Natural Language Processing

第七周 自注意力机制

庞彦

yanpang@gzhu.edu.cn

Spring 2023

Overview

CONTENTS

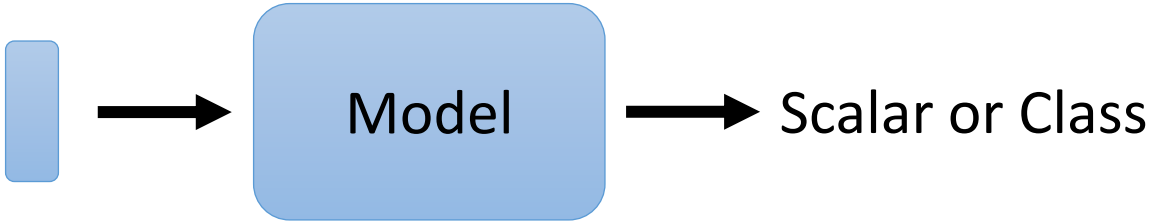
01 Self-attention Mechanism

自注意力机制

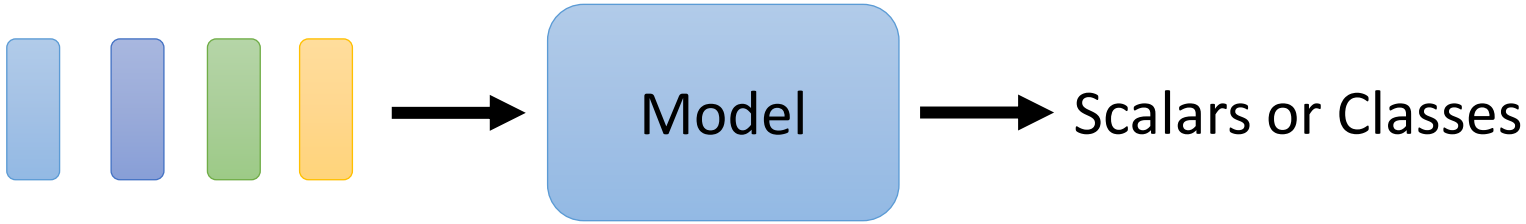
Spring 2023

Sophisticated Input

- Input is **a vector** 输入是一个矢量

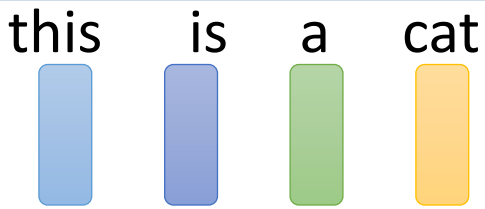


- Input is **a set of vectors**输入是一组矢量



(may change length)
长度可变

Vector Set as Input



One-hot Encoding

apple = [1 0 0 0 0 ……]

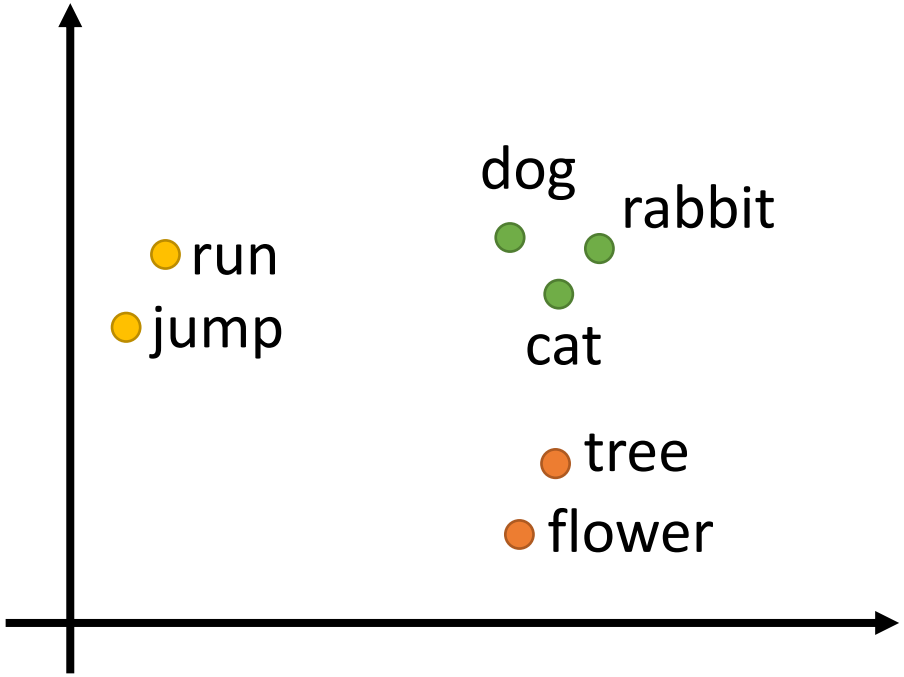
bag = [0 1 0 0 0 ……]

cat = [0 0 1 0 0 ……]

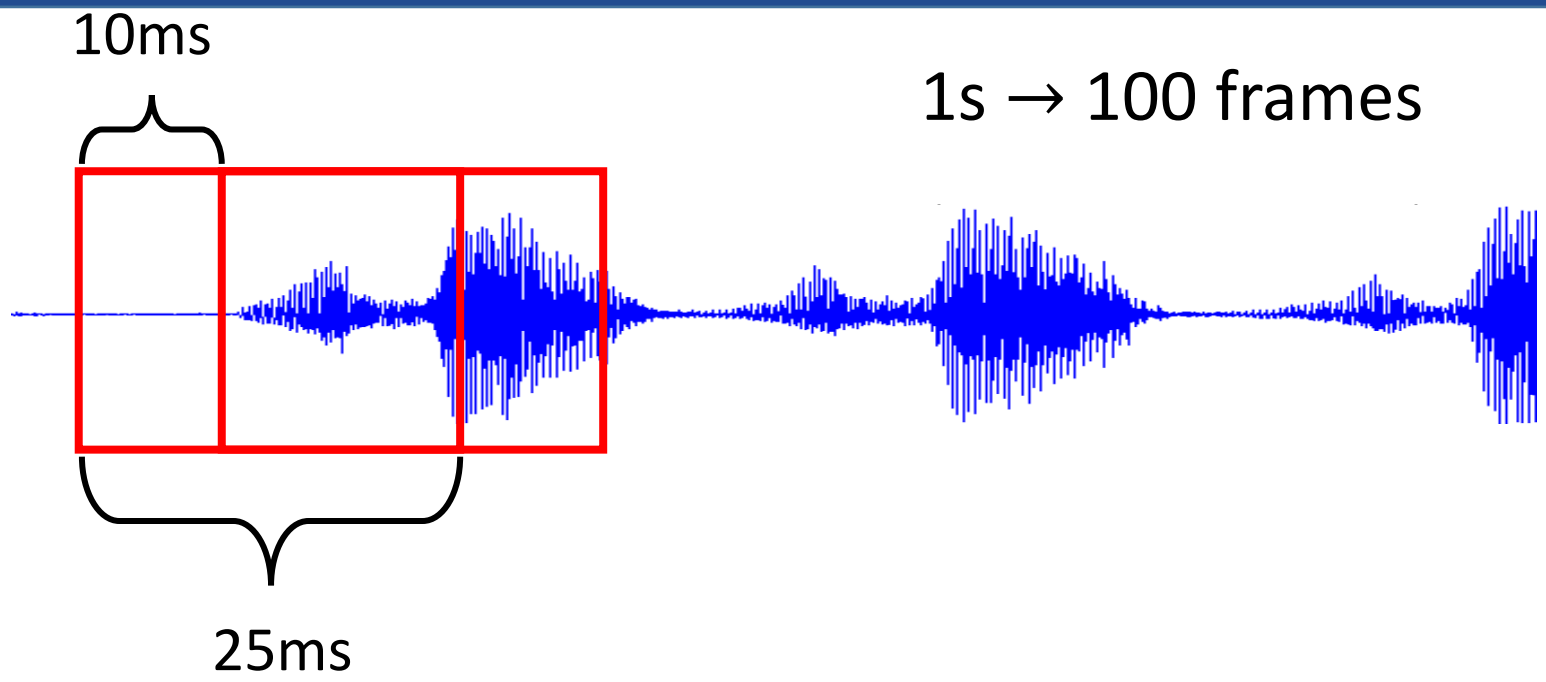
dog = [0 0 0 1 0 ……]

elephant = [0 0 0 0 1 ……]

Word Embedding



Vector Set as Input



1s → 100 frames

frame

400 sample points (16KHz)
39-dim MFCC
80-dim filter bank output

Vector Set as Input

Graph is also a set of vectors (consider each **node** as **a vector**)

图也是一组矢量

Each profile is a vector

每个人都是一个矢量

Vector Set as Input

Graph is also a set of vectors (consider each **node** as **a vector**)

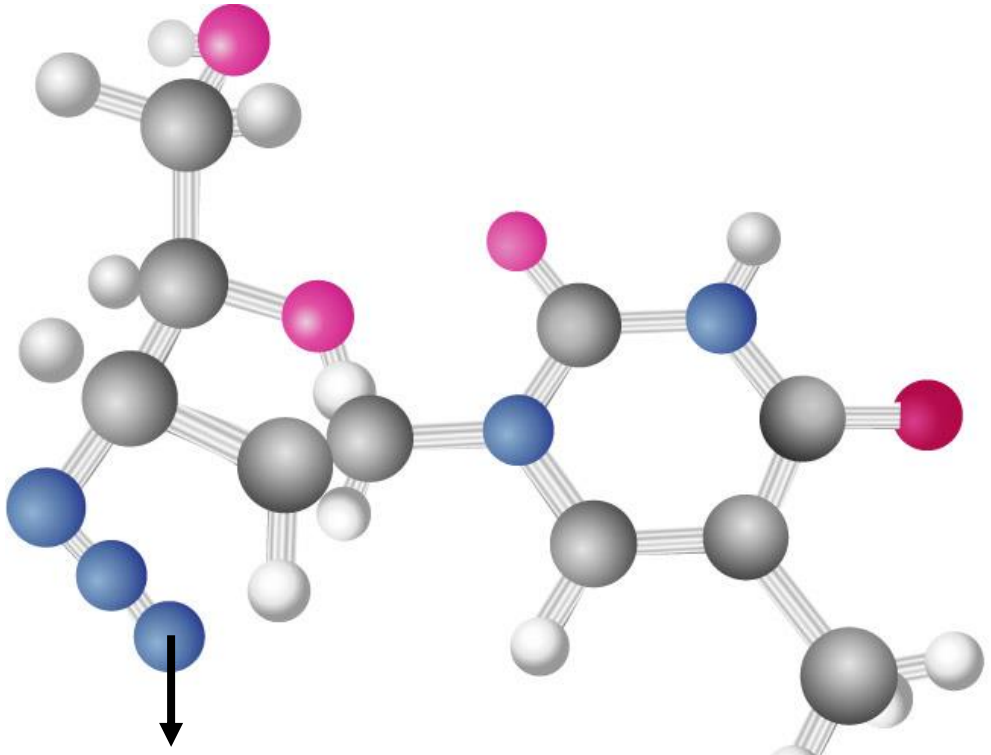
图也是一组矢量

H = [1 0 0 0 0 ……]

C = [0 1 0 0 0 ……]

O = [0 0 1 0 0 ……]

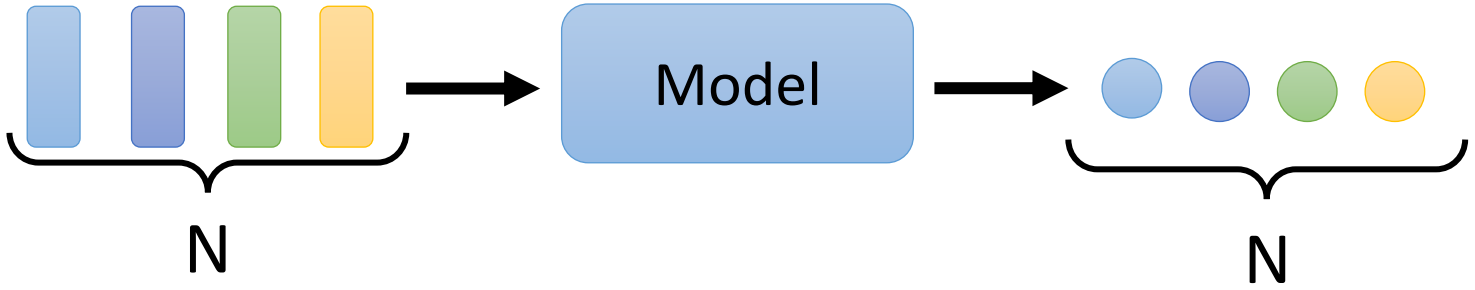
⋮



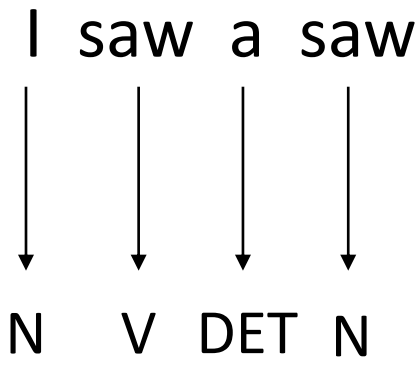
One-hot vector

What is the output?

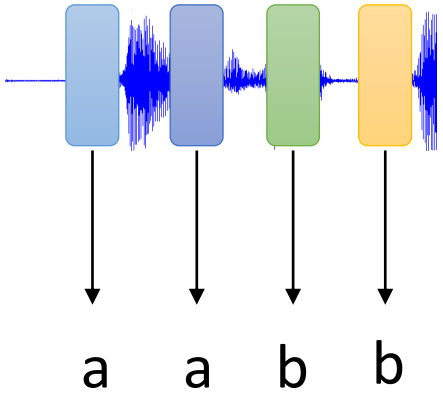
- Each vector has a label. 每个矢量含有一个标签。

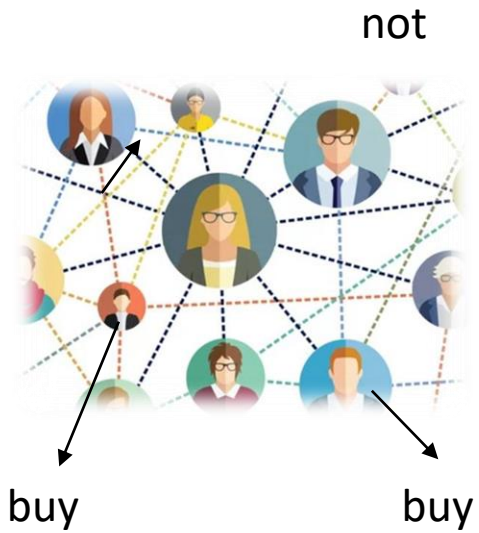


Example Applications



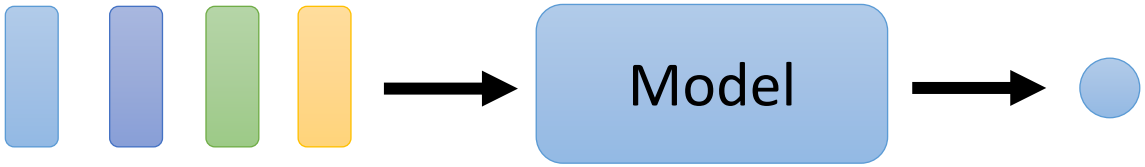
POS tagging





What is the output?

- The whole sequence has a label.整个句子含有一个标签。

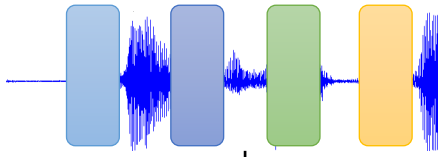


Example Applications

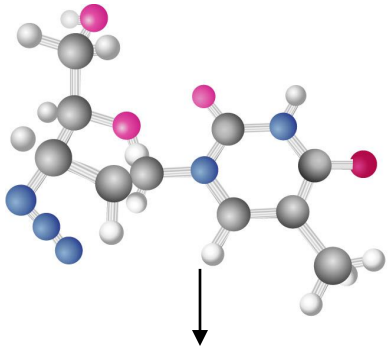
this is good

Sentiment analysis

positive



speaker

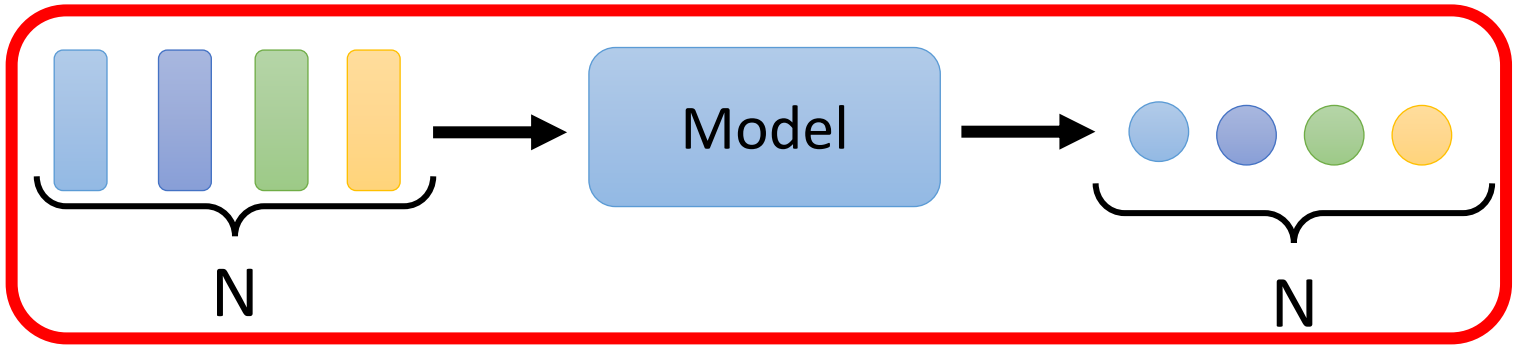


hydrophilicity

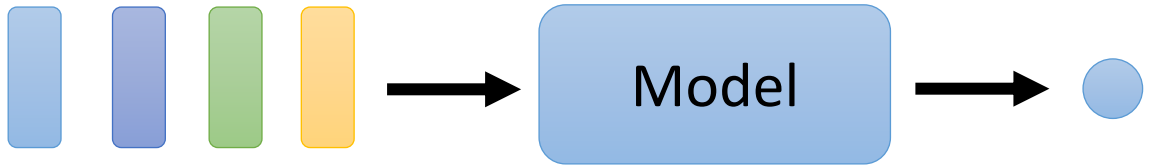
What is the output?

- Each vector has a label.

focus of this lecture



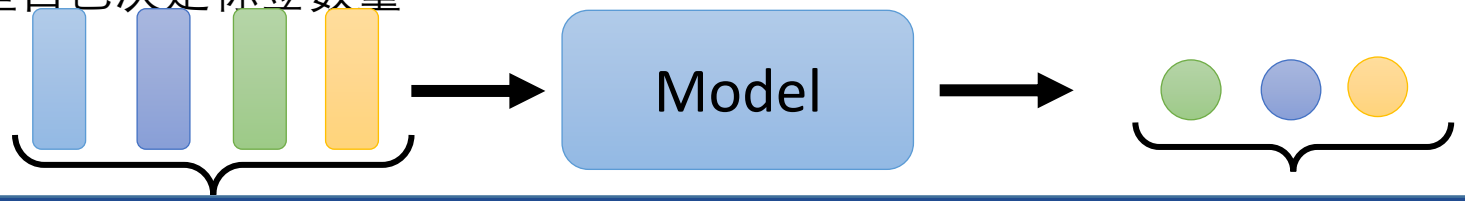
- The whole sequence has a label.



- Model decides the number of labels itself.

seq2seq

模型自己决定标签数量



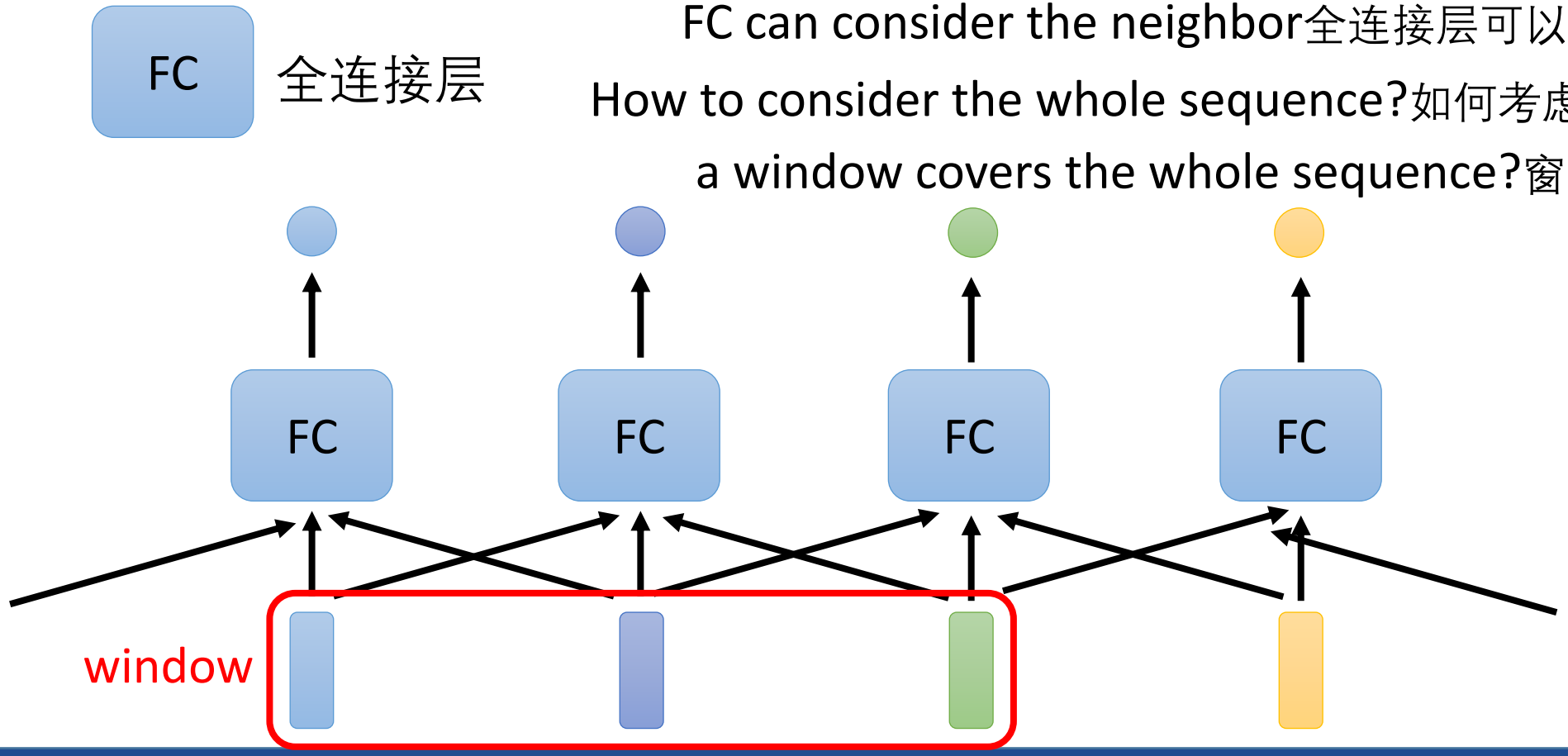
Sequence Labeling

Is it possible to consider the context? 需要考虑内容吗?

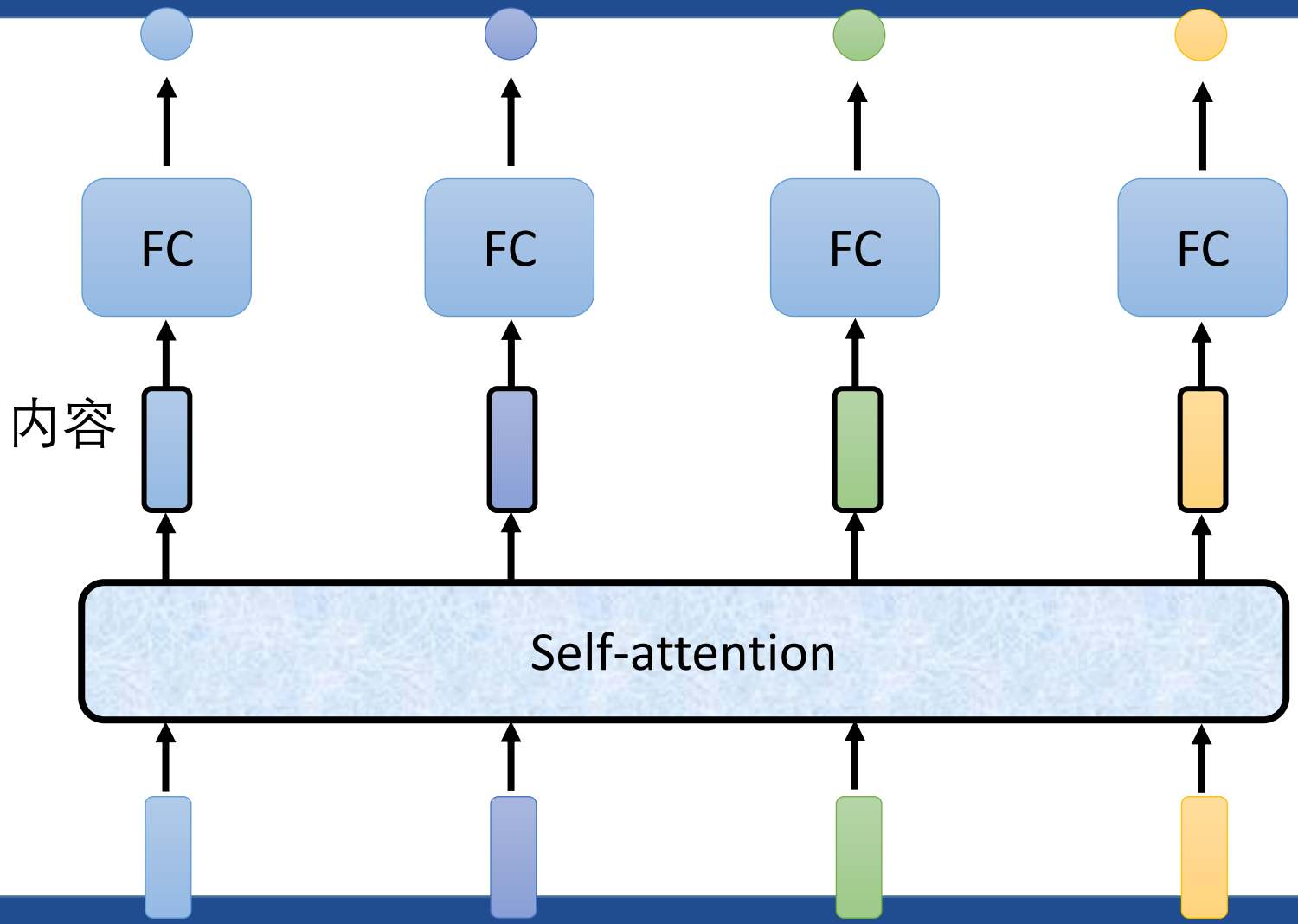
FC can consider the neighbor全连接层可以考虑邻居节点

How to consider the whole sequence?如何考虑整个句子?

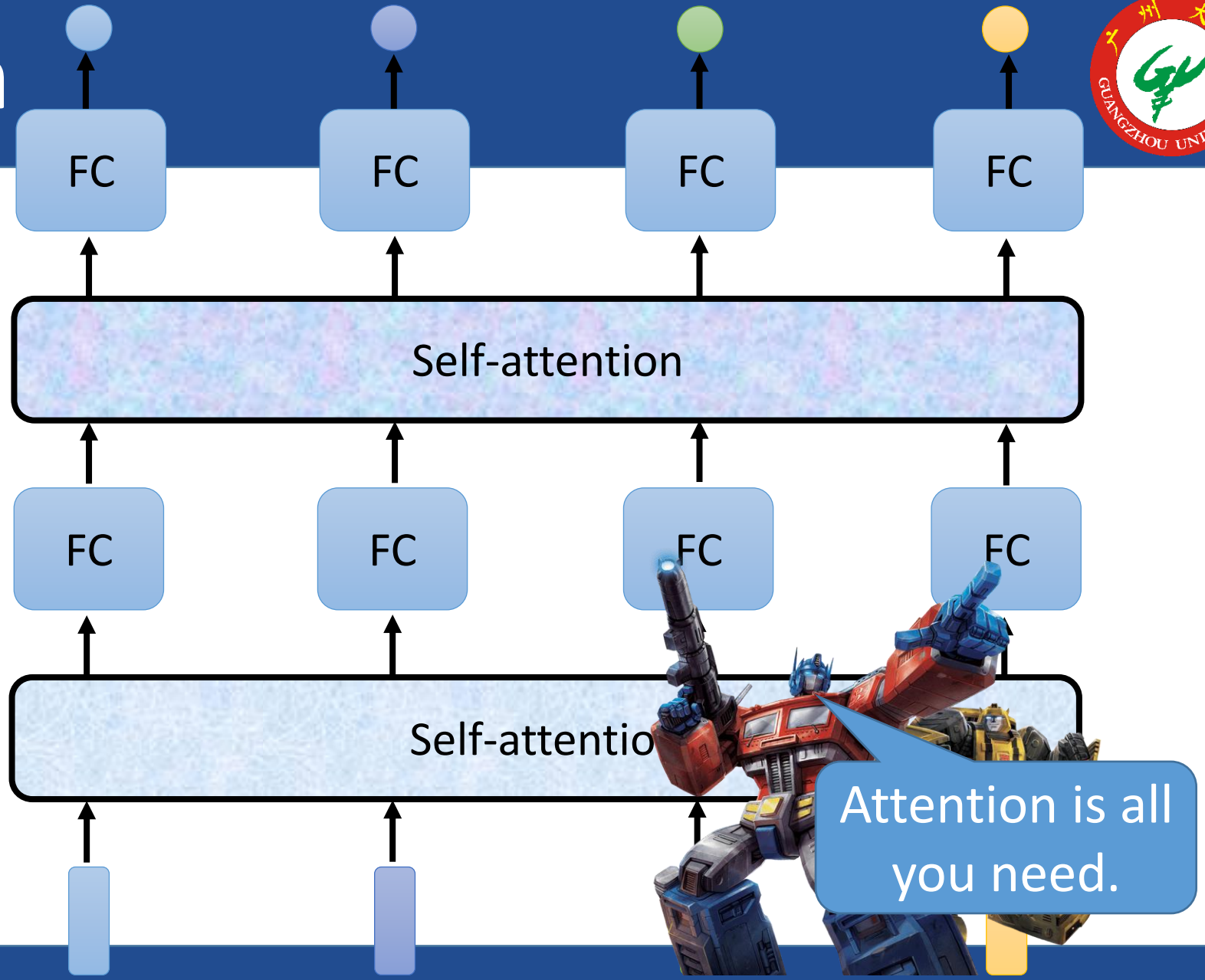
a window covers the whole sequence?窗口



Self-attention

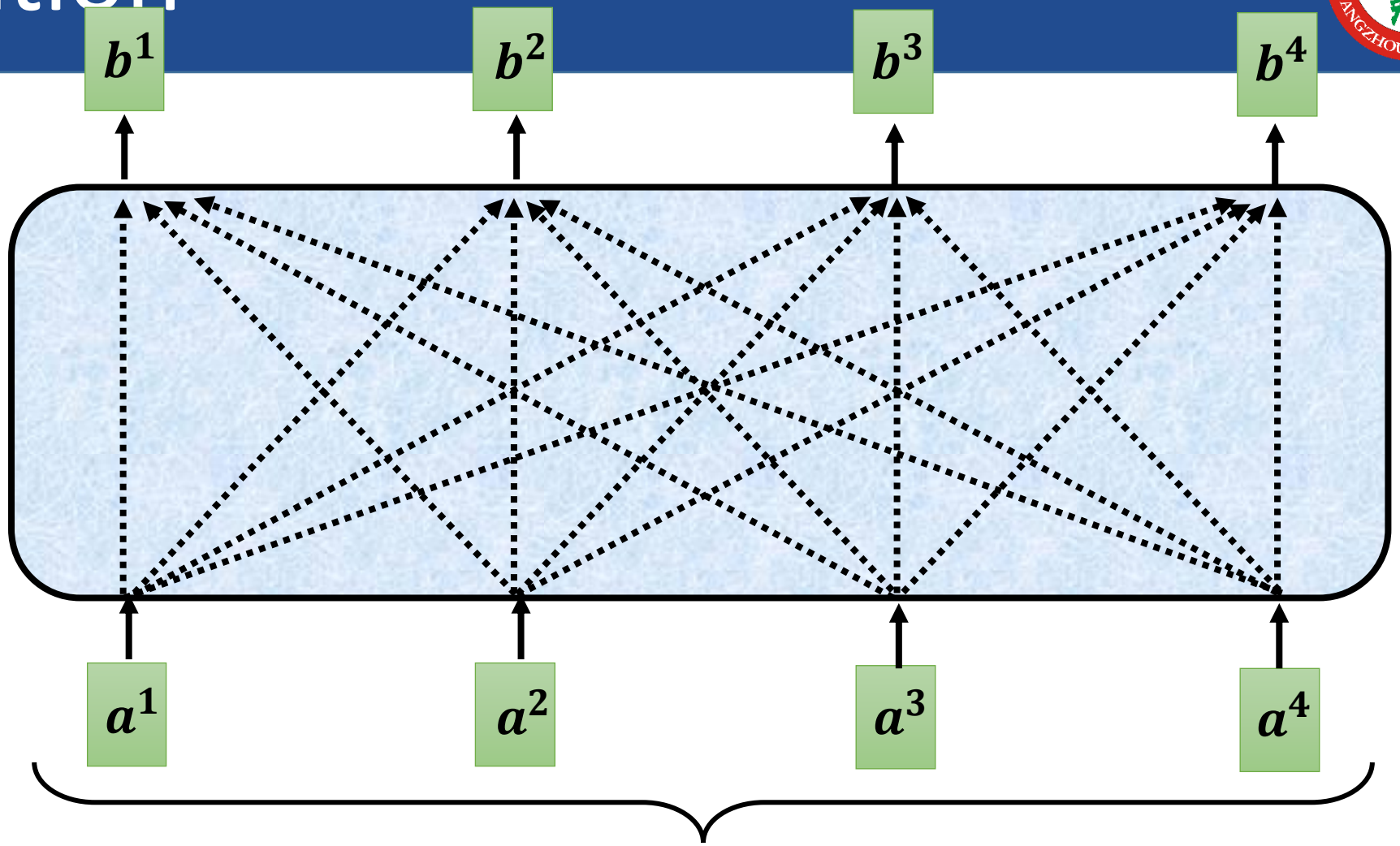


Self-attention



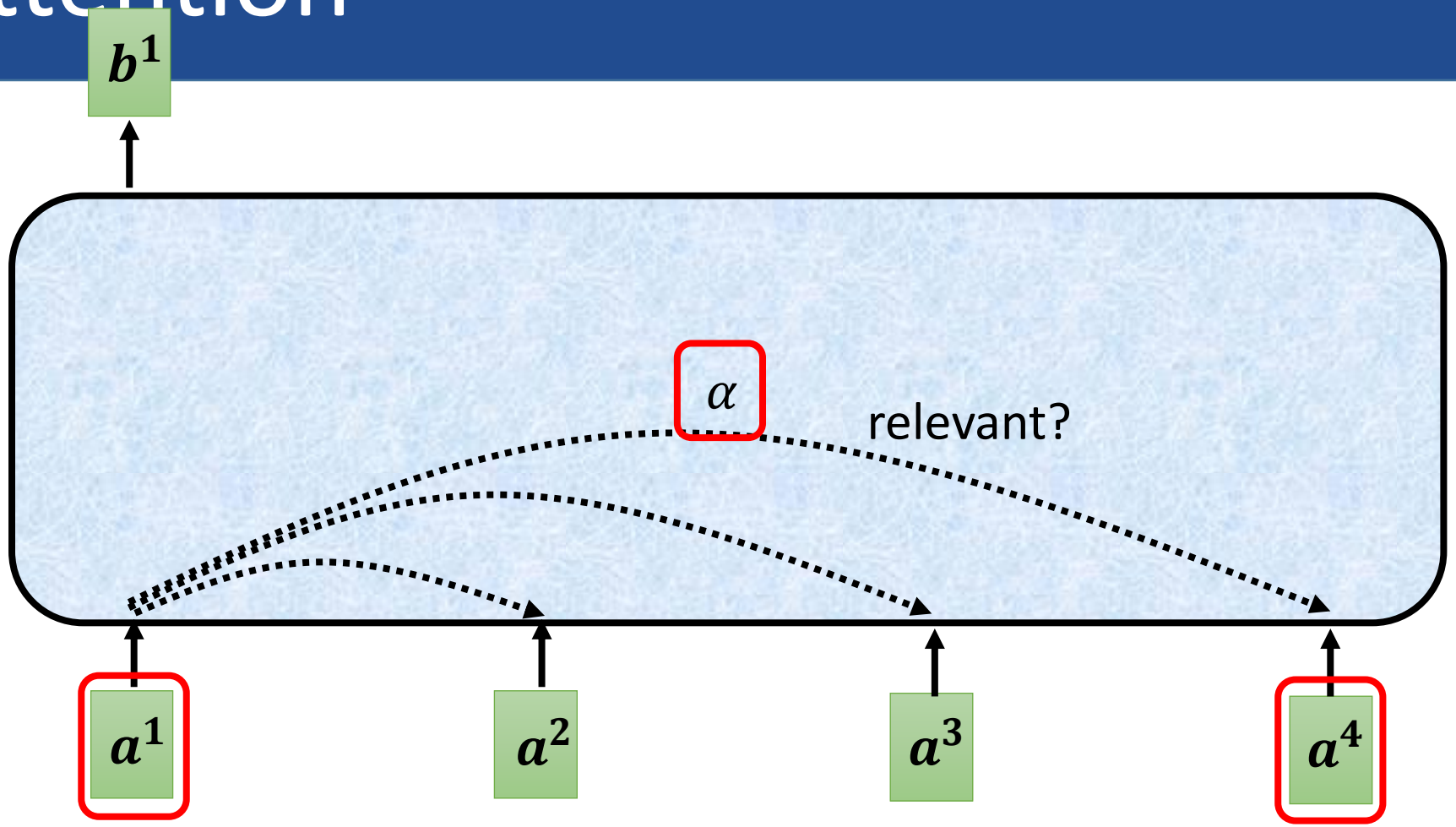
https://arxiv.org/abs/1706.03762

Self-attention



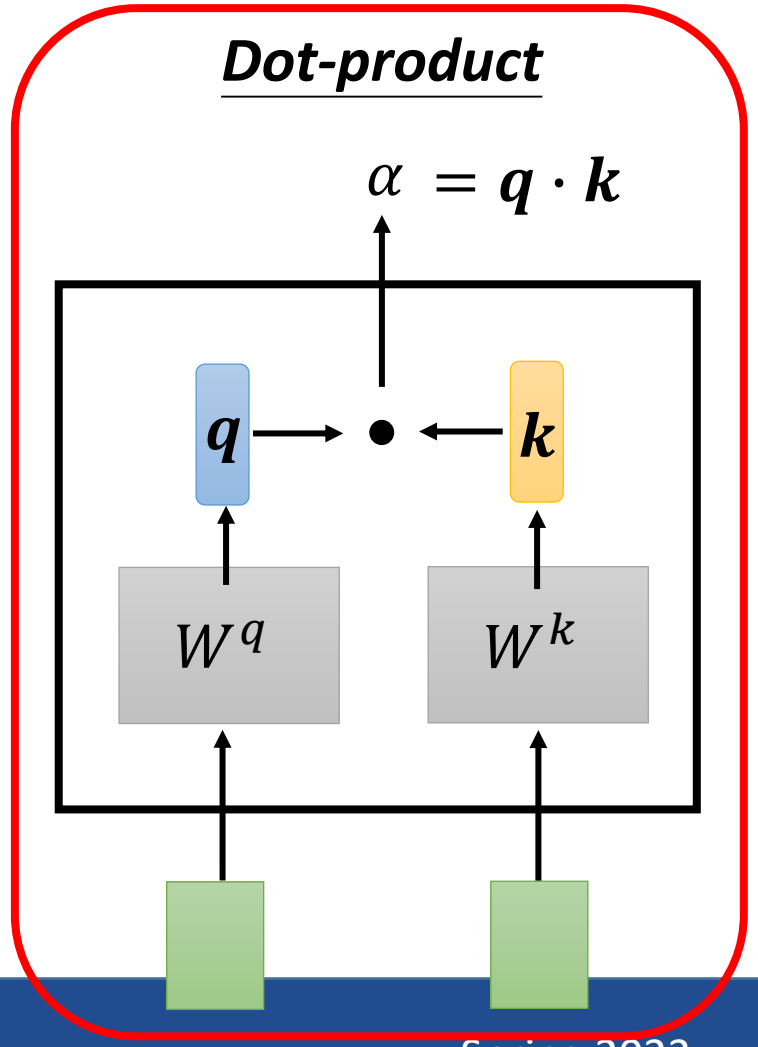
Can be either **input** or **a hidden layer**可为输入或隐藏层

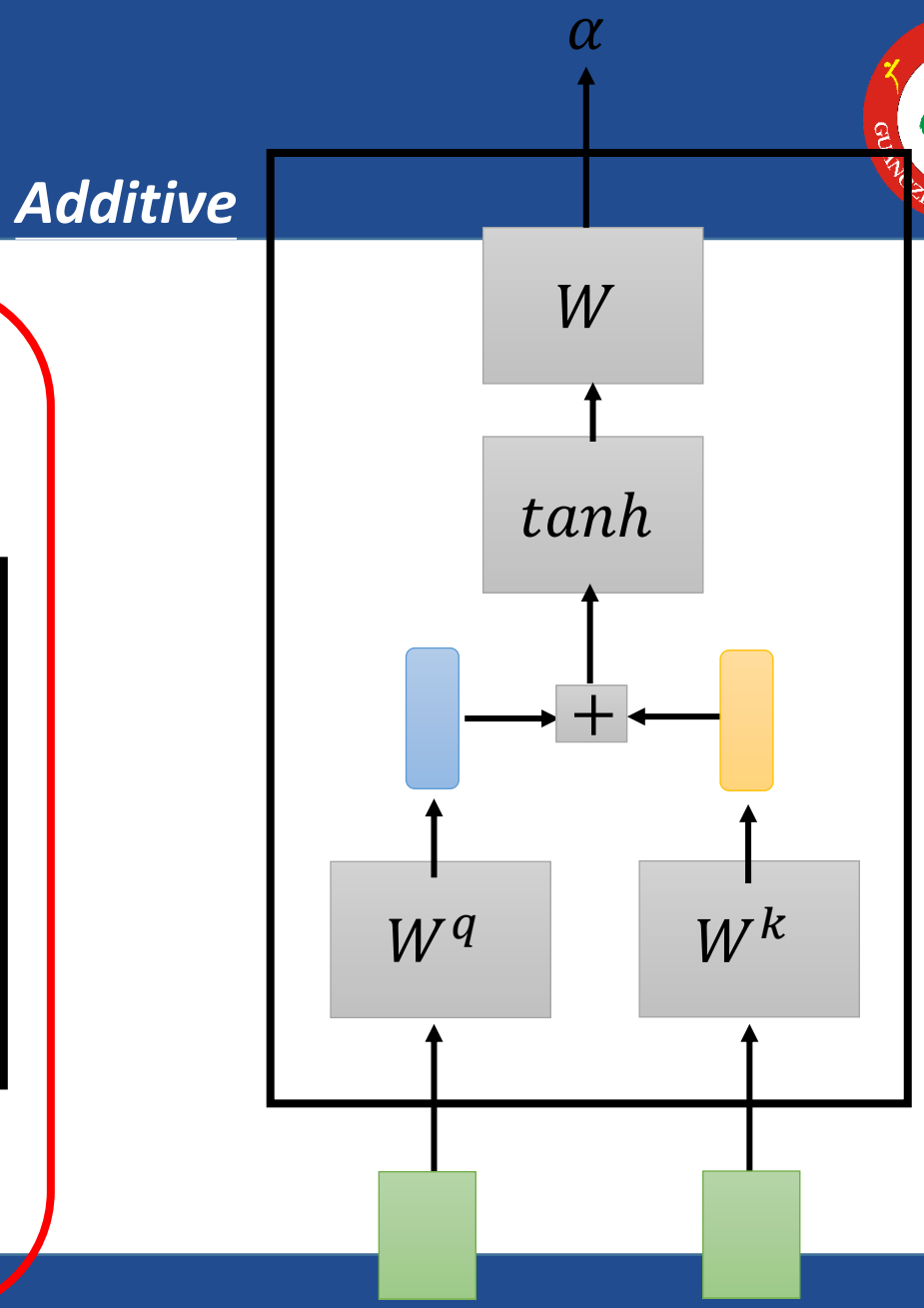
Self-attention



Find the relevant vectors in a sequence找到居中最相关的矢量

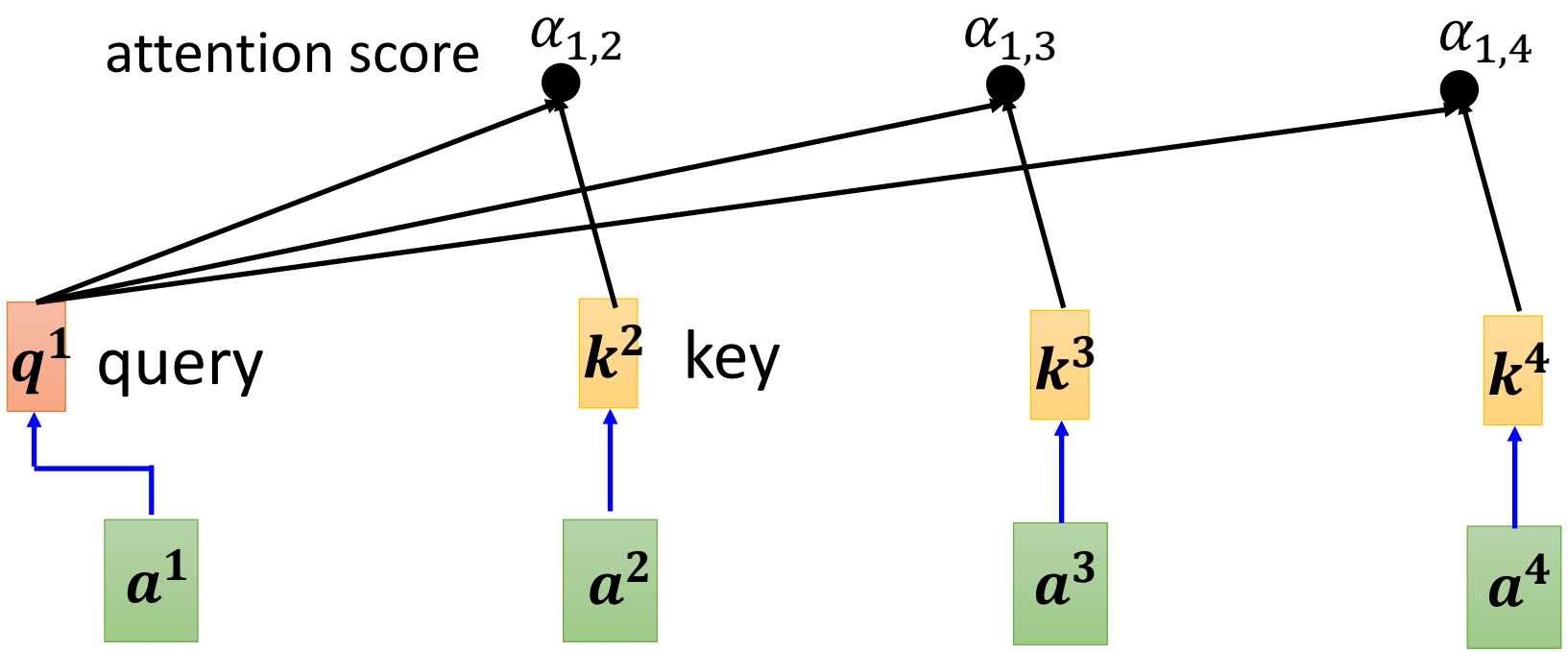
Self-attention





Self-attention

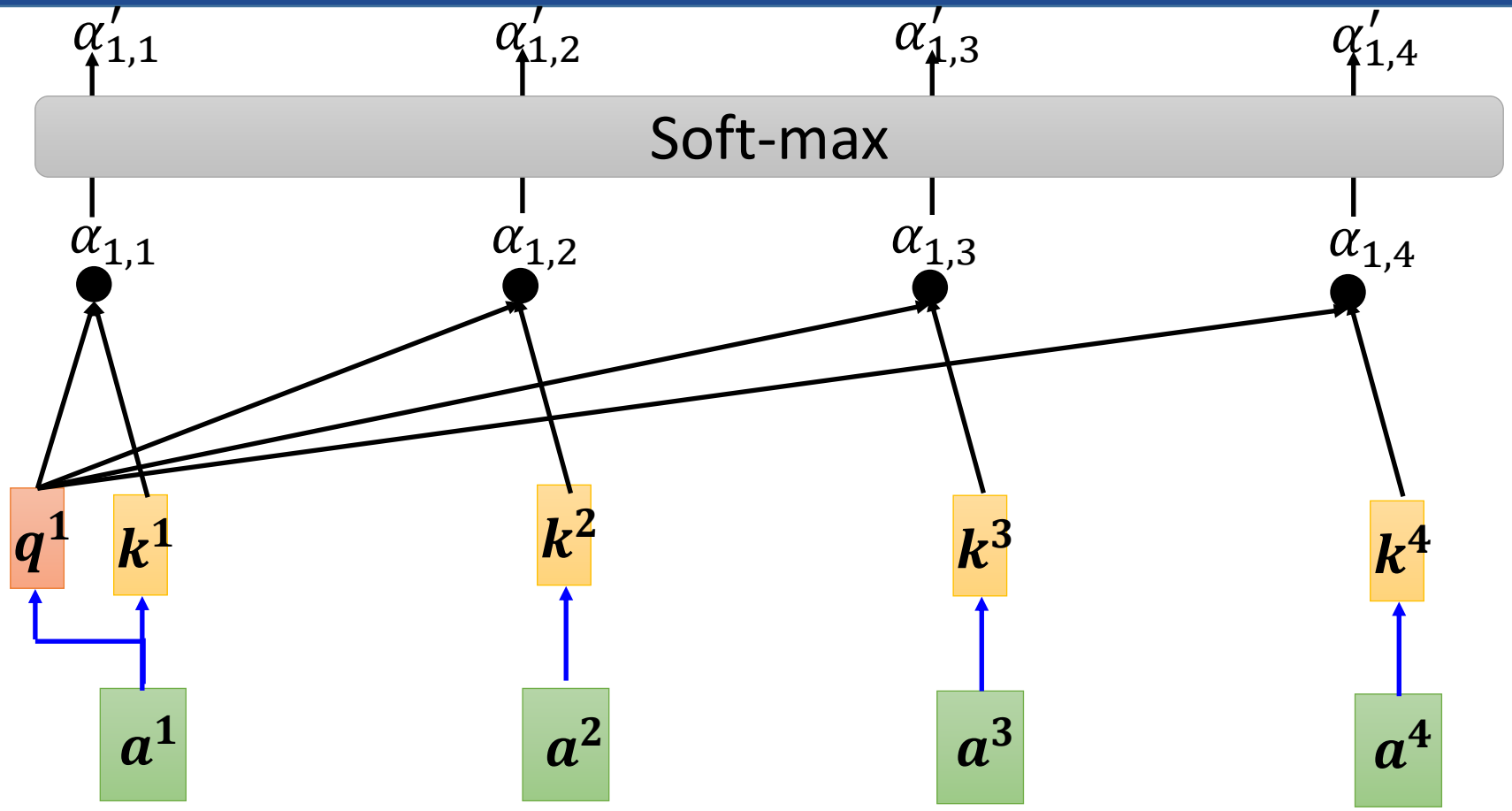
$$\alpha_{1,2} = \boldsymbol{q}^1 \cdot \boldsymbol{k}^2 \qquad \alpha_{1,3} = \boldsymbol{q}^1 \cdot \boldsymbol{k}^3 \qquad \alpha_{1,4} = \boldsymbol{q}^1 \cdot \boldsymbol{k}^4$$



$$\boldsymbol{q}^1 = W^q \boldsymbol{a}^1 \qquad \boldsymbol{k}^2 = W^k \boldsymbol{a}^2 \qquad \boldsymbol{k}^3 = W^k \boldsymbol{a}^3 \qquad \boldsymbol{k}^4 = W^k \boldsymbol{a}^4$$

Self-attention

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



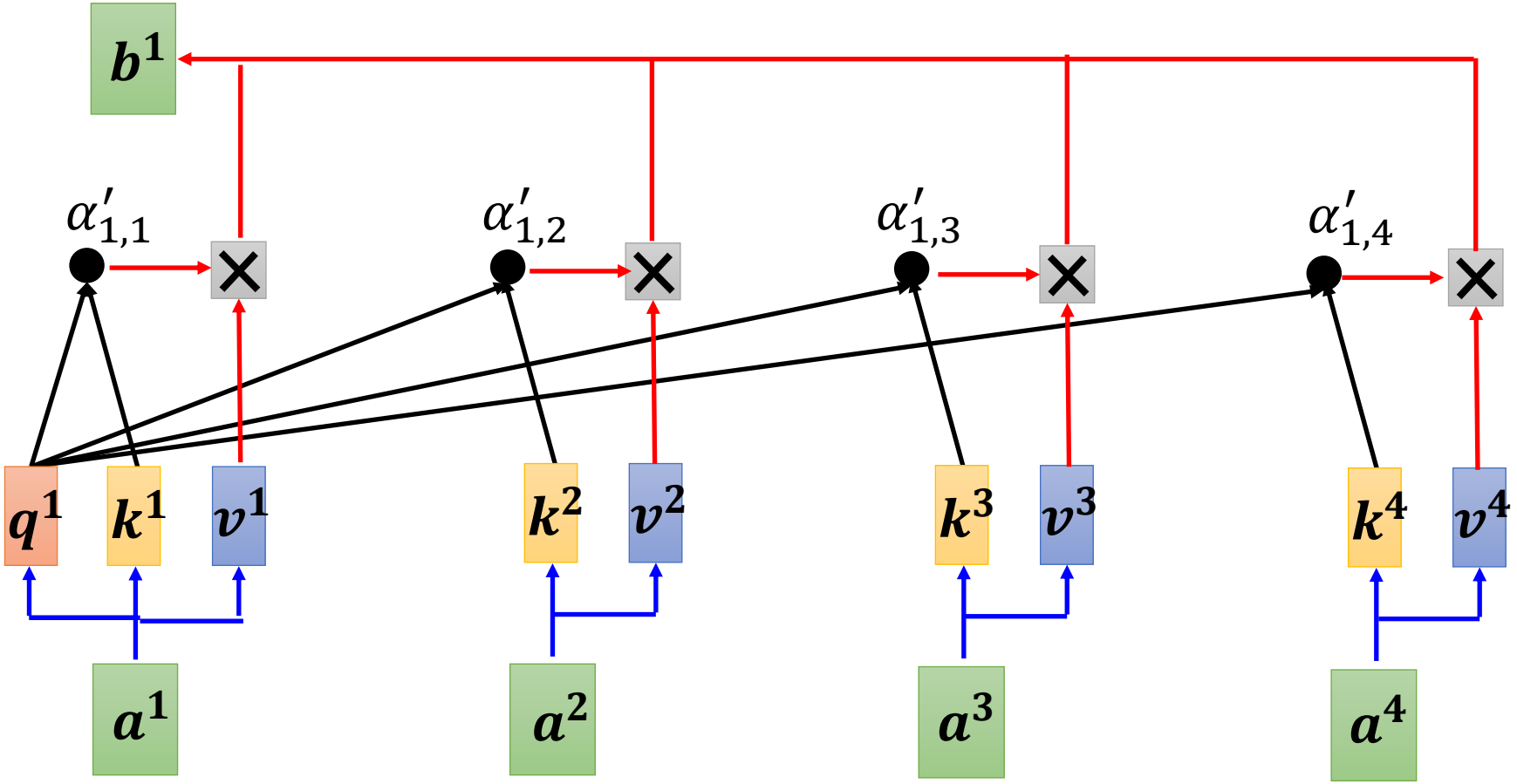
$$q^1 = W^q a^1 \qquad k^2 = W^k a^2 \qquad k^3 = W^k a^3 \qquad k^4 = W^k a^4$$

$$k^1 = W^k a^1$$

Self-attention

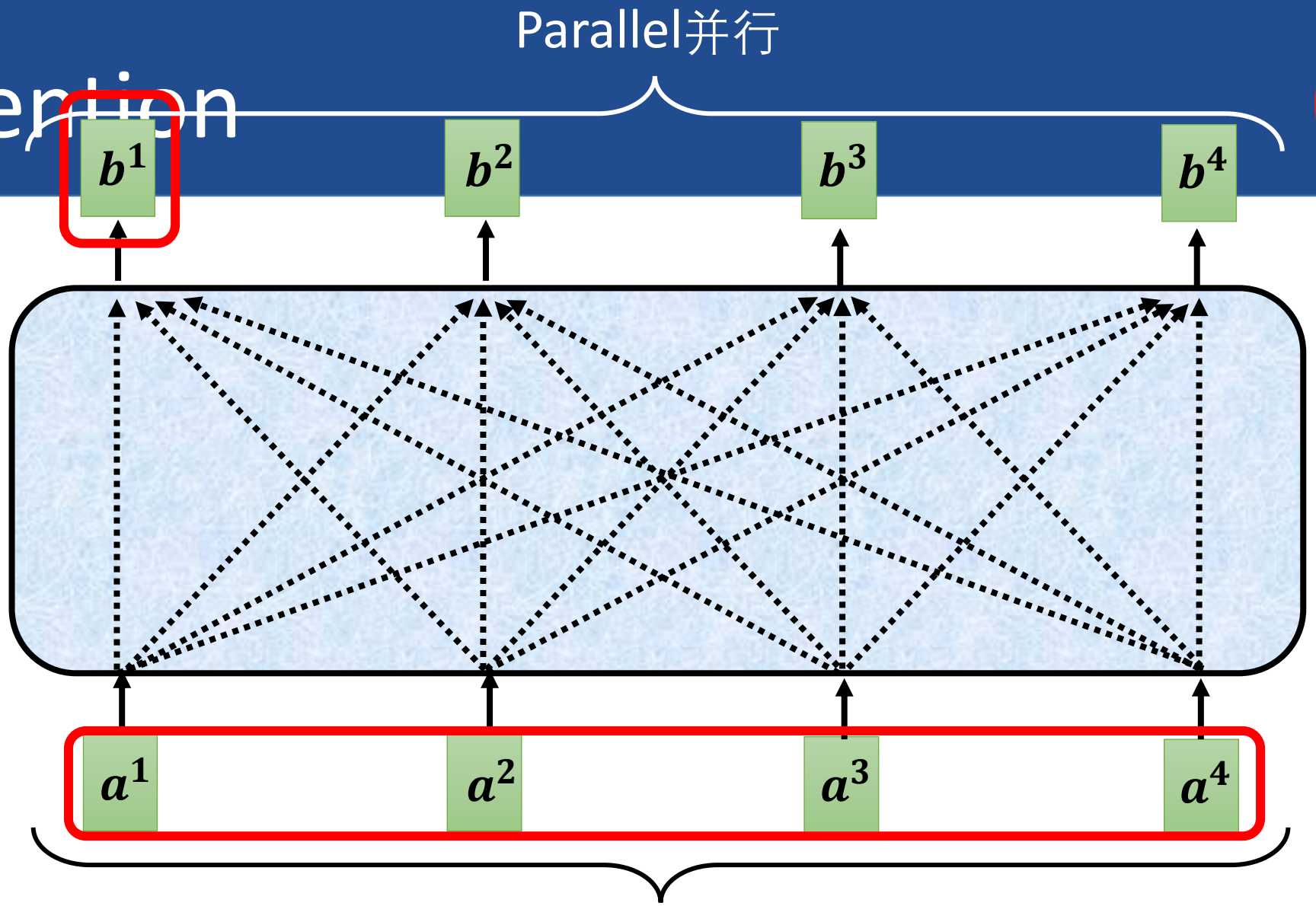
基于注意力系数来提取特征

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



$v^1 = W^v a^1$ $v^2 = W^v a^2$ $v^3 = W^v a^3$ $v^4 = W^v a^4$

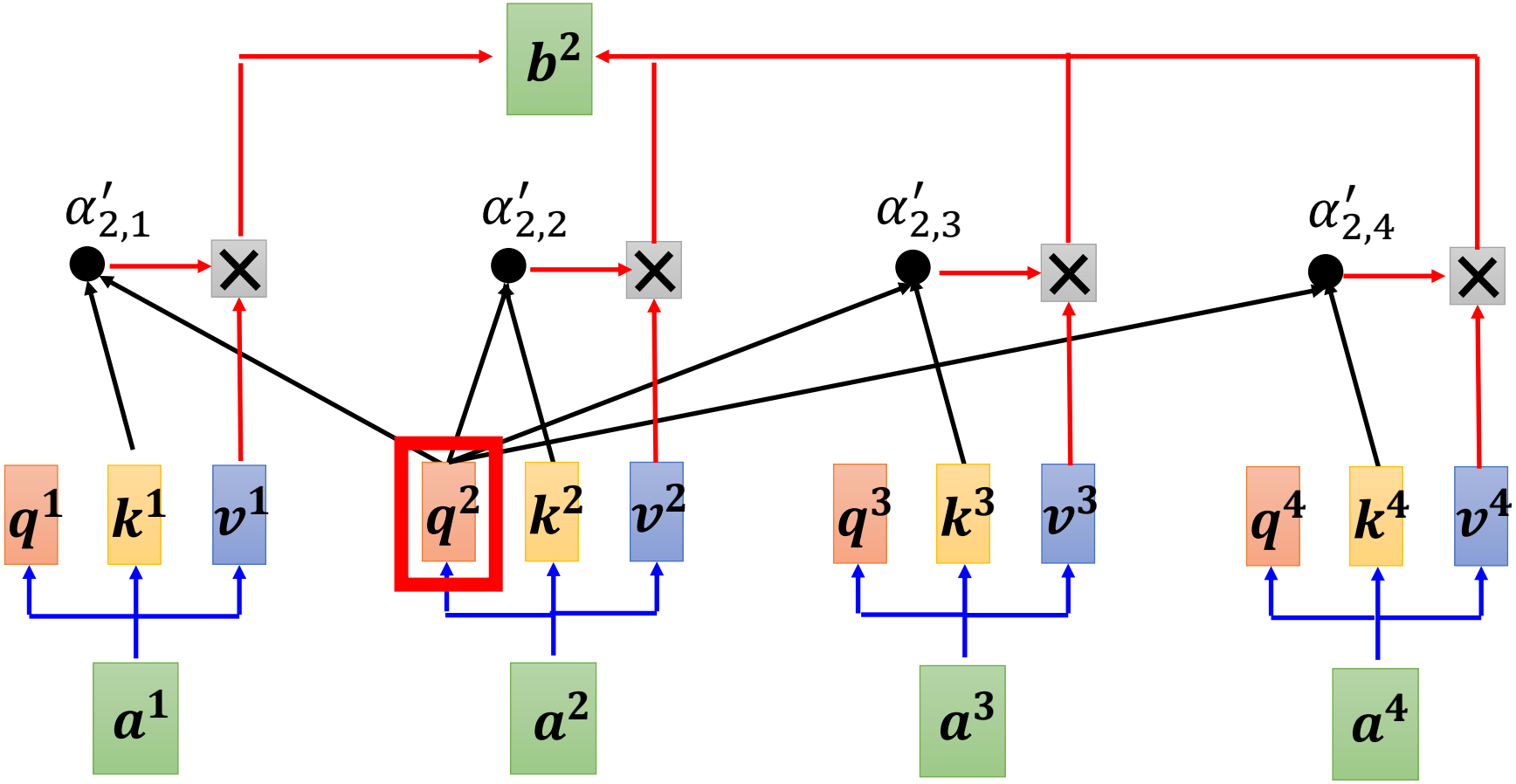
Self-attention



Find the relevant vectors in a sequence找到居中最相关的矢量

Self-attention

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

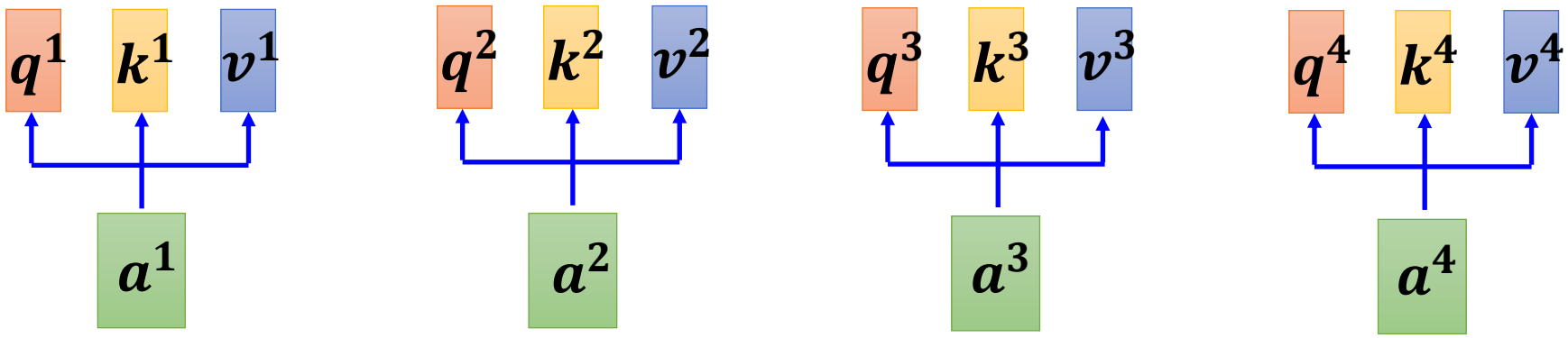


Self-attention

$$q^i = W^q a^i \qquad \underbrace{q^1\ q^2\ q^3\ q^4}_{Q} = W^q \underbrace{a^1\ a^2\ a^3\ a^4}_{I}$$

$$k^i = W^k a^i \qquad \underbrace{k^1\ k^2\ k^3\ k^4}_{K} = W^k \underbrace{a^1\ a^2\ a^3\ a^4}_{I}$$

$$v^i = W^v a^i \qquad \underbrace{v^1\ v^2\ v^3\ v^4}_{V} = W^v \underbrace{a^1\ a^2\ a^3\ a^4}_{I}$$

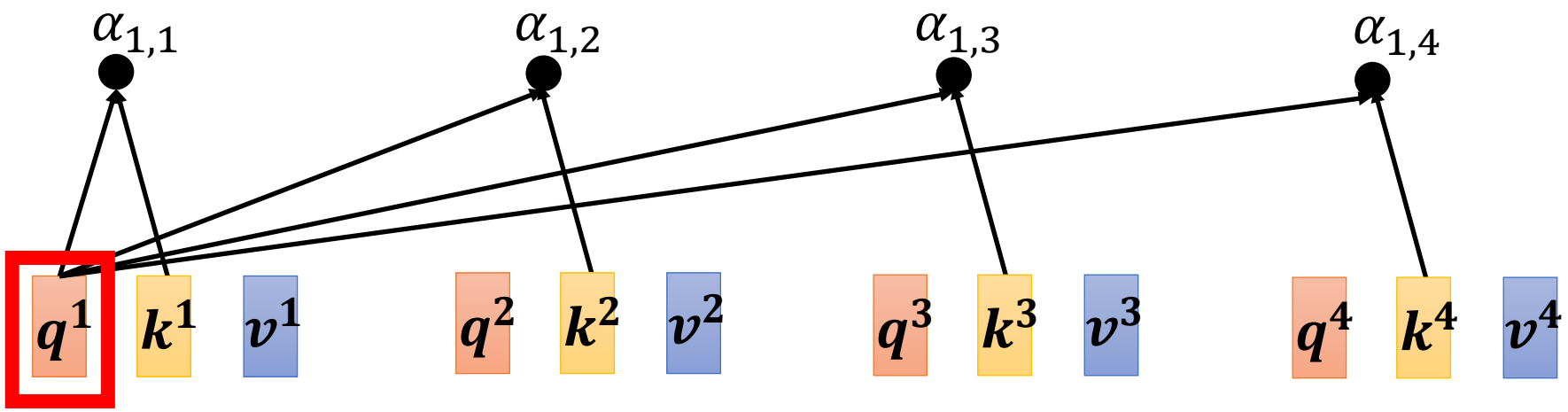


Self-attention

$$\alpha_{1,1} = \boldsymbol{k}^1 \boldsymbol{q}^1 \qquad \alpha_{1,2} = \boldsymbol{k}^2 \boldsymbol{q}^1$$

$$\alpha_{1,3} = \boldsymbol{k}^3 \boldsymbol{q}^1 \qquad \alpha_{1,4} = \boldsymbol{k}^4 \boldsymbol{q}^1$$

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

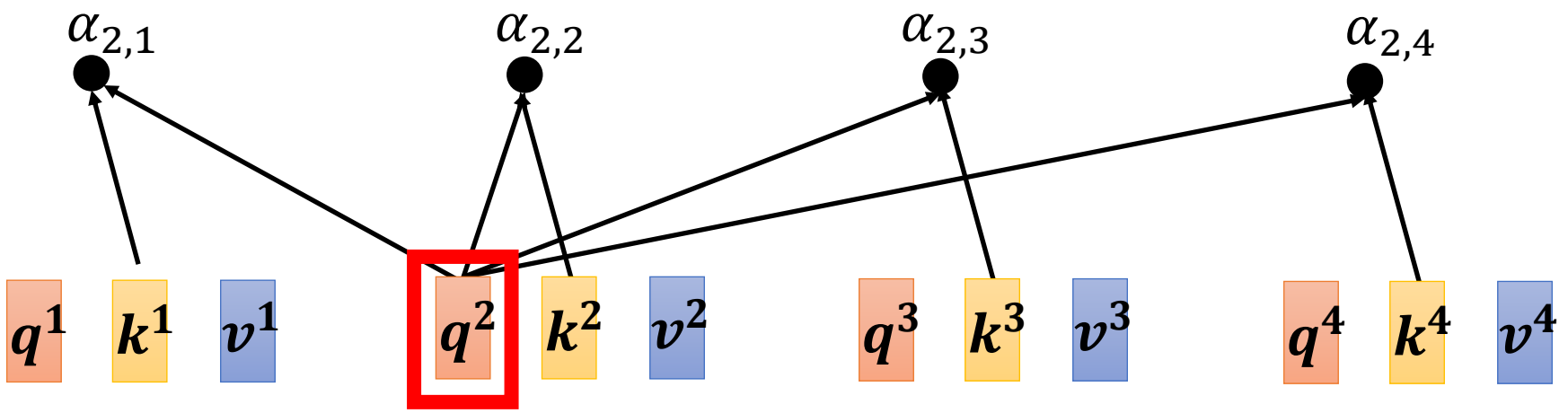


Self-attention

$\alpha_{1,1} = k^1 q^1$ $\alpha_{1,2} = k^2 q^1$

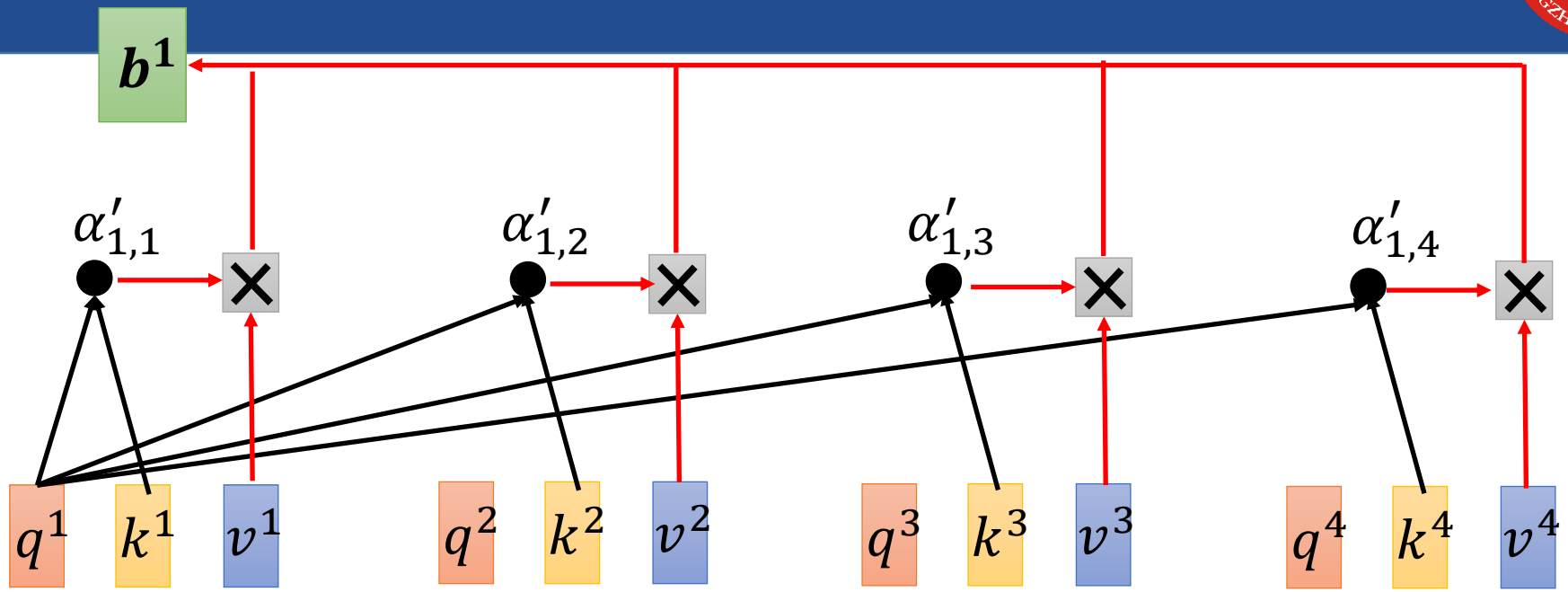
$\alpha_{1,3} = k^3 q^1$ $\alpha_{1,4} = k^4 q^1$

$$\begin{bmatrix} \alpha_{1,1} \\ \alpha_{1,2} \\ \alpha_{1,3} \\ \alpha_{1,4} \end{bmatrix} = \begin{bmatrix} k^1 \\ k^2 \\ k^3 \\ k^4 \end{bmatrix} q^1$$



$$\underbrace{\begin{bmatrix} \alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1} \\ \alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2} \\ \alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3} \\ \alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4} \end{bmatrix}}_{A'} \xleftarrow{\text{softmax}} \underbrace{\begin{bmatrix} \alpha_{1,1} & \alpha_{2,1} & \alpha_{3,1} & \alpha_{4,1} \\ \alpha_{1,2} & \alpha_{2,2} & \alpha_{3,2} & \alpha_{4,2} \\ \alpha_{1,3} & \alpha_{2,3} & \alpha_{3,3} & \alpha_{4,3} \\ \alpha_{1,4} & \alpha_{2,4} & \alpha_{3,4} & \alpha_{4,4} \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} k^1 \\ k^2 \\ k^3 \\ k^4 \end{bmatrix}}_{K^T} \underbrace{\begin{bmatrix} q^1 & q^2 & q^3 & q^4 \end{bmatrix}}_{Q}$$

Self-attention



$$\underbrace{\boldsymbol{b}^1\ \boldsymbol{b}^2\ \boldsymbol{b}^3\ \boldsymbol{b}^4}_{O} = \underbrace{\boldsymbol{v}^1\ \boldsymbol{v}^2\ \boldsymbol{v}^3\ \boldsymbol{v}^4}_{V} \underbrace{\begin{bmatrix} \alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1} \\ \alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2} \\ \alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3} \\ \alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4} \end{bmatrix}}_{A'}$$

Self-attention

$$Q = W^q I$$

$$K = W^k I$$

$$V = W^v I$$

Parameters to be learned

超参

$$A = K^T Q$$

$$A' \leftarrow A$$

Attention Matrix

注意力矩阵

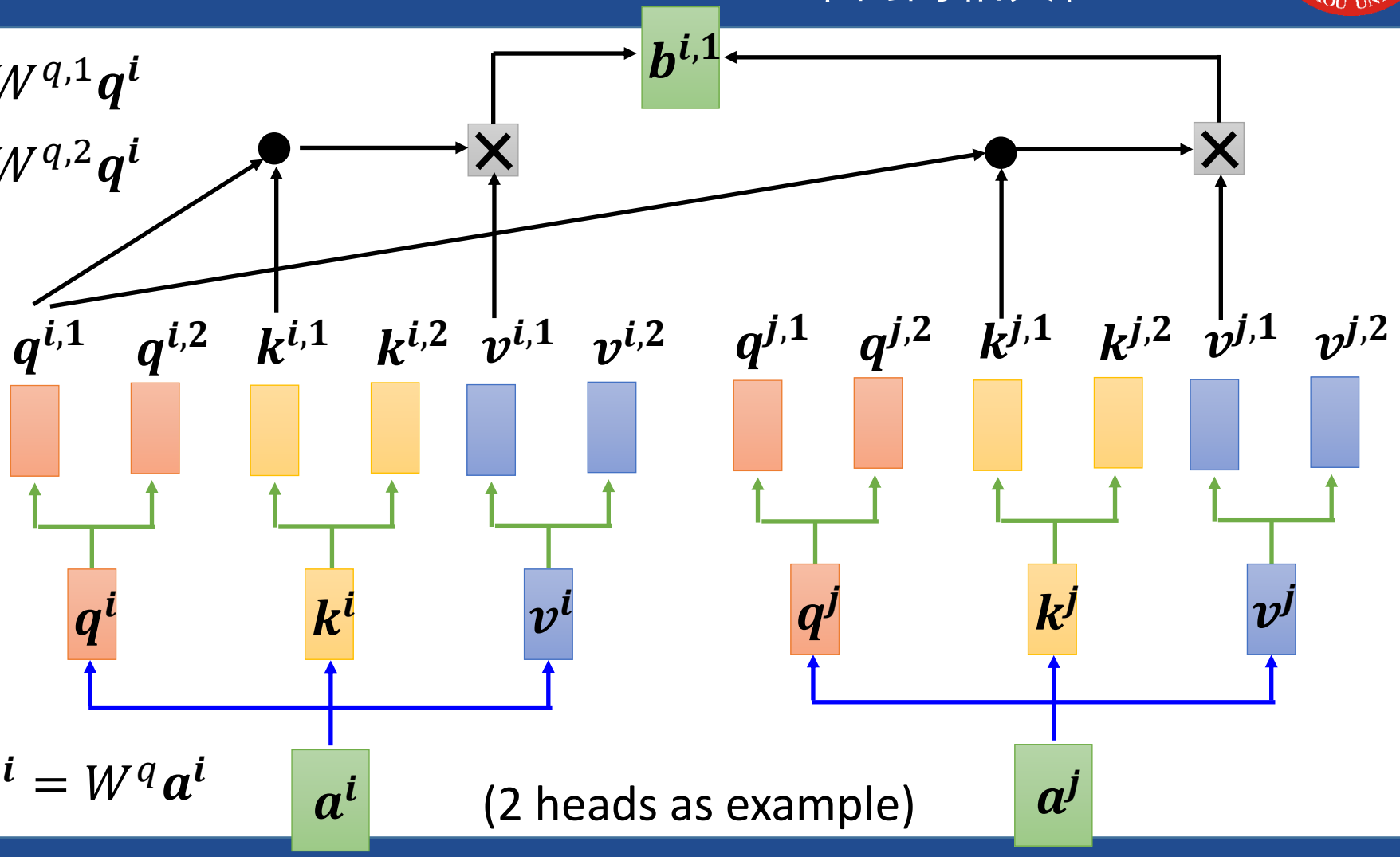
$$O = V A'$$

Multi-head Self-attention

不同的相关性

$$q^{i,1} = W^{q,1} q^i$$

$$q^{i,2} = W^{q,2} q^i$$



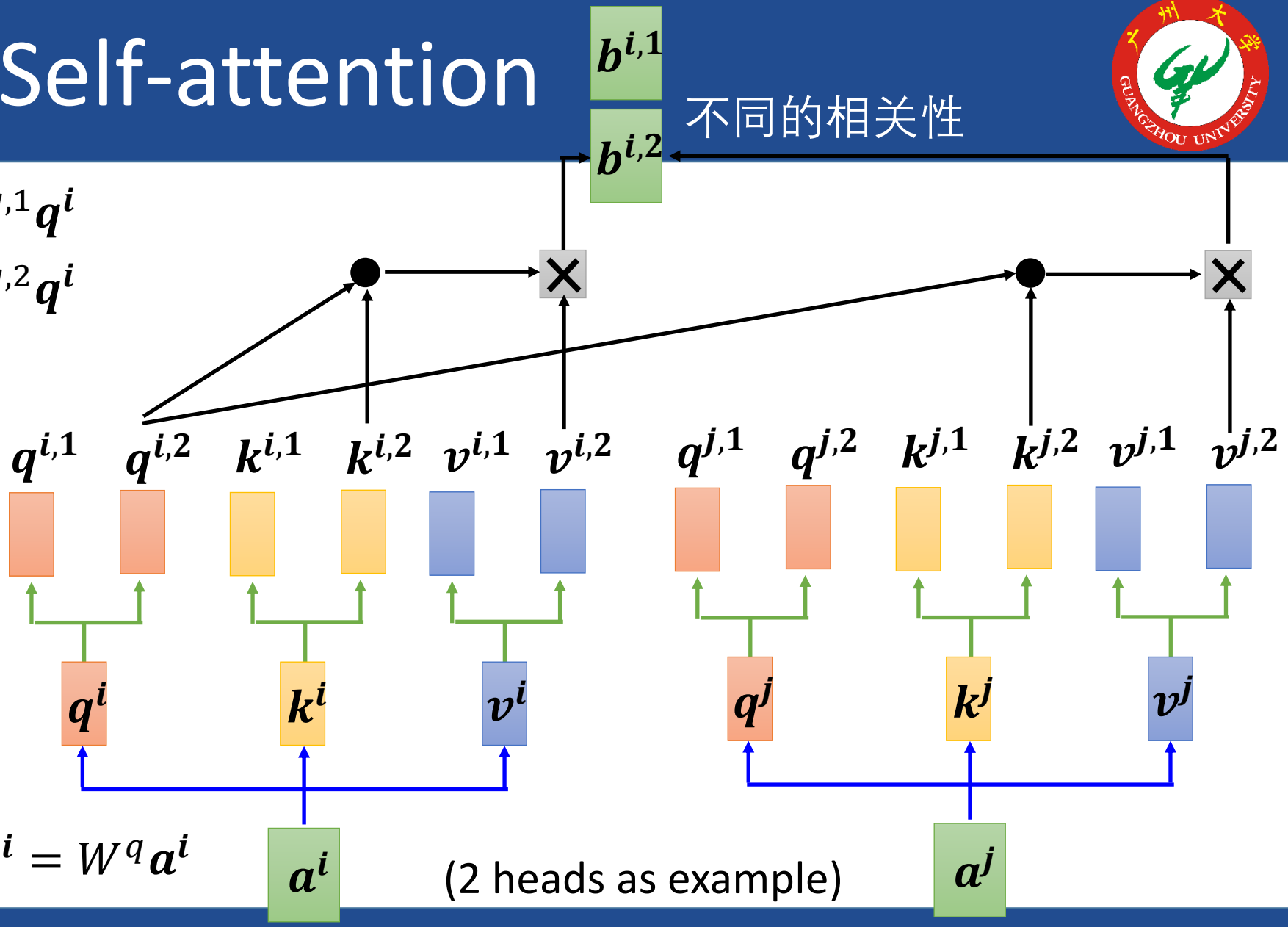
$$q^i = W^q a^i$$

(2 heads as example)

Multi-head Self-attention

$$q^{i,1} = W^{q,1} q^i$$

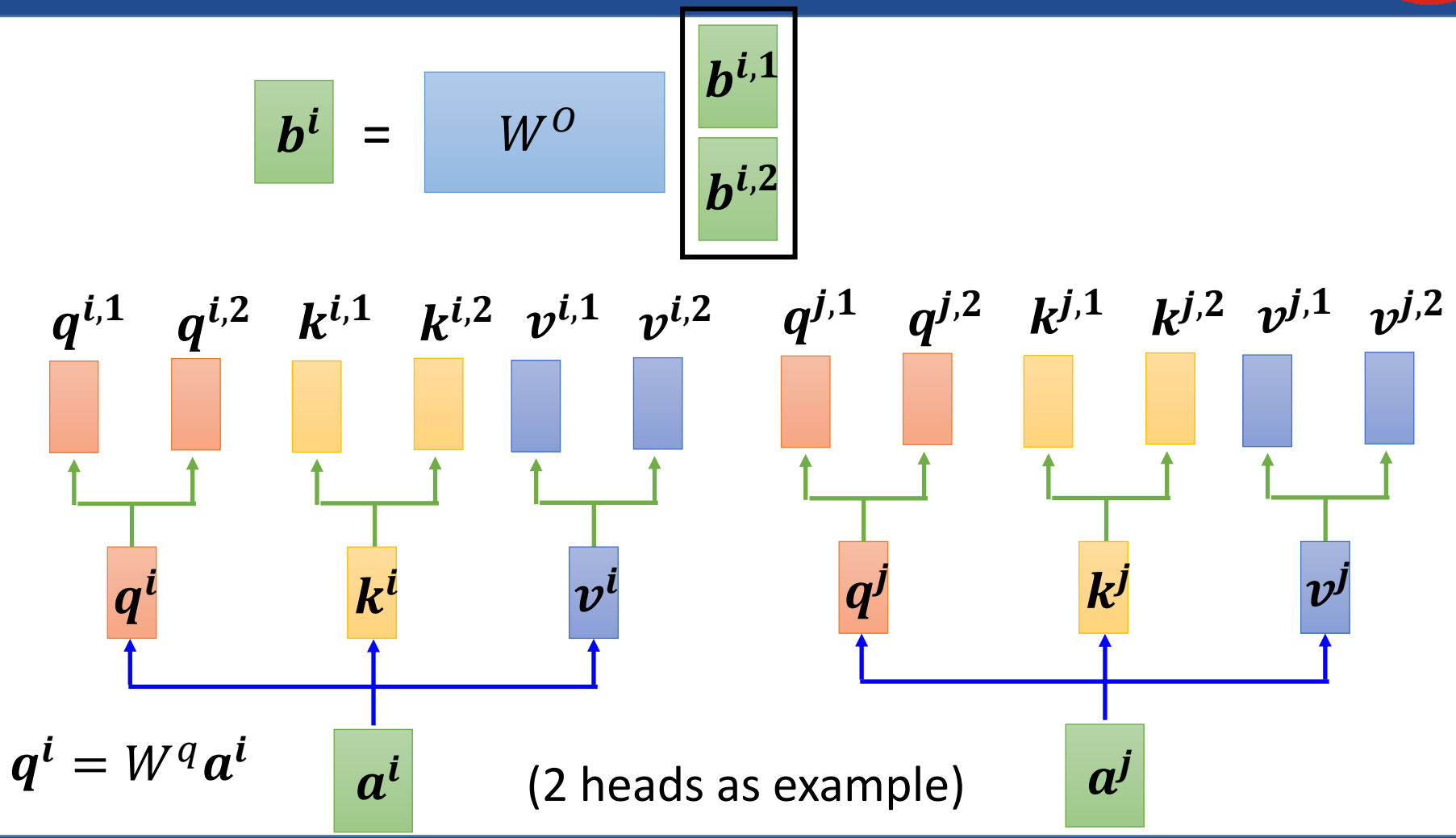
$$q^{i,2} = W^{q,2} q^i$$



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

(2 heads as example)

Multi-head Self-attention



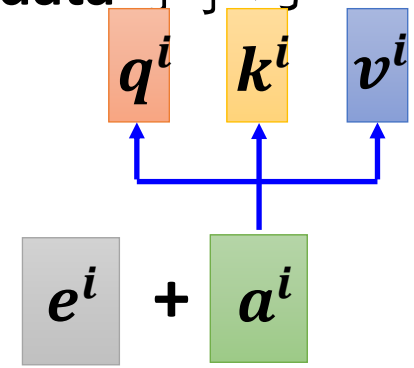
$q^i = W^q a^i$

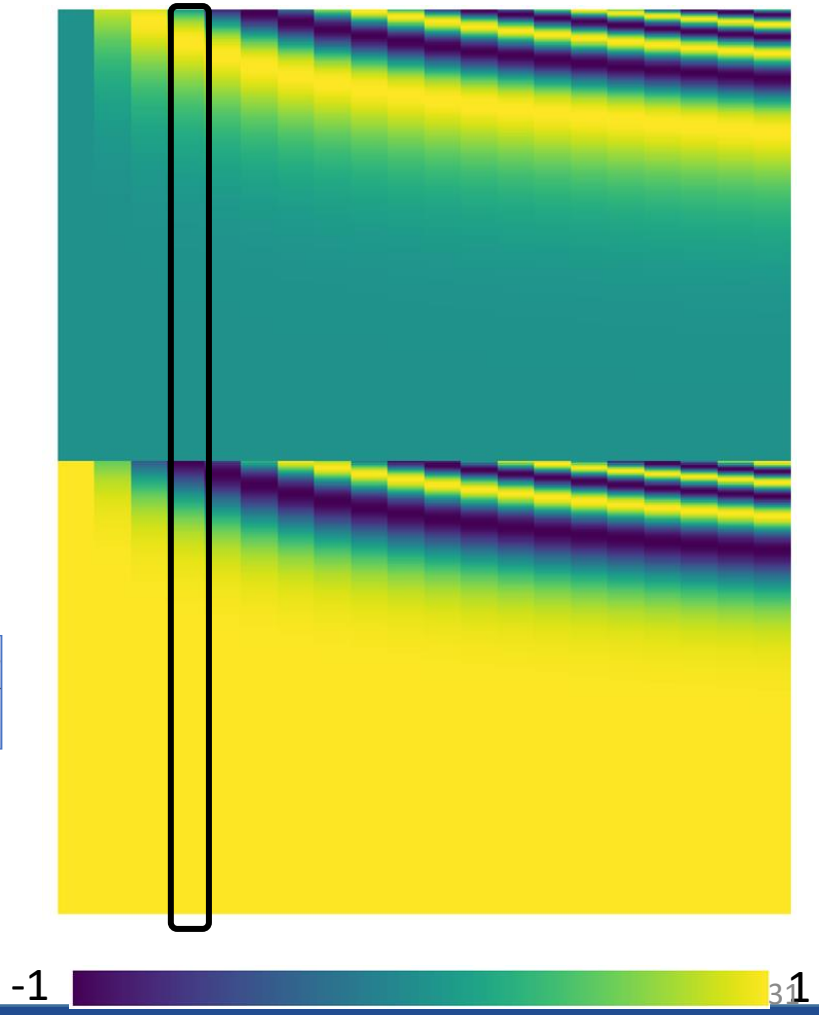
(2 heads as example)

Positional Encoding

Each column represents a positional vector e^i
每列代表不同的位置矢量

- No position information in self-attention. 自注意力机制无位置信息
- Each position has a unique positional vector e^i 每位需加独特的位置矢量。
- **hand-crafted** 手动增加
- **learned from data**可学习





Many applications …

Transformer

https://arxiv.org/abs/1706.03762

BERT

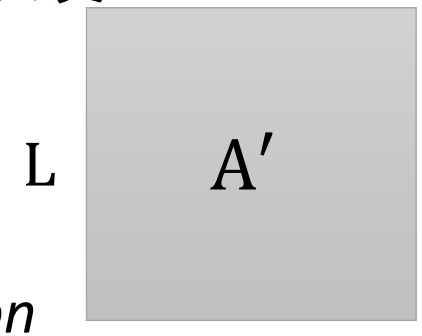
https://arxiv.org/abs/1810.04805

Self-attention for Speech

Speech is a very long vector sequence.语音长度较长

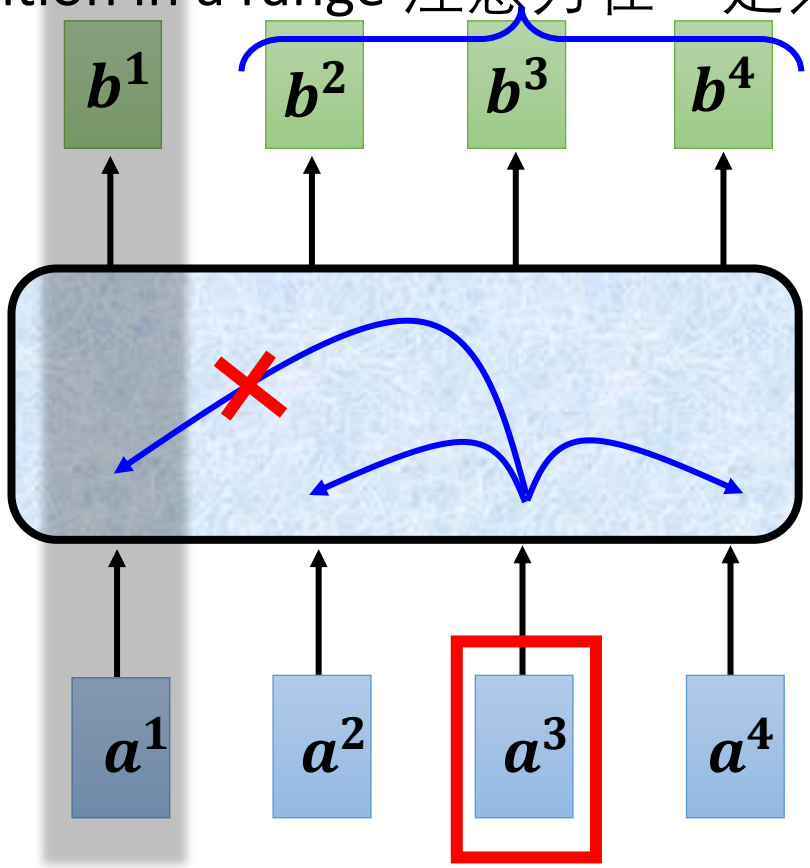
10ms

If input sequence is length L
输入序列长度L



Attention Matrix

Attention in a range 注意力在一定范围内



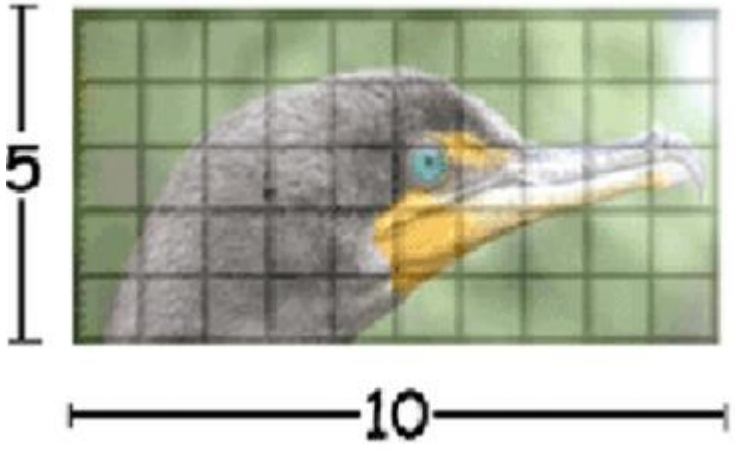
Truncated Self-attention

截断自注意力

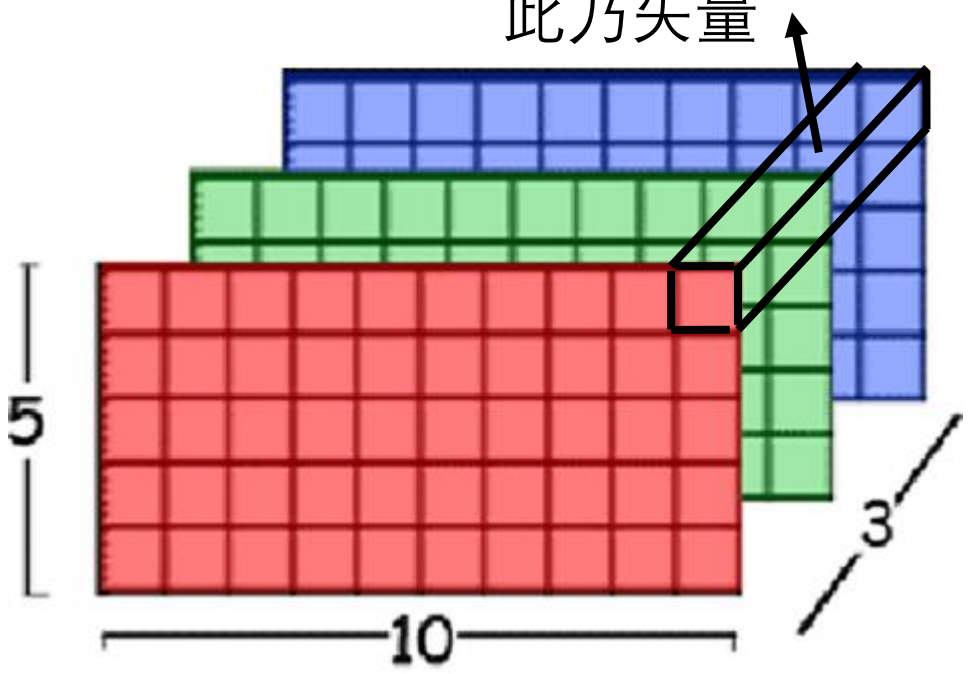
https://arxiv.org/abs/1910.12977

Self-attention for Image

An **image** can also be considered as a **vector set**.
图像也是矢量集合

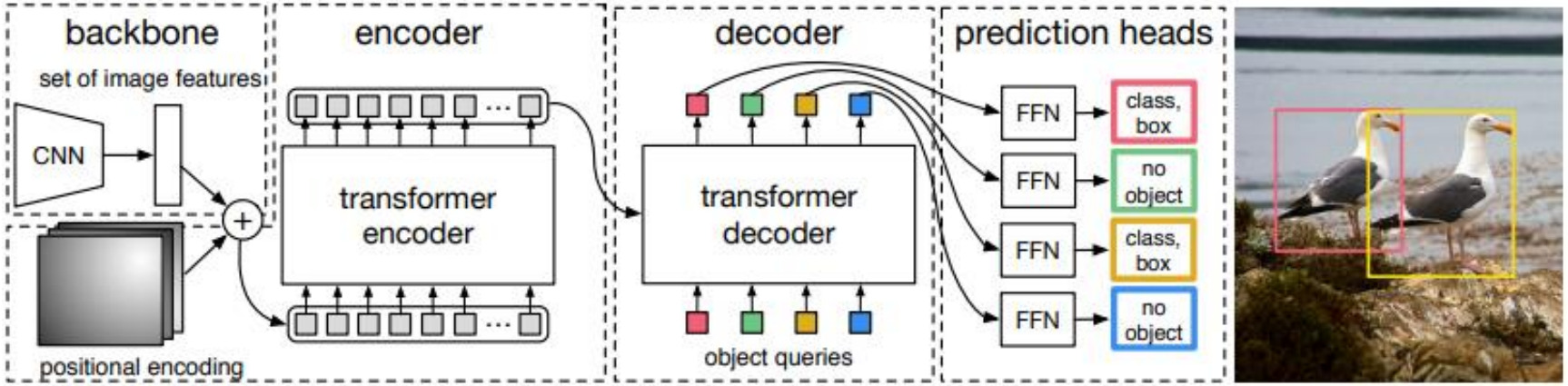


This is a vector.
此乃矢量



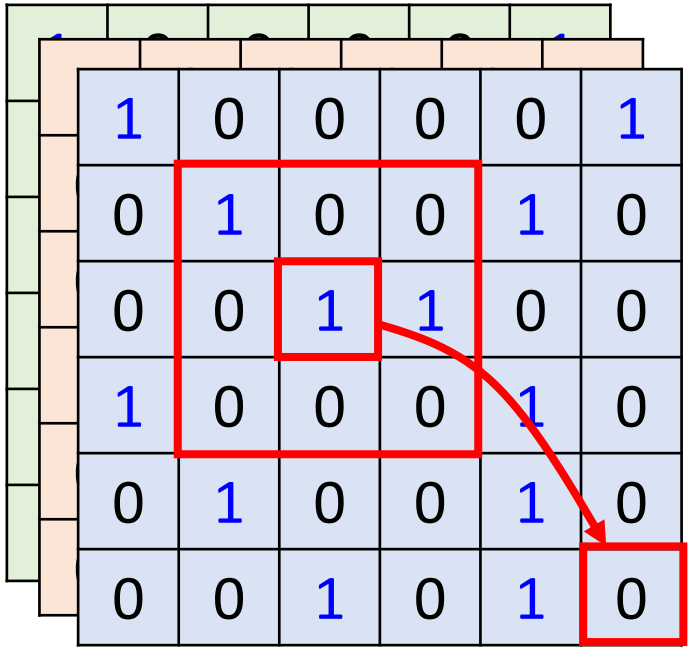
Self-Attention GAN

DEtection Transformer (DETR)



https://arxiv.org/abs/2005.12872

Self-attention v.s. CNN



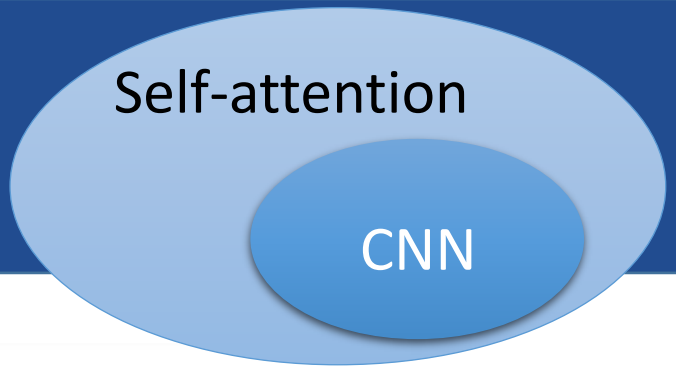
CNN: self-attention that can only attends in a receptive field
CNN:感受野内部计算

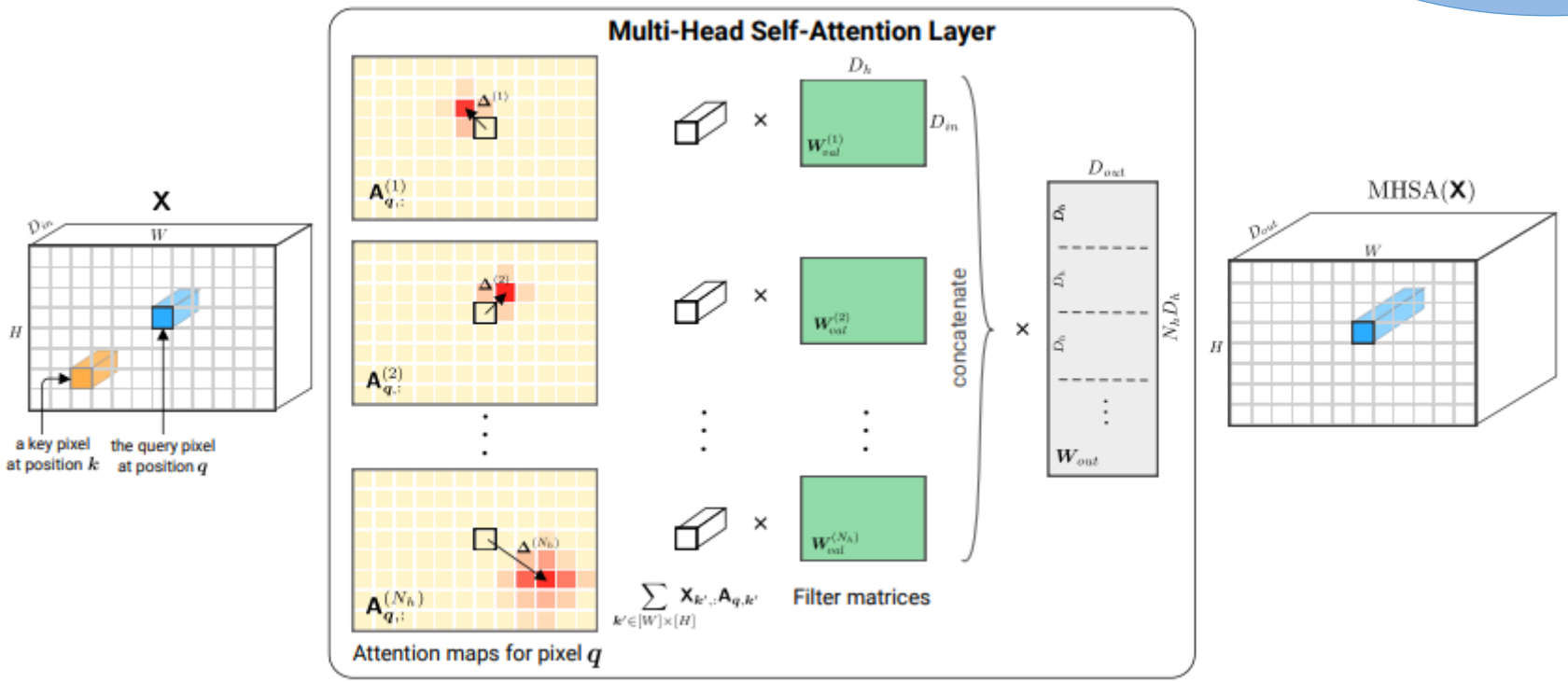
- CNN is simplified self-attention.
  CNN是种简单的自注意力机制

Self-attention: CNN with learnable receptive field
Self-attention:可学习感受野的特殊CNN

- Self-attention is the complex version of CNN.
  Self-attention是复杂版本的CNN

Self-attention v.s. CNN



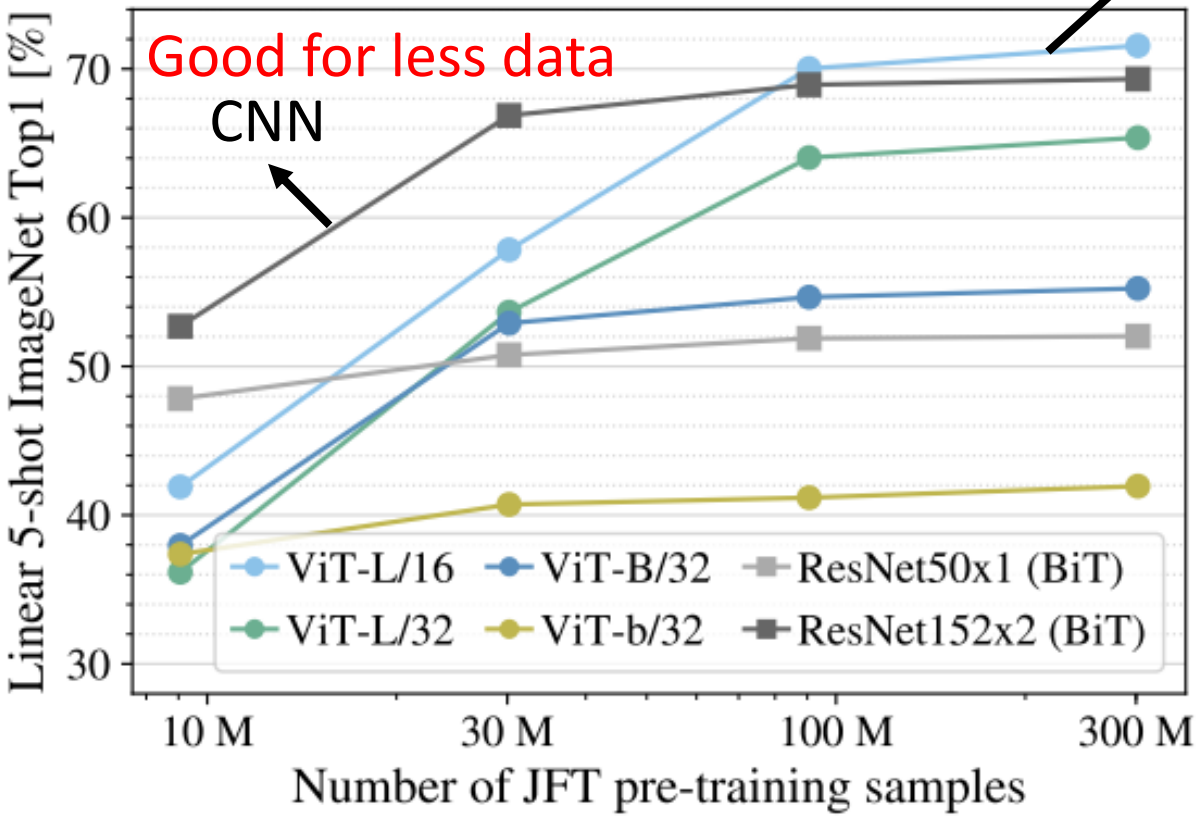


https://arxiv.org/abs/1911.03584

Self-attention v.s. CNN

Good for more data

Self-attention

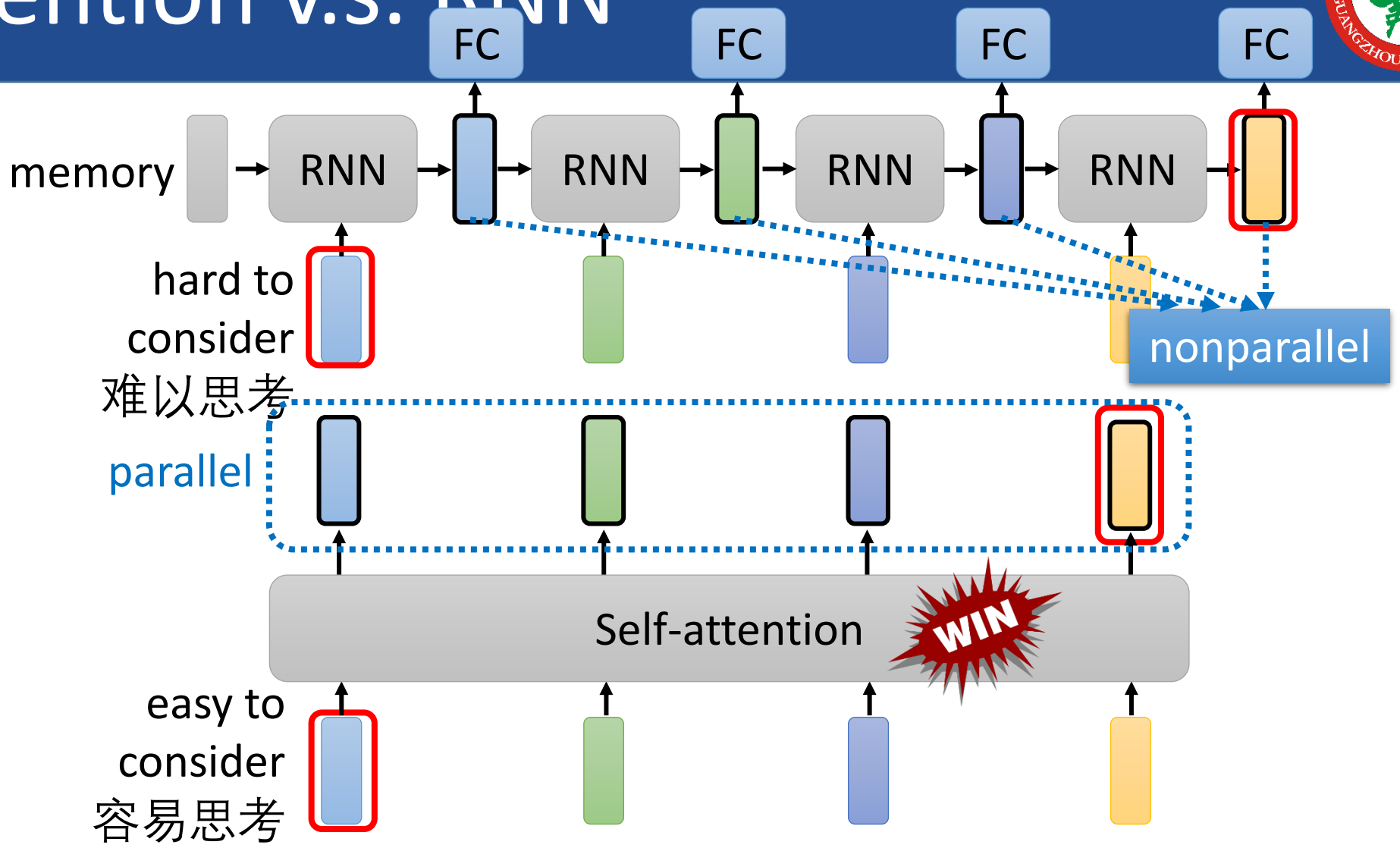


An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

https://arxiv.org/pdf/2010.11929.pdf

Self-attention v.s. RNN

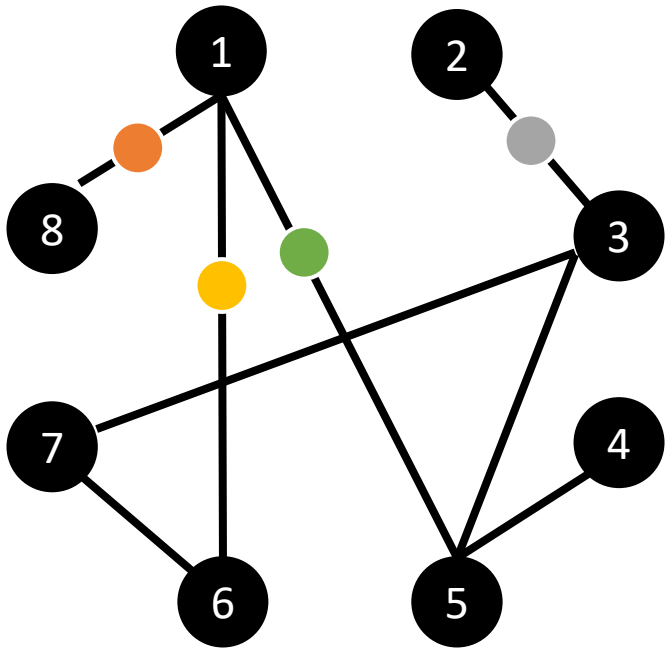
Recurrent Neural Network (RNN)



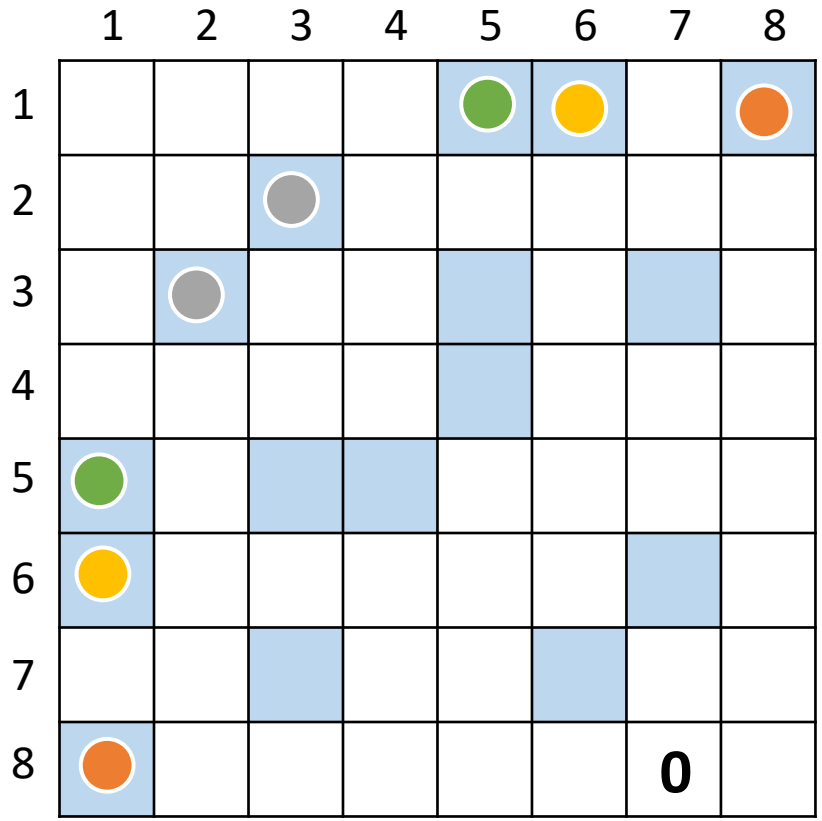
Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention

https://arxiv.org/abs/2006.16236

Self-attention v.s. Graph



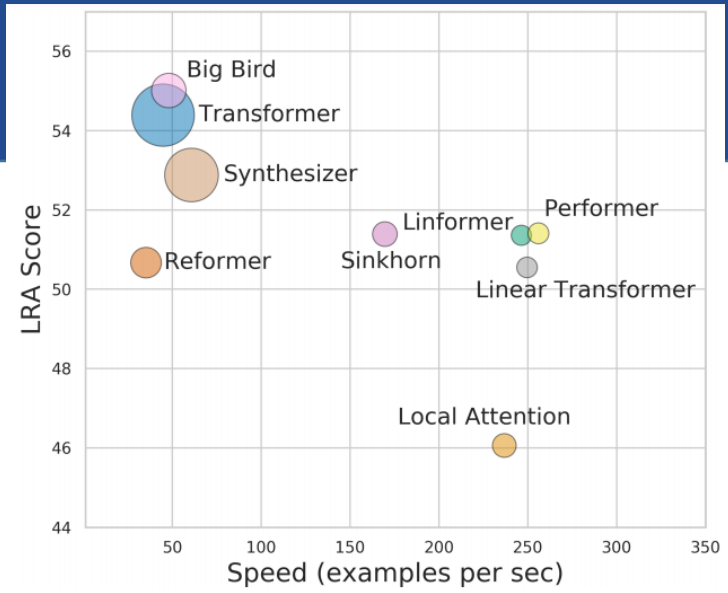
Attention Matrix



Consider **edge**: only attention to connected nodes

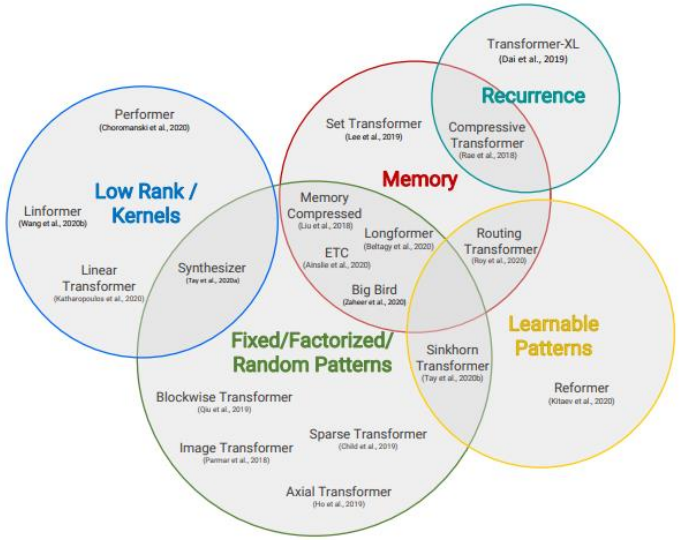
This is one type of **Graph Neural Network (GNN)**.

To Learn More …



Long Range Arena: A Benchmark for Efficient Transformers

https://arxiv.org/abs/2011.04006



Efficient Transformers: A Survey

https://arxiv.org/abs/2009.06732

Q&A