tel: 18687840023

地址: 安宁校区 诚远楼201

南院 智能应用研究院A306-2