道路工程深度学习技术

第七周 Transformer 原理及应用

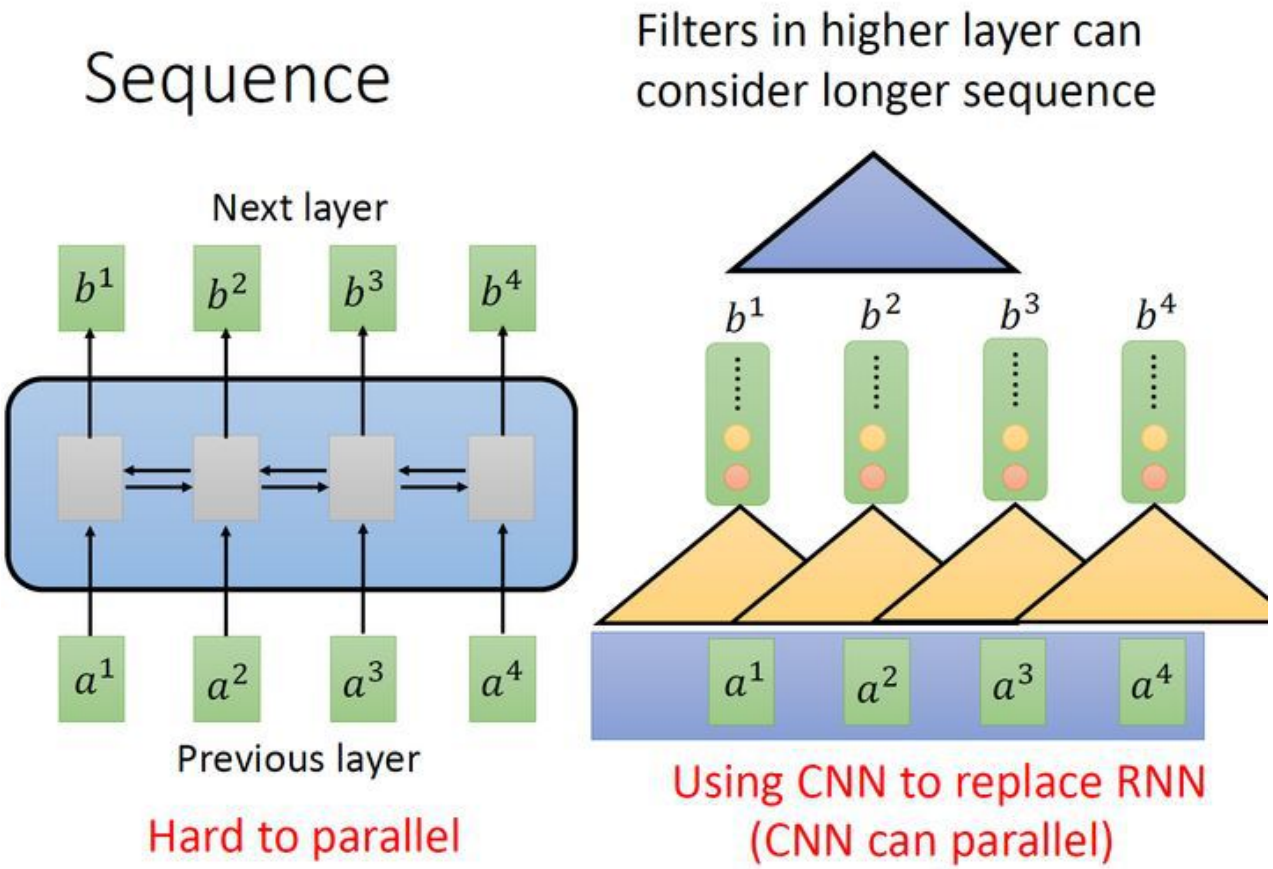
童峥

东南大学 交通学院

2023 年 4 月 25 日

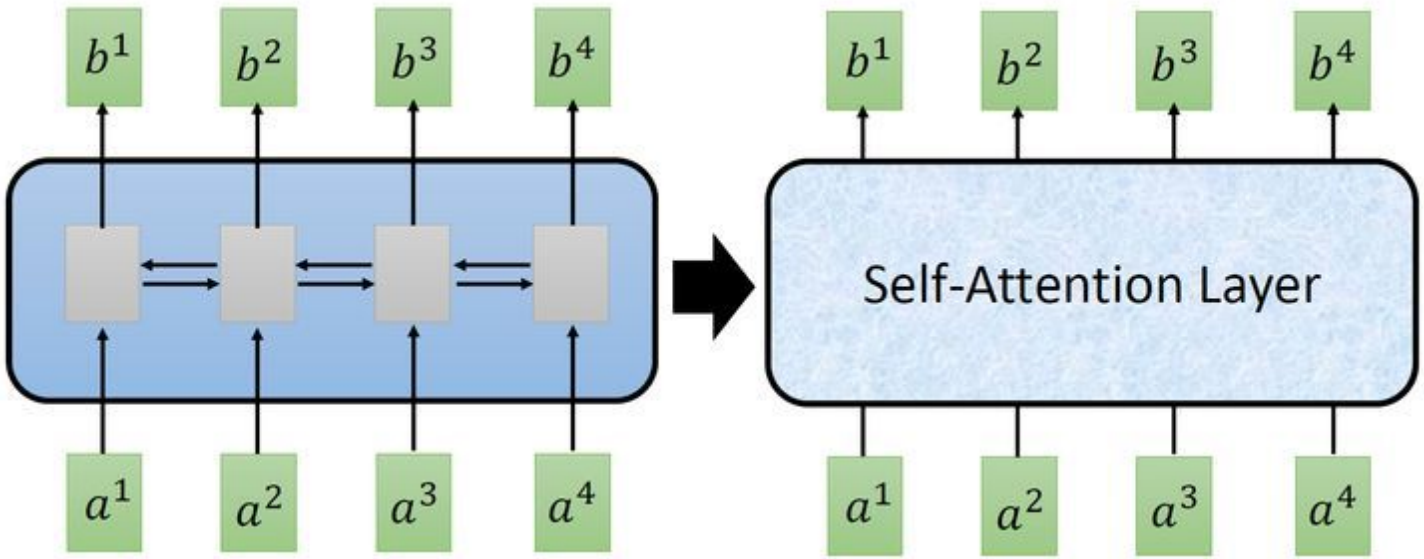
Sequence to Sequence model

- Transformer 是一个 Sequence to Sequence model
- RNN 很不容易并行化 (hard to parallel)



Self-attention 结构层

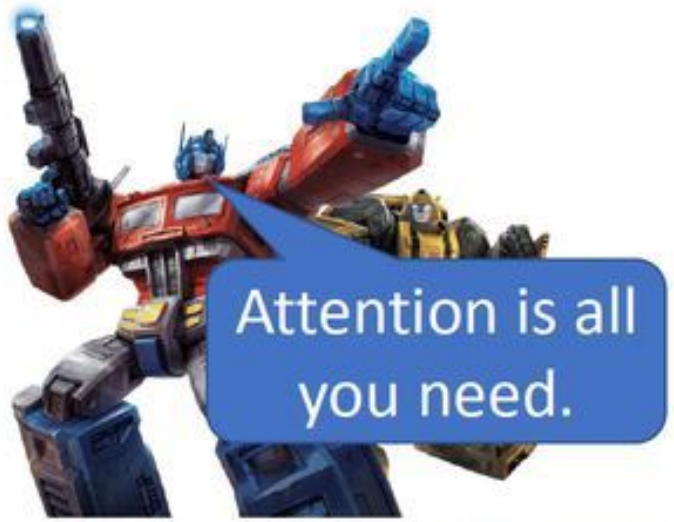
- 输入和输出和 RNN 是一模一样的
- 输出 b_1-b_4 可以并行化计算



Self-attention 结构层

Self-attention

https://arxiv.org/abs/1706.03762



q: query (to match others)

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

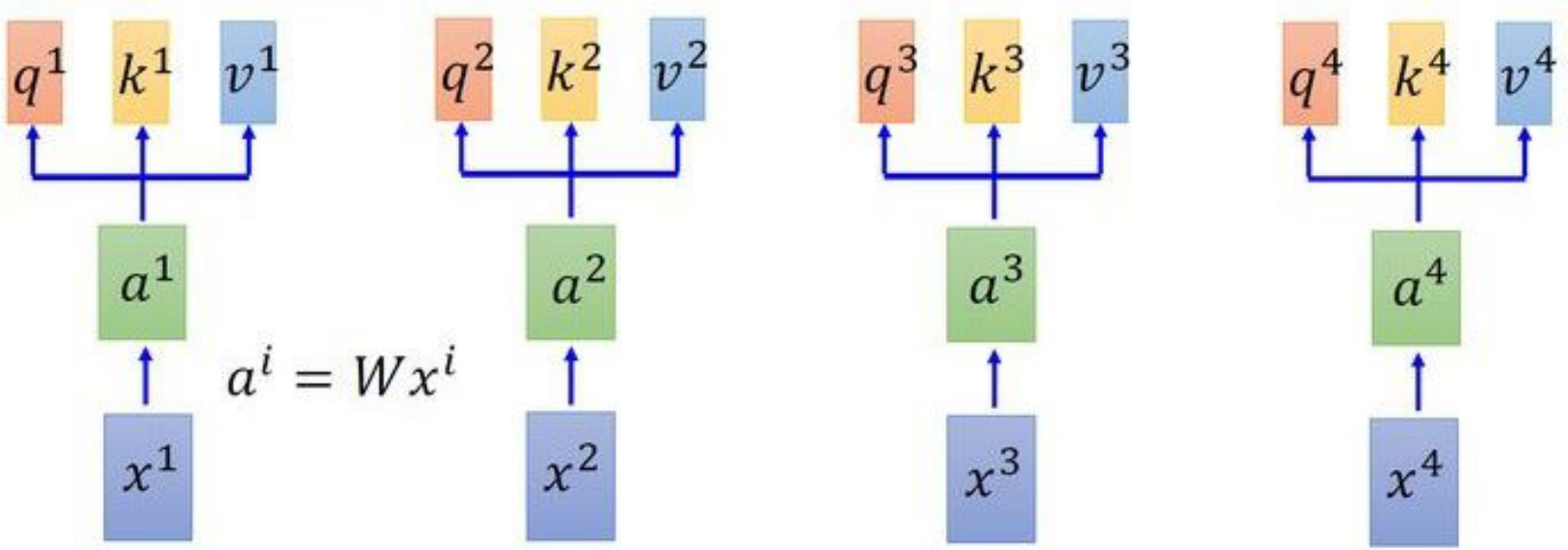
k: key (to be matched)

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

v: information to be extracted

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

Self-attention 结构层



注意力机制

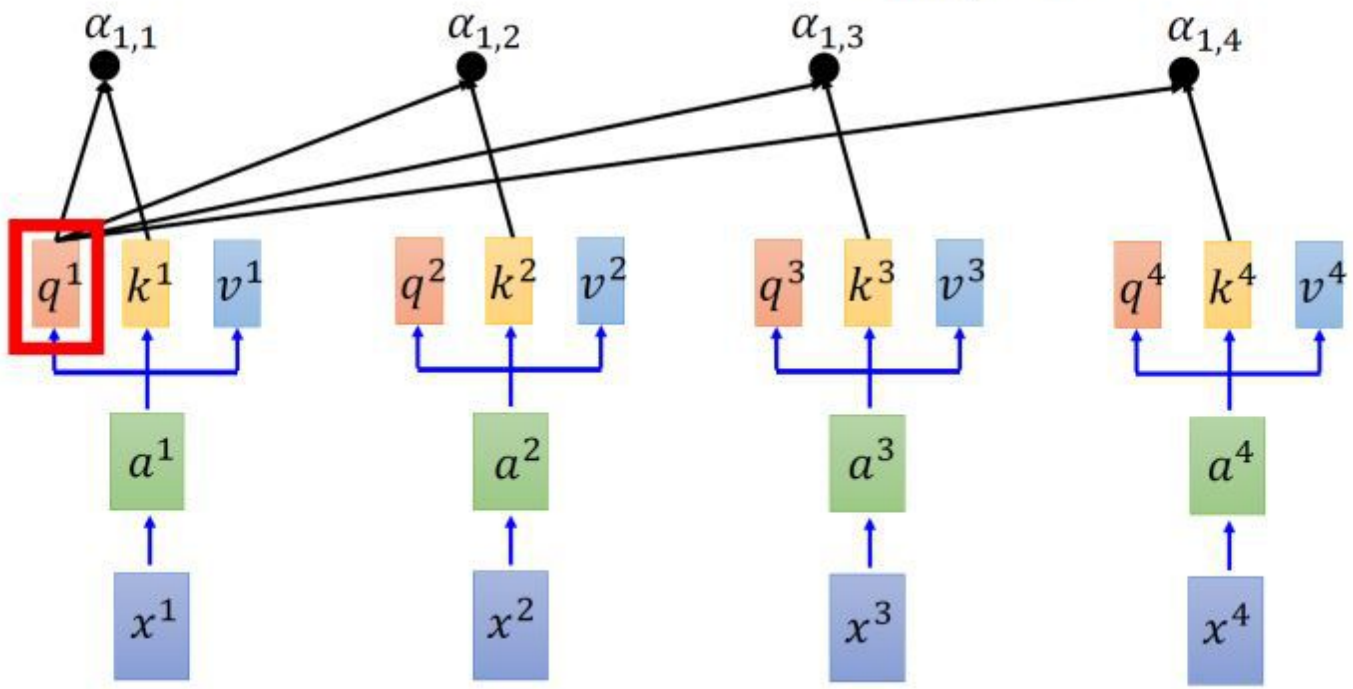
Self-attention

拿每個 query q 去對每個 key k 做 attention

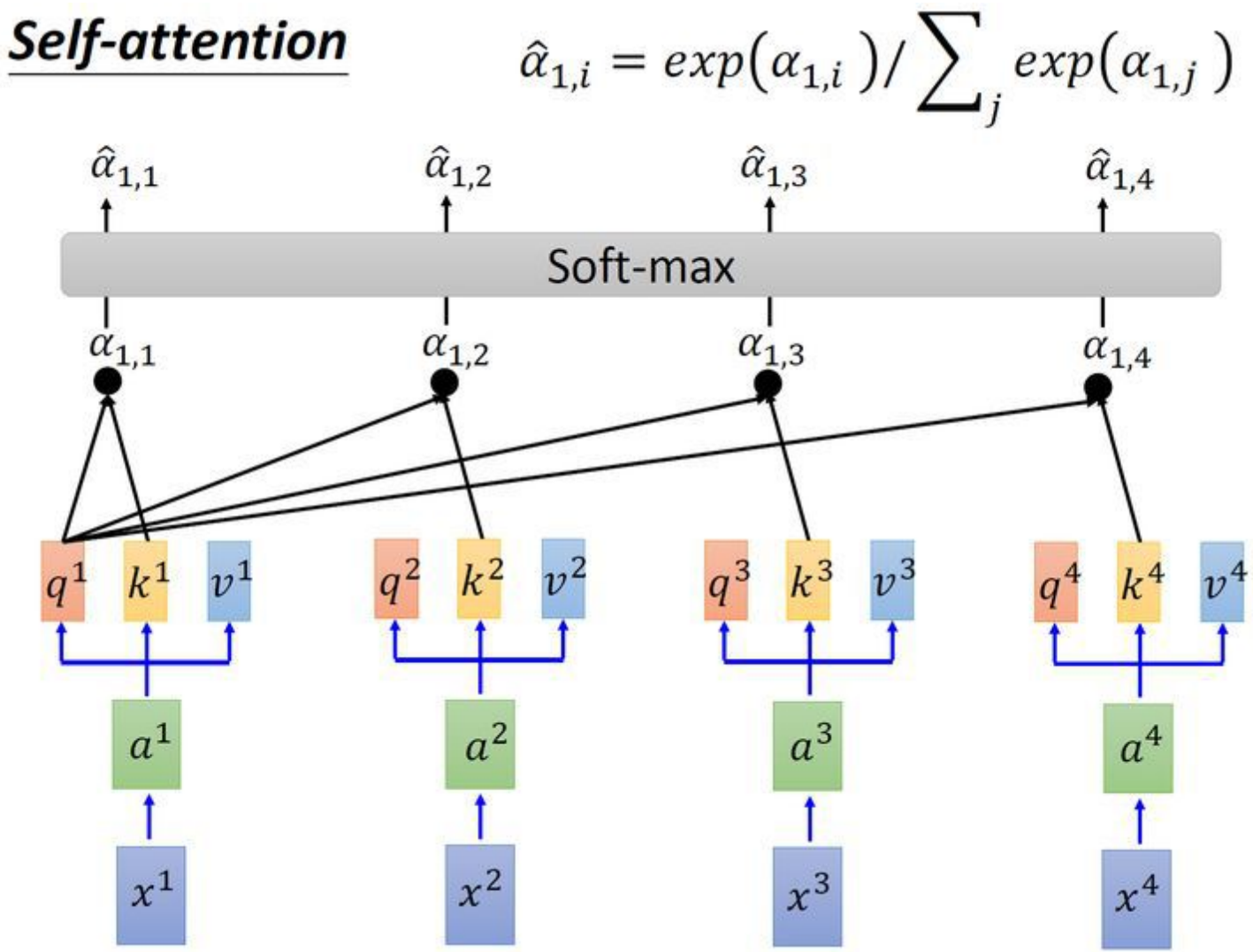
d is the dim of q and k

Scaled Dot-Product Attention: $\alpha_{1,i} = q^1 \cdot k^i / \sqrt{d}$

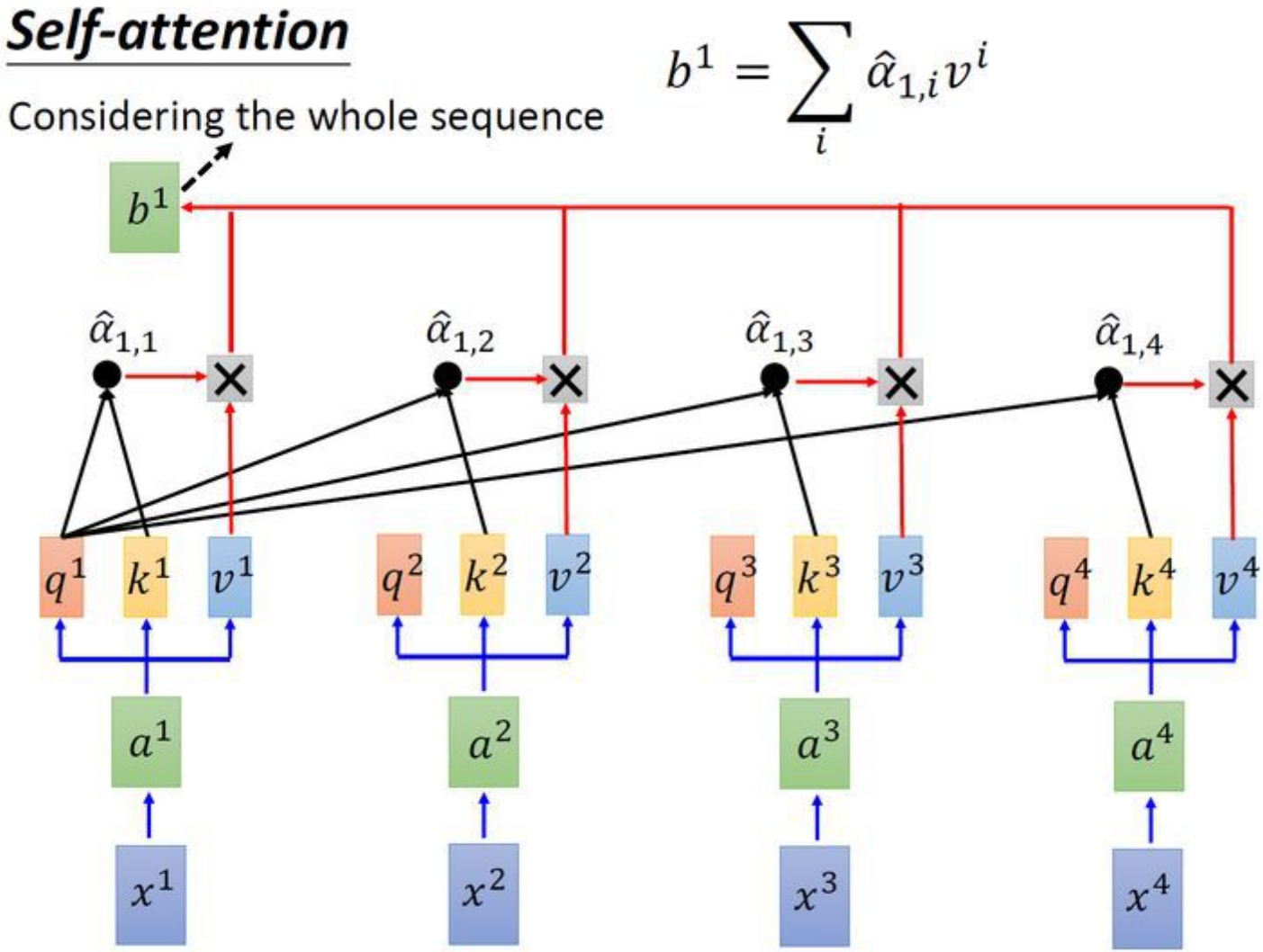
dot product



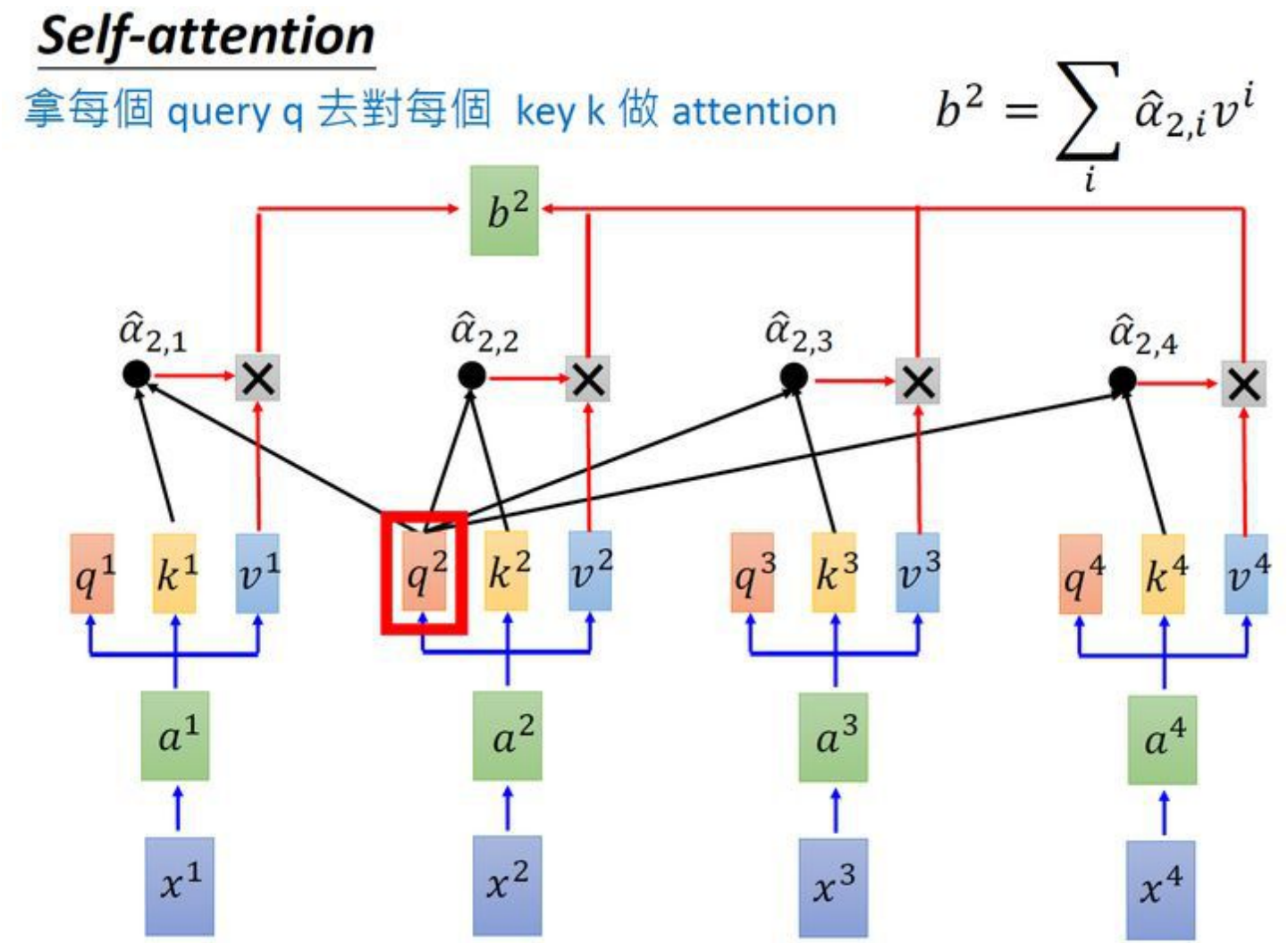
Softmax 标准化


Self-attention
$\hat{\alpha}_{1,i} = exp(\alpha_{1,i}) / \sum_j exp(\alpha_{1,j})$
$\hat{\alpha}_{1,1}$
$\hat{\alpha}_{1,2}$
$\hat{\alpha}_{1,3}$
$\hat{\alpha}_{1,4}$
Soft-max
$\alpha_{1,1}$
$\alpha_{1,2}$
$\alpha_{1,3}$
$\alpha_{1,4}$
q^1 k^1 v^1
q^2 k^2 v^2
q^3 k^3 v^3
q^4 k^4 v^4
a^1 a^2 a^3 a^4
x^1 x^2 x^3 x^4

Sequence 组合


Self-attention
Considering the whole sequence
$b^1 = \sum_i \hat{\alpha}_{1,i} v^i$
b^1
$\hat{\alpha}_{1,1}$
$\hat{\alpha}_{1,2}$
$\hat{\alpha}_{1,3}$
$\hat{\alpha}_{1,4}$
q^1 k^1 v^1
q^2 k^2 v^2
q^3 k^3 v^3
q^4 k^4 v^4
a^1 a^2 a^3 a^4
x^1 x^2 x^3 x^4

Sequence 组合


Self-attention
拿每個 query q 去對每個 key k 做 attention
$b^2 = \sum_i \hat{\alpha}_{2,i} v^i$
b^2
$\hat{\alpha}_{2,1}$
$\hat{\alpha}_{2,2}$
$\hat{\alpha}_{2,3}$
$\hat{\alpha}_{2,4}$
q^1 k^1 v^1
q^2 k^2 v^2
q^3 k^3 v^3
q^4 k^4 v^4
a^1 a^2 a^3 a^4
x^1 x^2 x^3 x^4

Sequence 组合

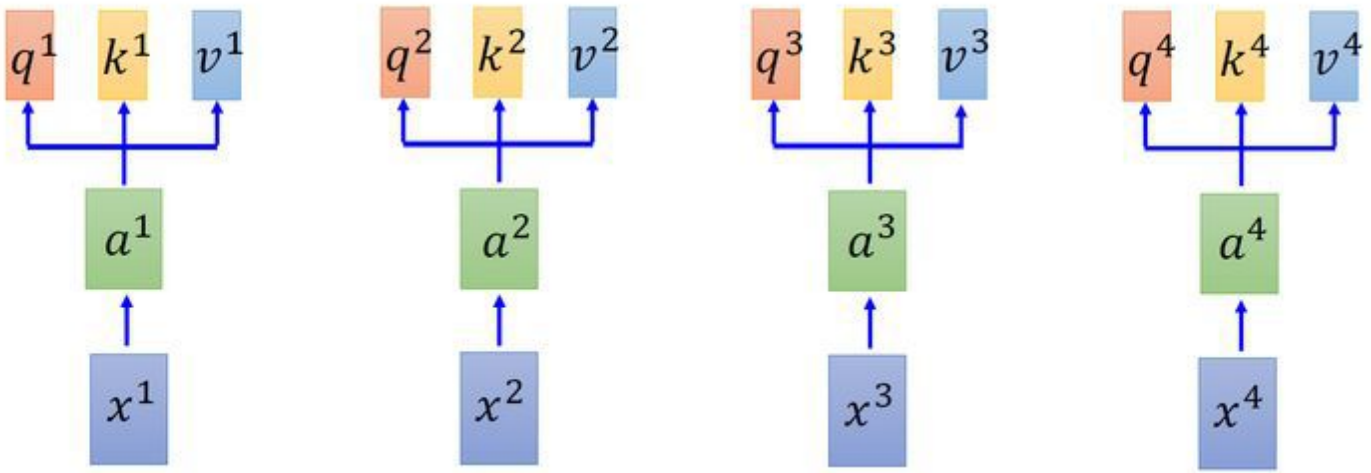
Self-attention

$$q^i = W^q a^i$$
$$k^i = W^k a^i$$
$$v^i = W^v a^i$$

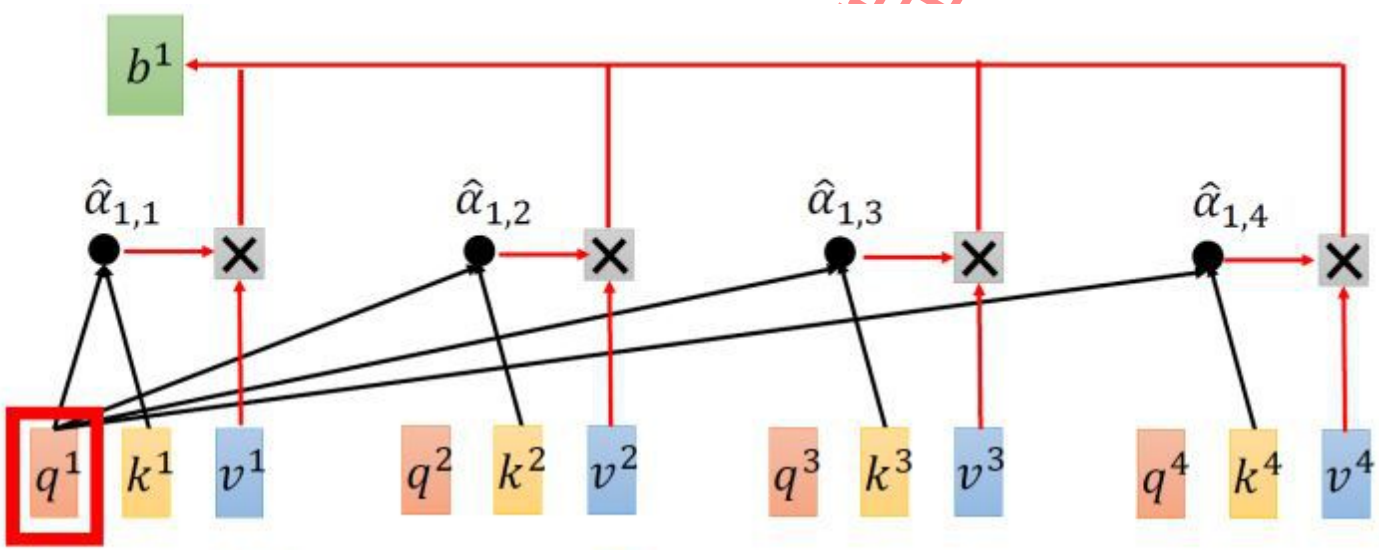
$$\underbrace{q^1 q^2 q^3 q^4}_{Q} = W^q \underbrace{a^1 a^2 a^3 a^4}_{I}$$

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

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



Sequence 组合流程



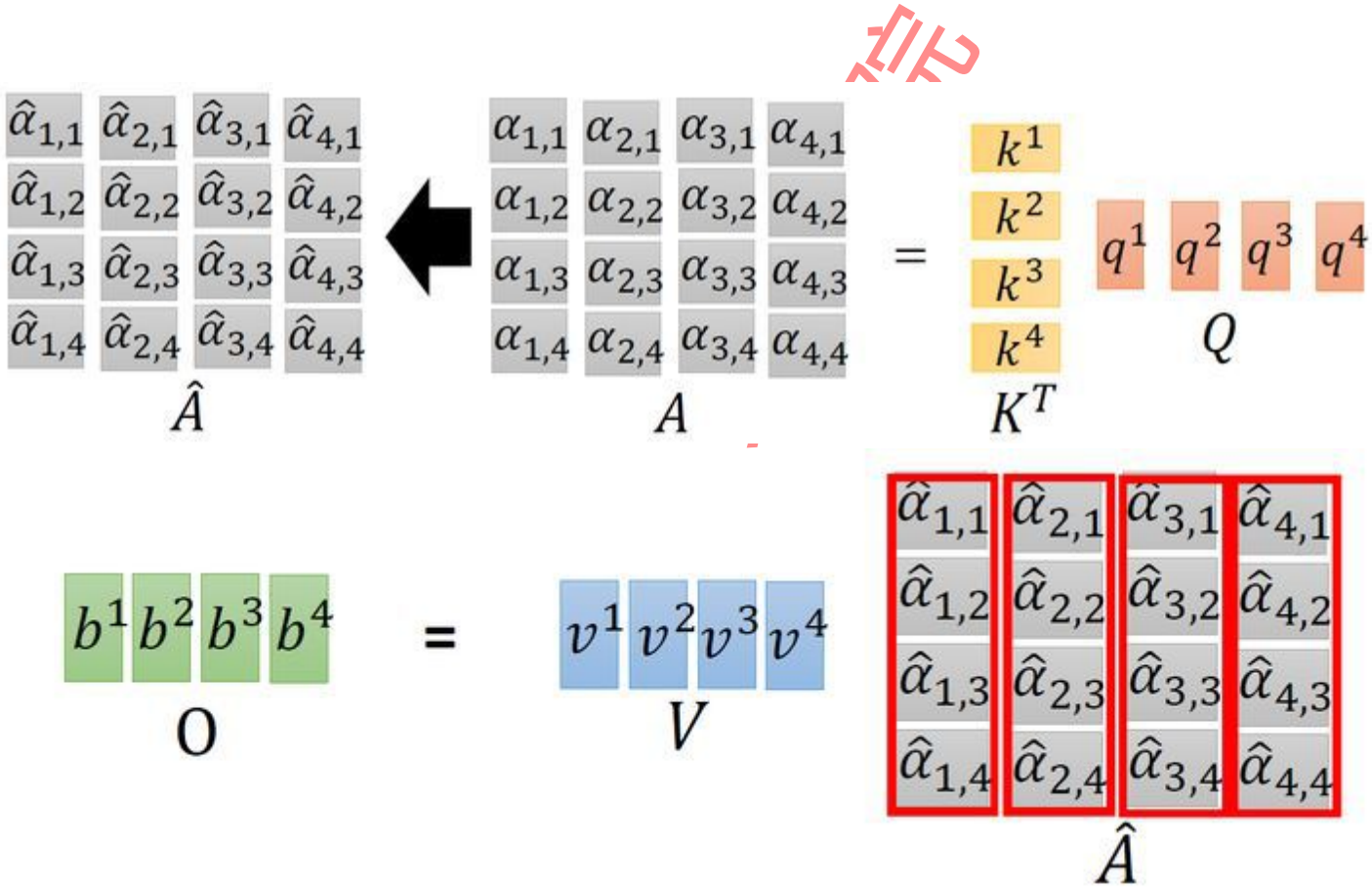
$$\alpha_{1,1} = k^1 q^1 \quad \alpha_{1,2} = k^2 q^1$$

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

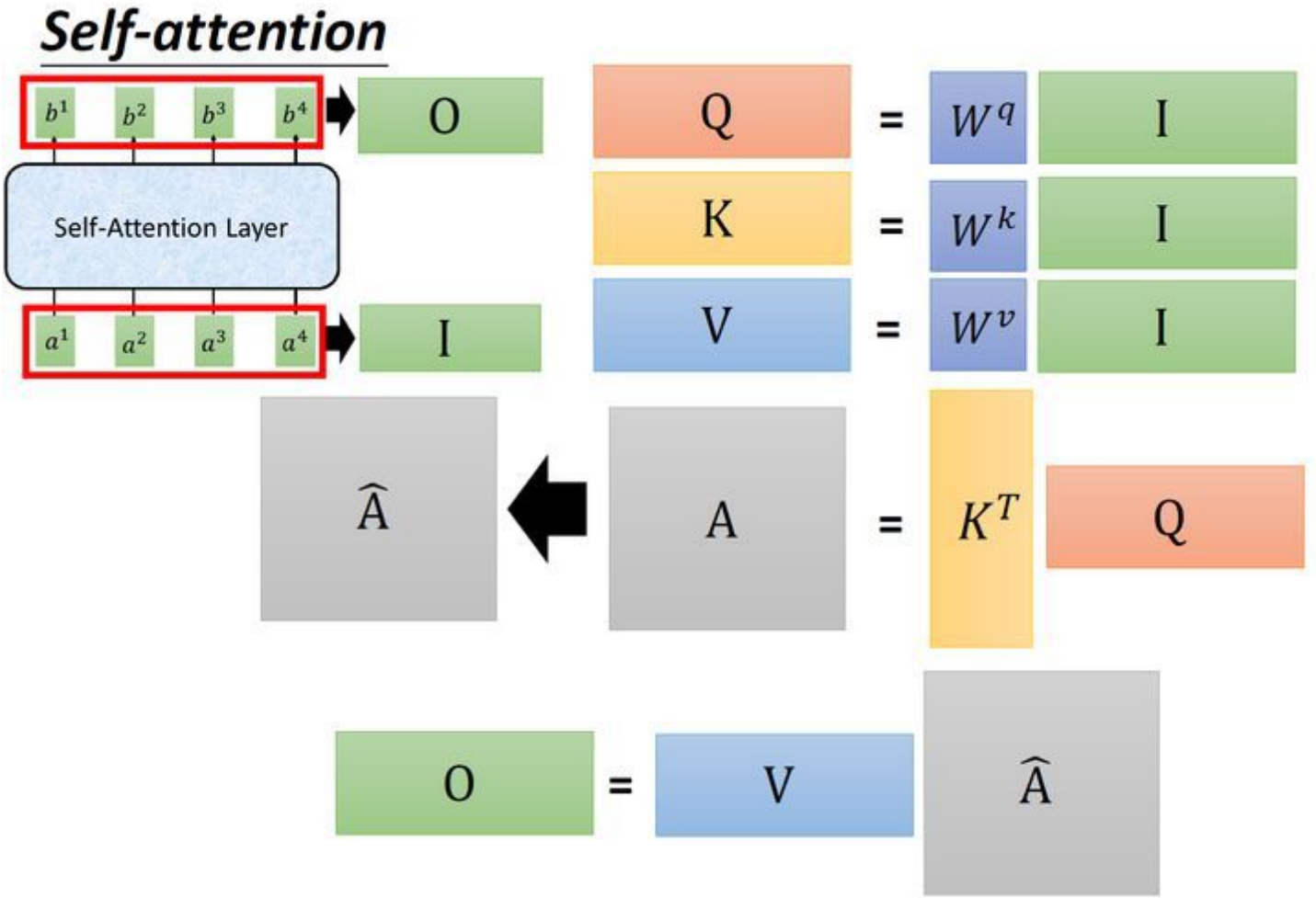
(ignore $\sqrt{d}$ for simplicity)

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

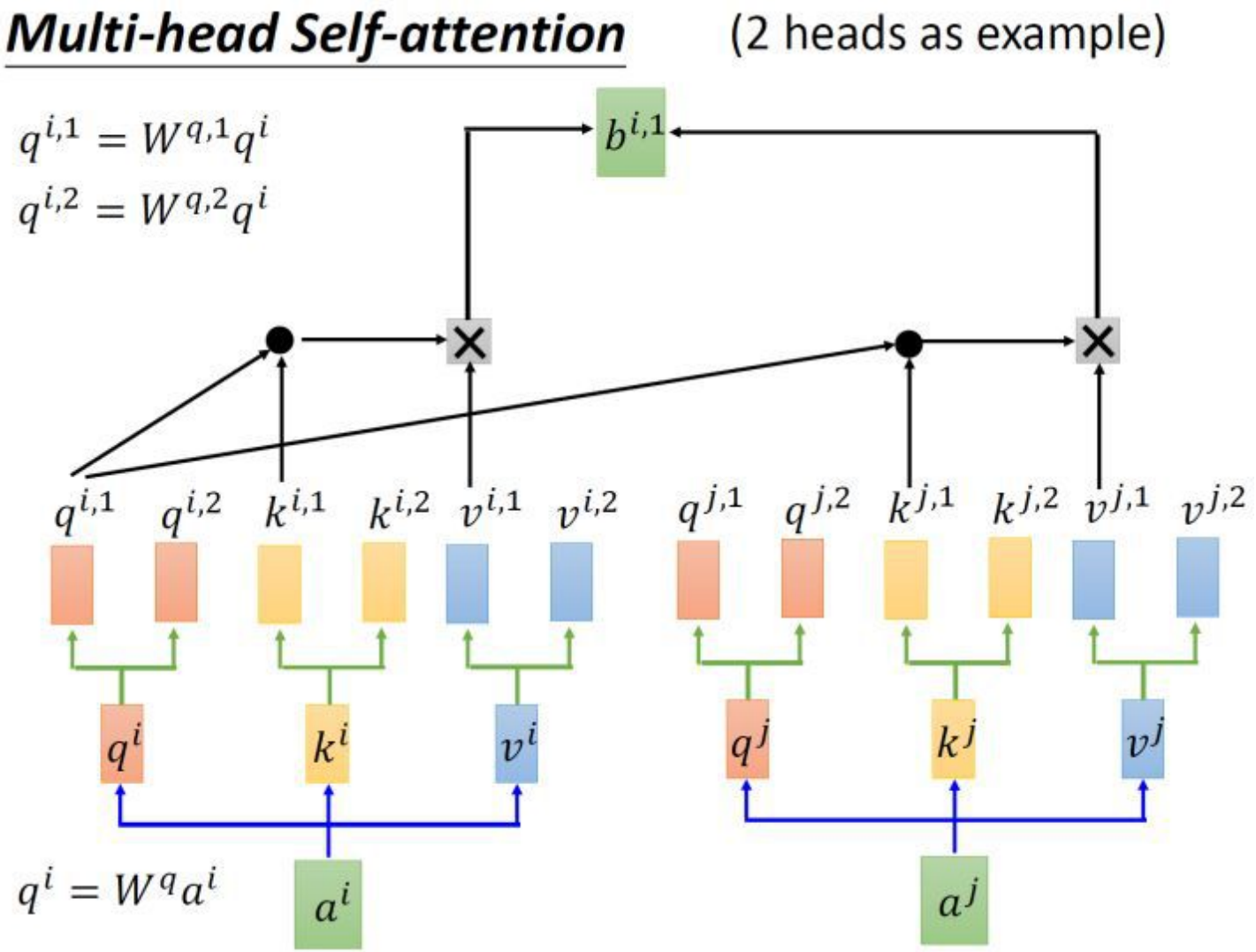
Sequence 组合流程



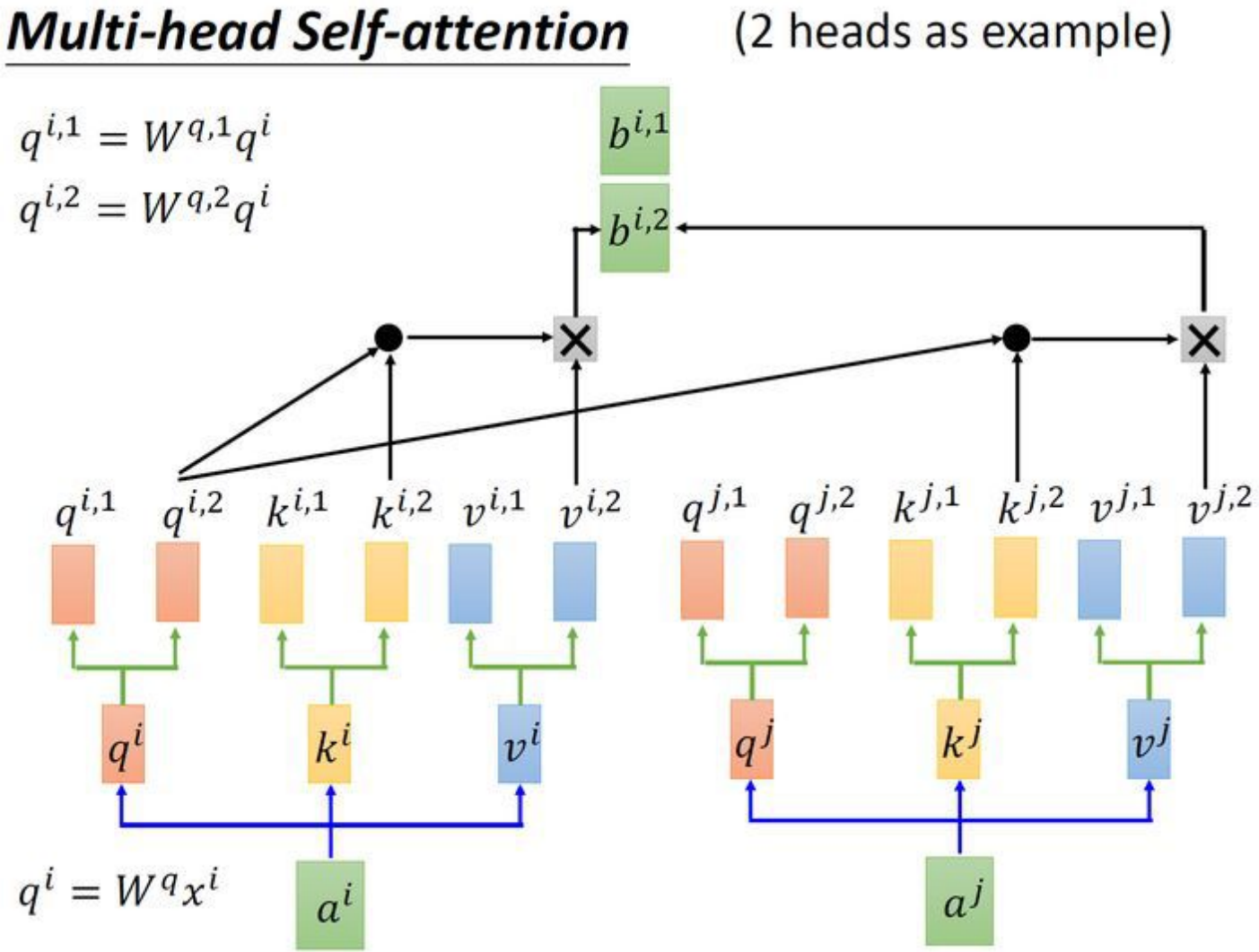
Self-attention 流程



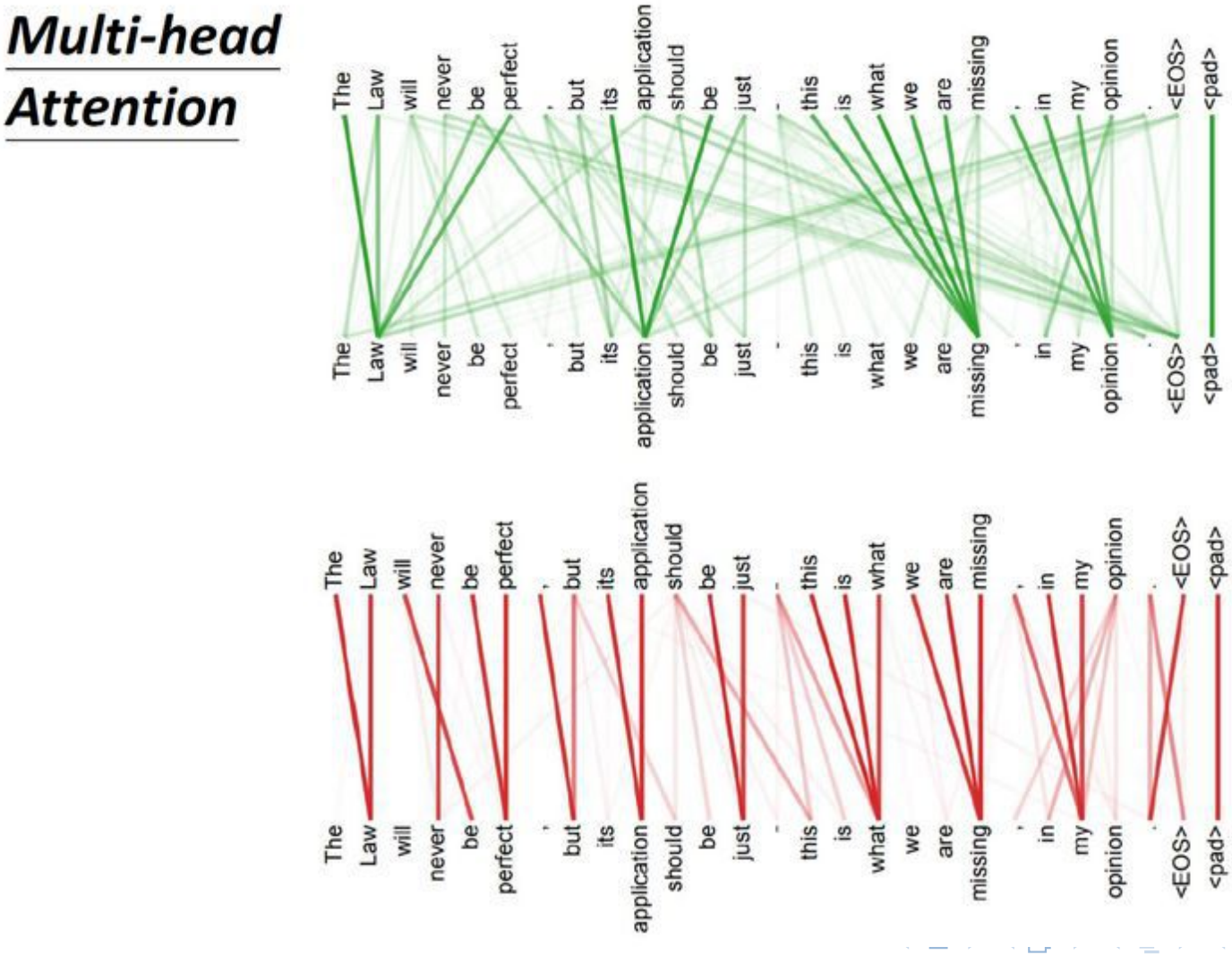
Multi-head Self-attention



Multi-head Self-attention



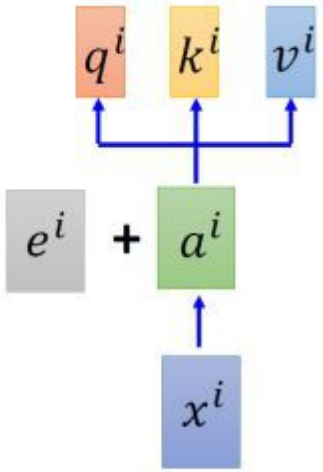
Multi-head Self-attention 案例

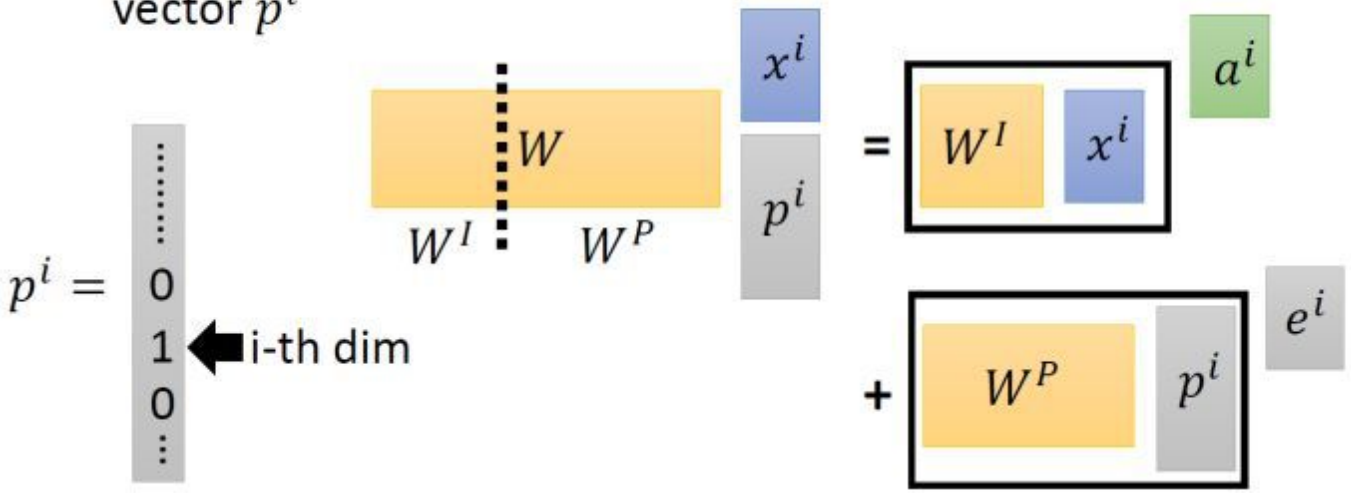


Multi-head positional encoding

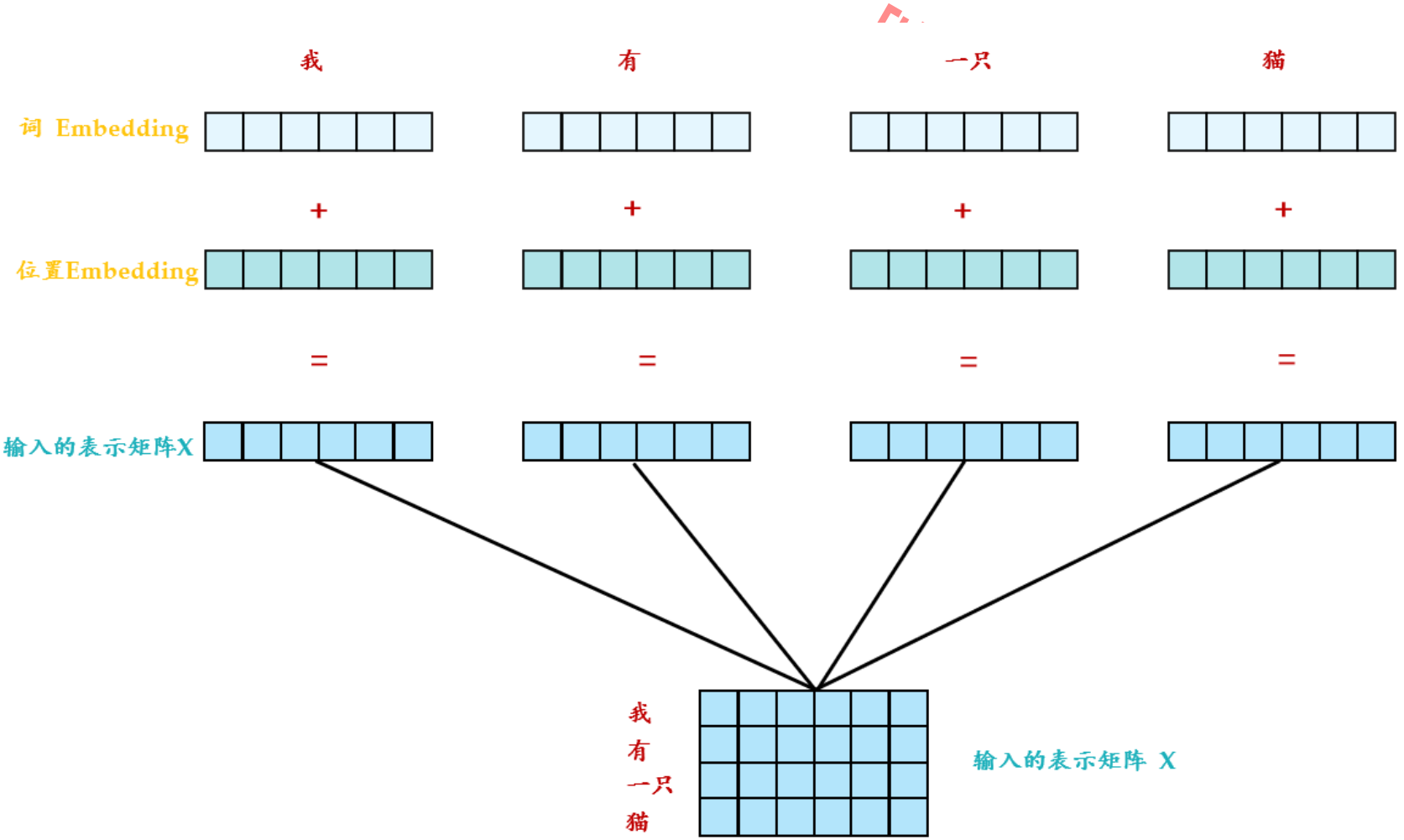
Positional Encoding

- No position information in self-attention.
- Original paper: each position has a unique positional vector e^i (not learned from data)
- In other words: each x^i appends a one-hot vector p^i



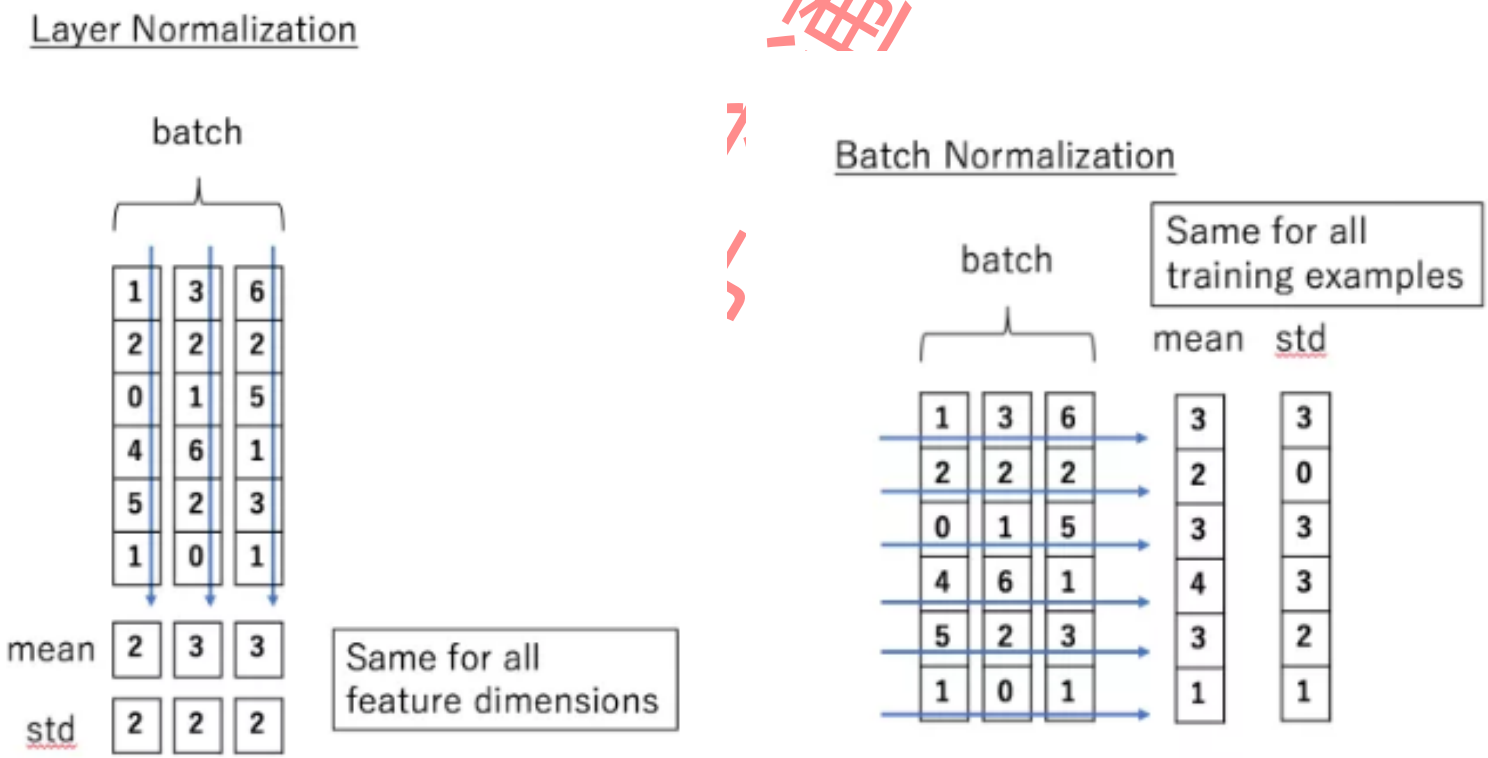


Multi-head positional encoding 案例

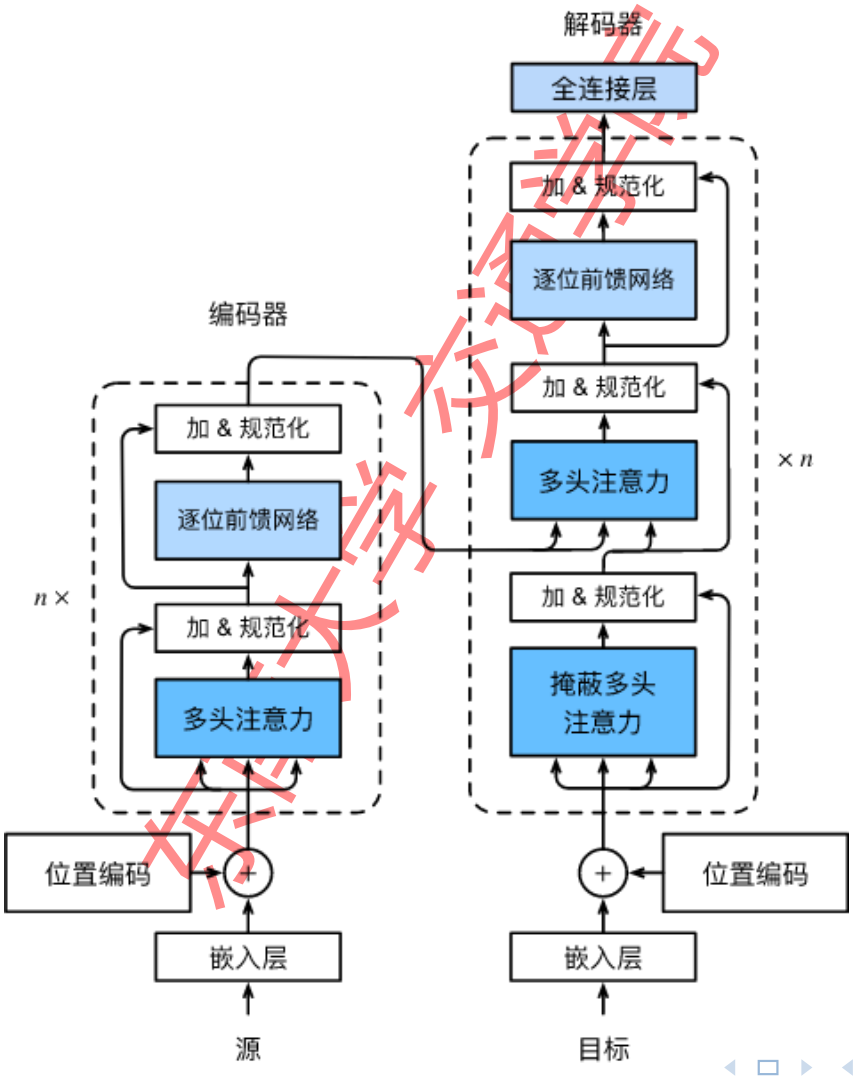


Layer normalization

- 对每一小批数据，在批这个方向上做归一化
- LN 是每一个样本上计算均值与方差
- Batch normalization（BN）是批样本上计算均均值与方差



Decoder 层



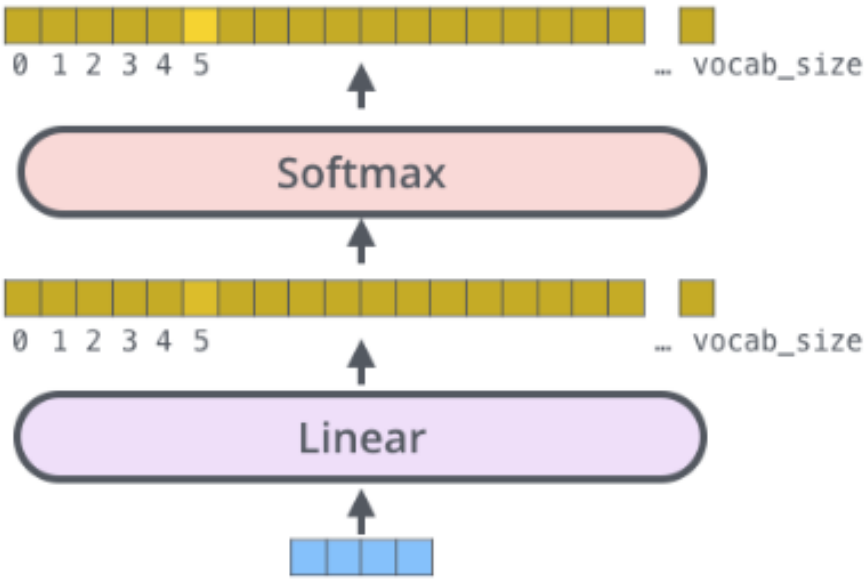
输出层

Which word in our vocabulary is associated with this index?

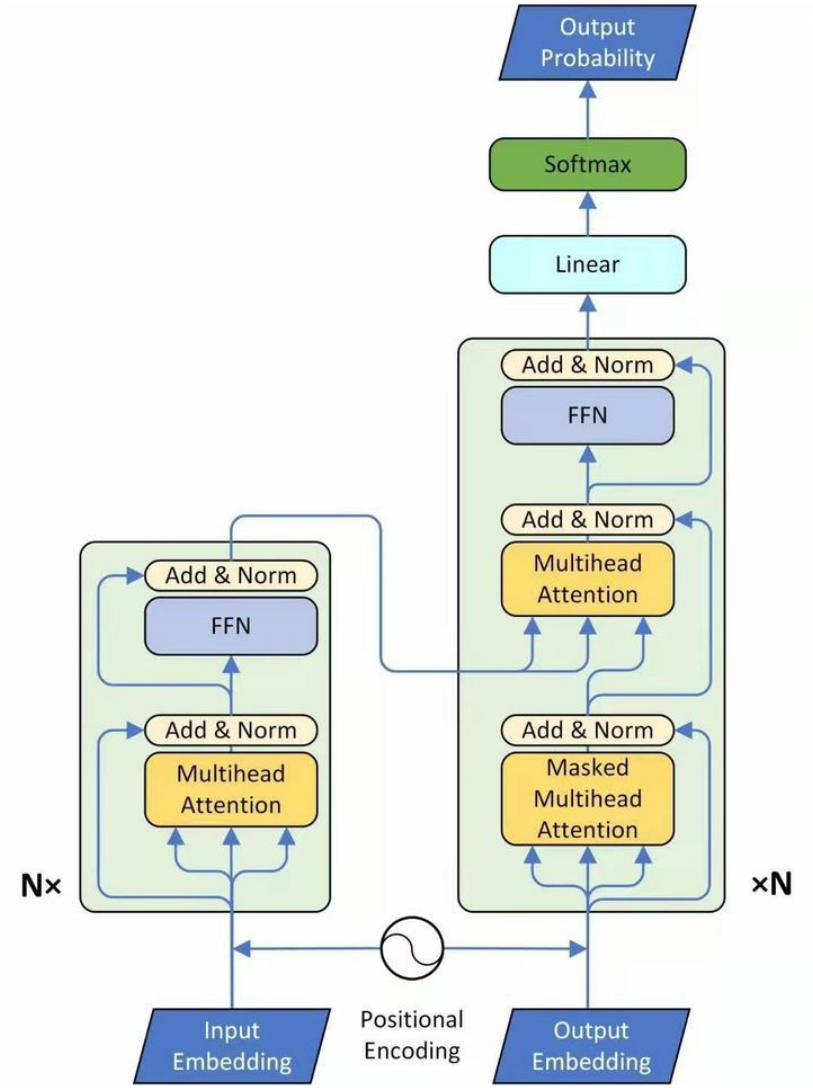
am

Get the index of the cell with the highest value (argmax)

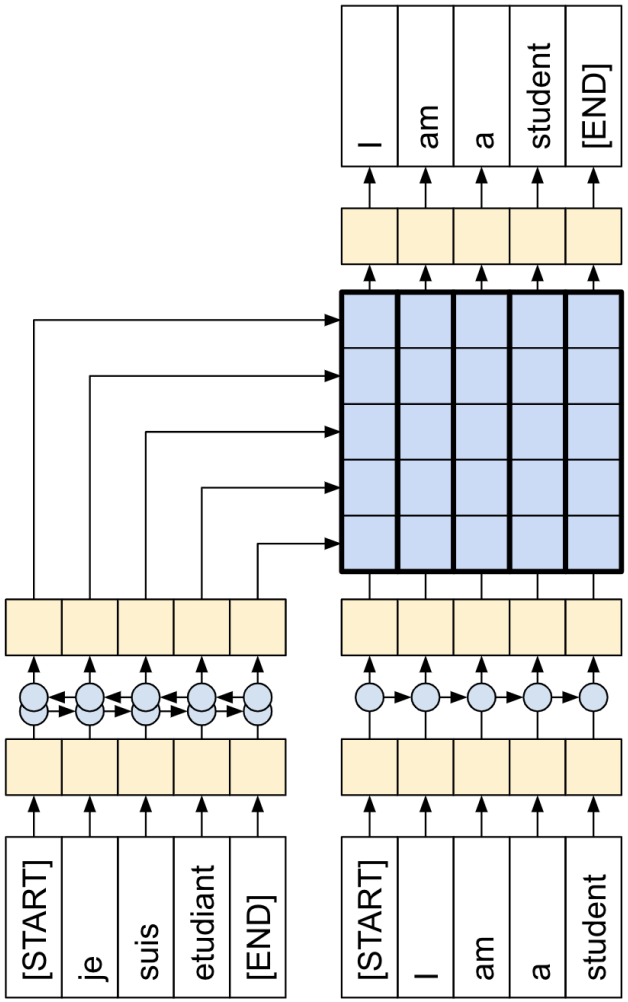
5



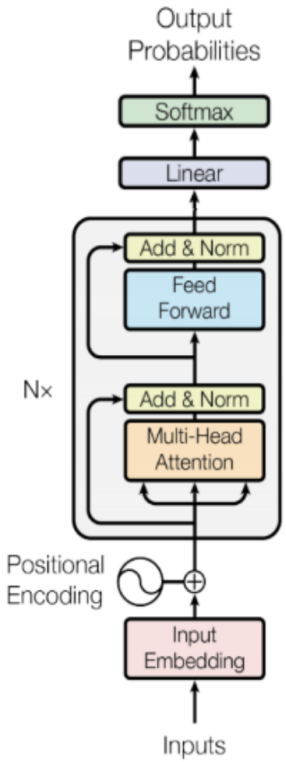
整体模型


Output Probability
Softmax
Linear
Add & Norm
FFN
Add & Norm
Multihead Attention
Add & Norm
Masked Multihead Attention
Add & Norm
FFN
Add & Norm
Multihead Attention
N×
×N
Input Embedding
Positional Encoding
Output Embedding

机器翻译


I
am
a
student
[END]
[START]
je
suis
etudiant
[END]
[START]
I
am
a
student

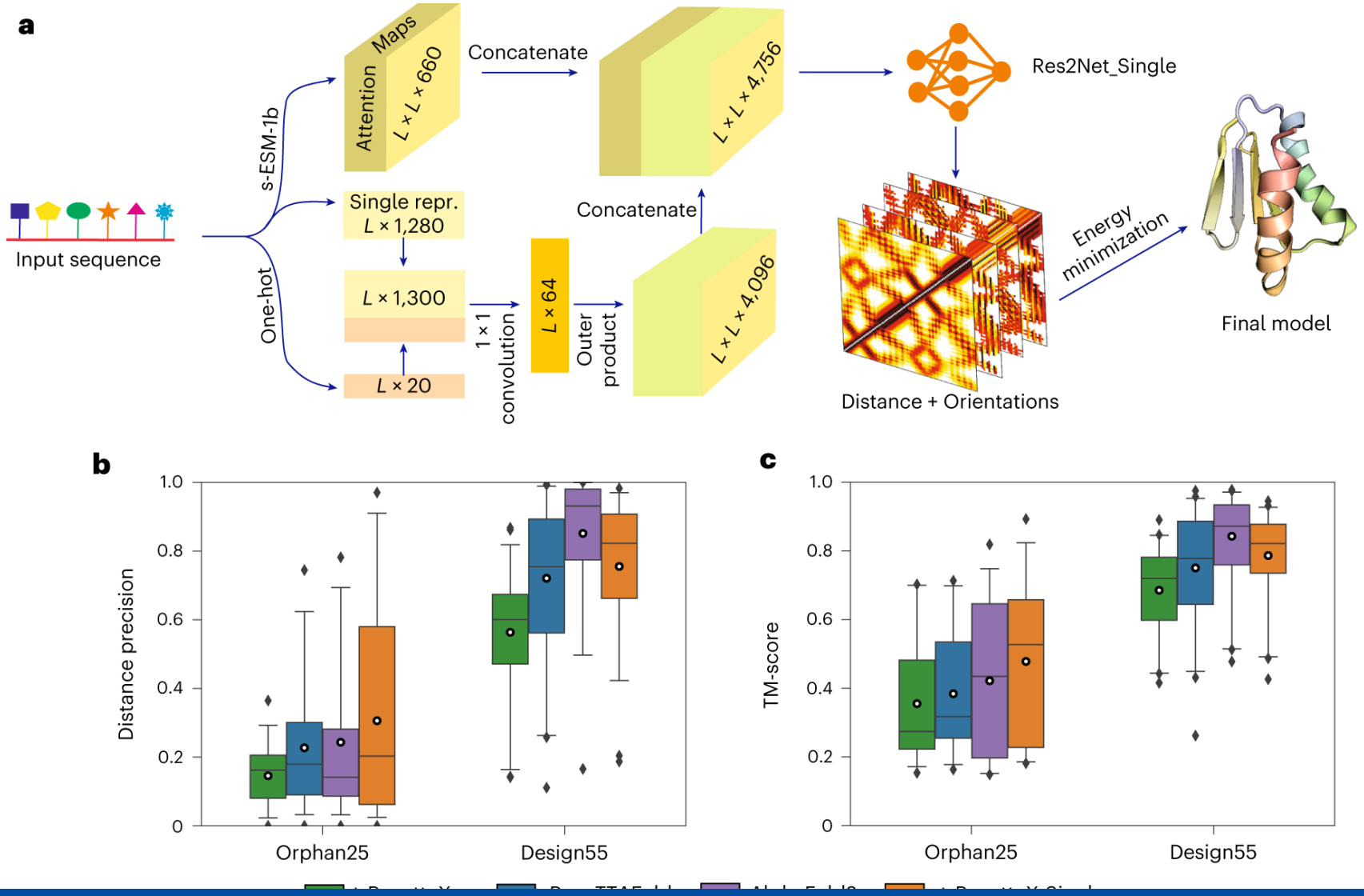
文本总结



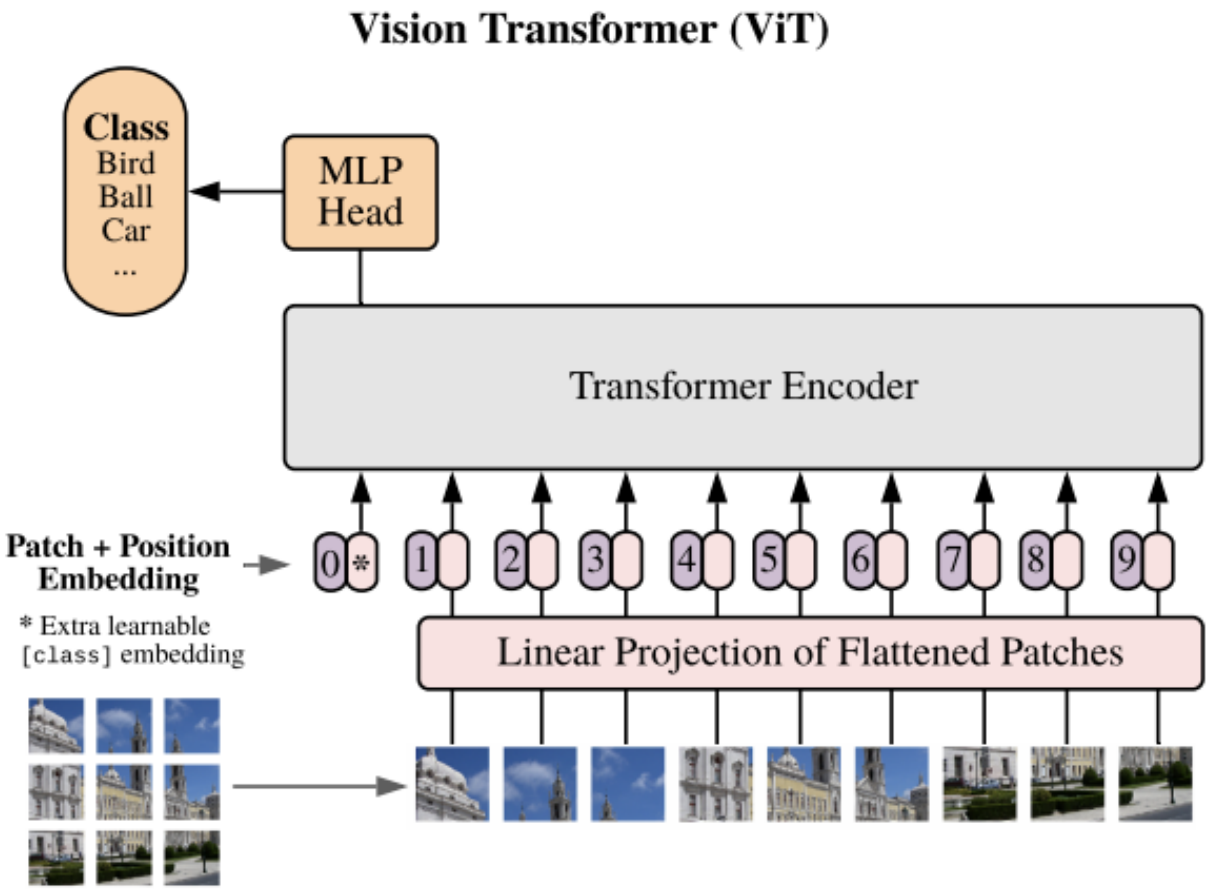
Input:

Output: ⟶ Summary

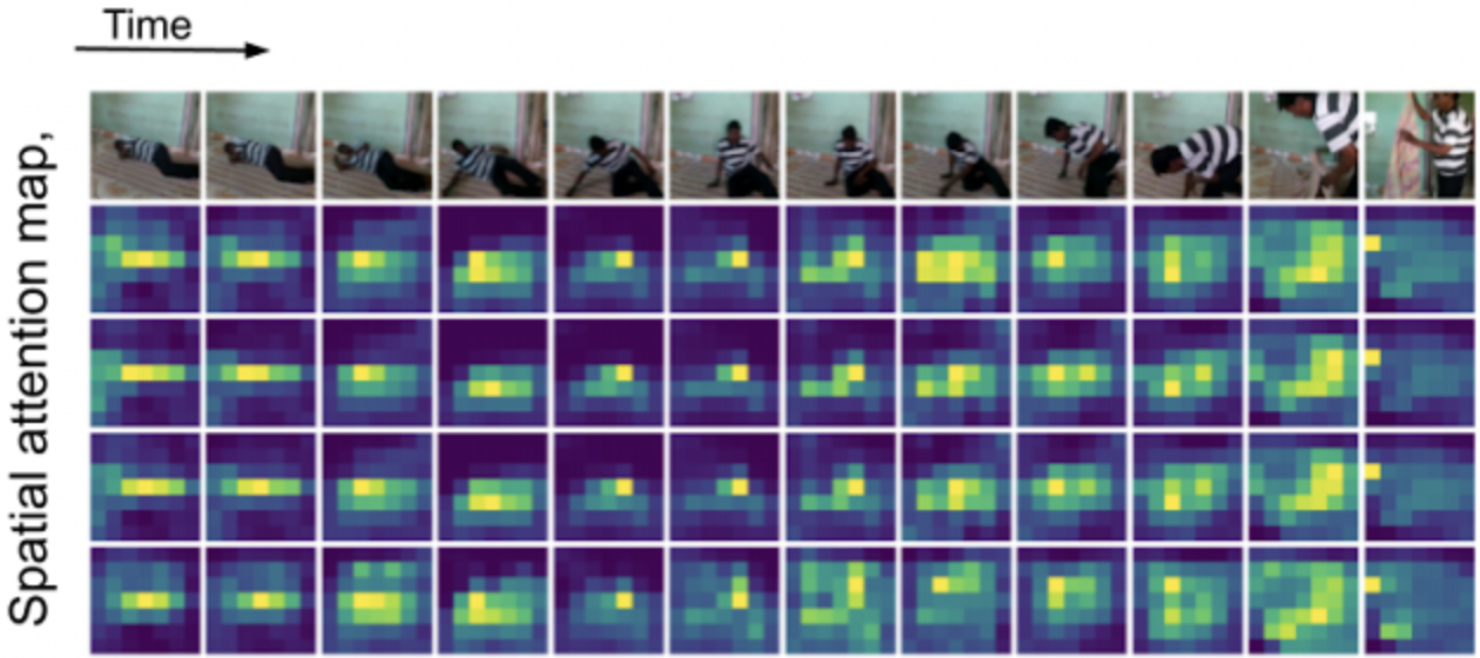
序列数据分析


a
Input sequence
s-ESM-1b
One-hot
Attention Maps
L × L × 660
Single repr.
L × 1,280
L × 1,300
L × 20
1 × 1 convolution
L × 64
Outer product
L × L × 4,096
Concatenate
L × L × 4,756
Concatenate
Res2Net_Single
Distance + Orientations
Energy minimization
Final model
b
Distance precision
1.0
0.8
0.6
0.4
0.2
0
Orphan25
Design55
c
TM-score
1.0
0.8
0.6
0.4
0.2
0
Orphan25
Design55

图像与视频处理


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP
Head
Transformer Encoder
Patch + Position
Embedding
* Extra learnable
[class] embedding
0 *
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

图像与视频处理



Visualization of the spatial attention maps in TokenLearner, over time. As the person is moving in the scene, TokenLearner pays attention to different spatial locations to tokenize.