自然语言处理

8. Transformer与预训练模型

虞剑飞

南京大学智能科学与技术学院

2025.4.30

本章内容

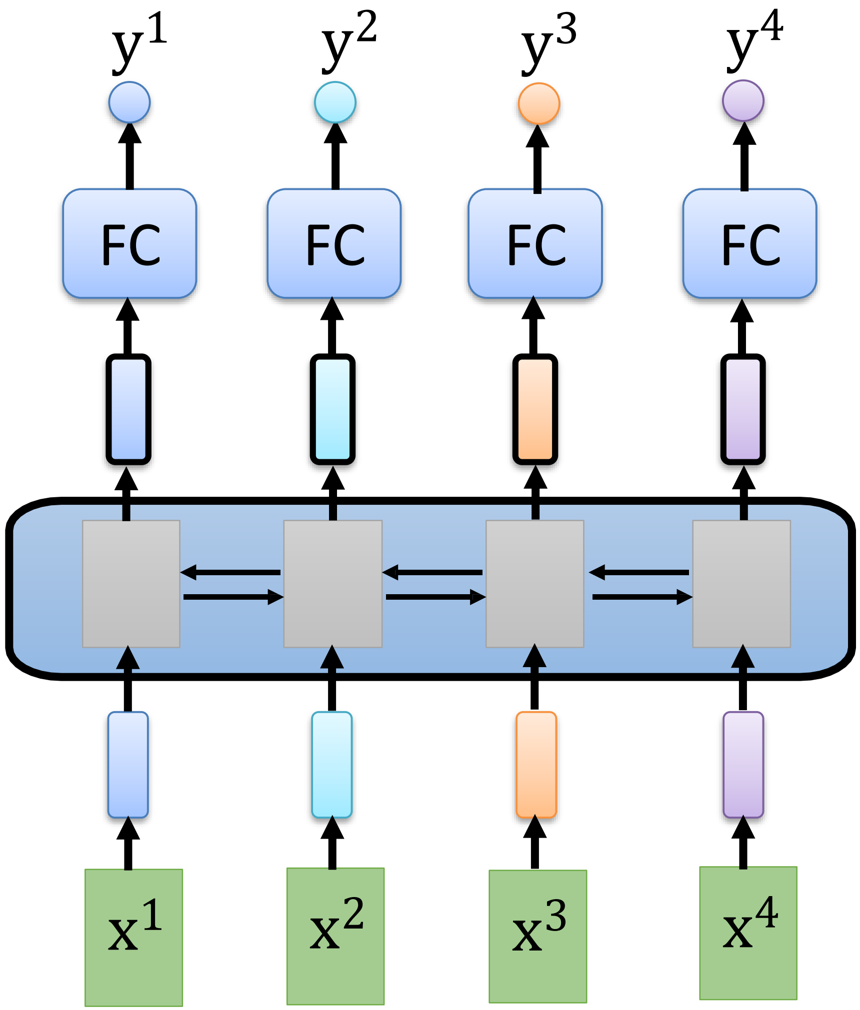
- 8.1 引言
- 8.2 自注意力机制
- 8.3 Transformer网络架构
- 8.4 预训练模型及其应用
  - 8.4.1 基于Transformer编码器的预训练语言模型
  - 8.4.2 基于Transformer的预训练Seq2Seq模型
  - 8.4.2 基于Transformer解码器的预训练语言模型
  - 8.4.3 基于Transformer的文本-视觉预训练模型
- 8.5 本章小结

本章内容

- 8.1 引言
- 8.2 自注意力机制
- 8.3 Transformer网络架构
- 8.4 预训练模型及其应用
  - 8.4.1 基于Transformer编码器的预训练语言模型
  - 8.4.2 基于Transformer的预训练Seq2Seq模型
  - 8.4.2 基于Transformer解码器的预训练语言模型
  - 8.4.3 基于Transformer的文本-视觉预训练模型
- 8.5 本章小结

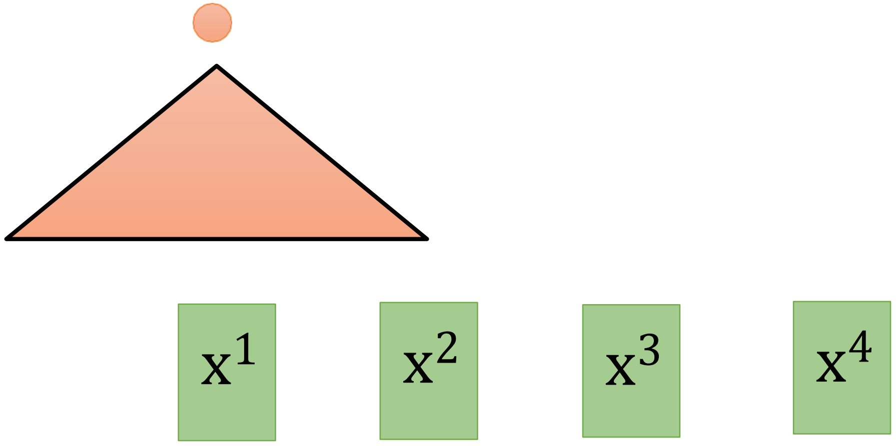
引言

- 循环神经网络缺陷
  - 很难并行化



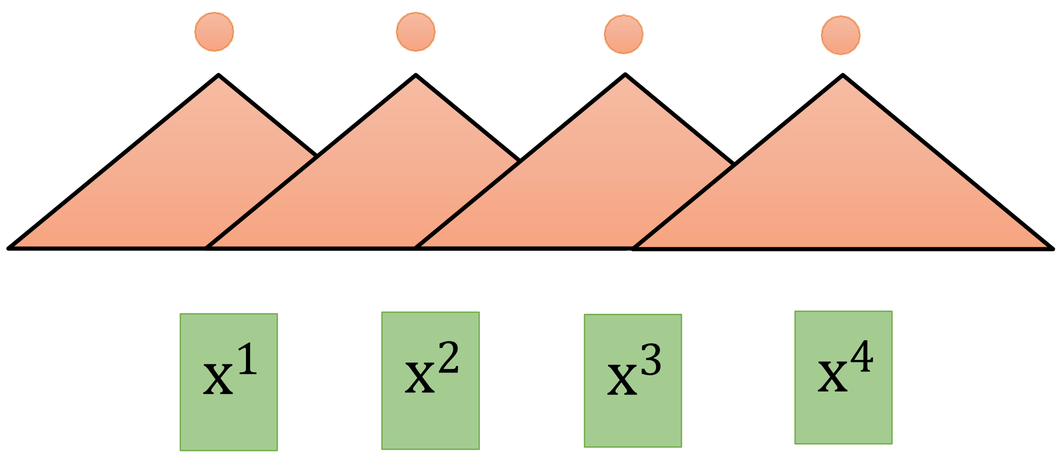
引言

- 循环神经网络缺陷
  - 很难并行化
    - 方案1：使用CNN来替换RNN



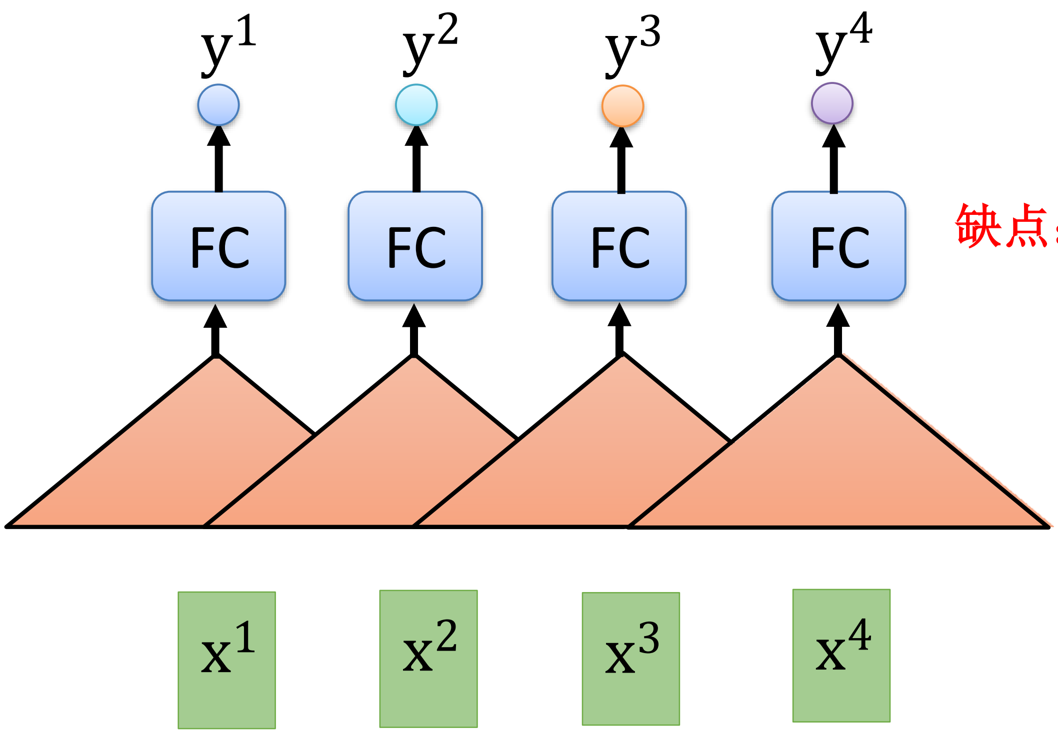
引言

- 循环神经网络缺陷
  - 很难并行化
    - 方案1：使用CNN来替换RNN



引言

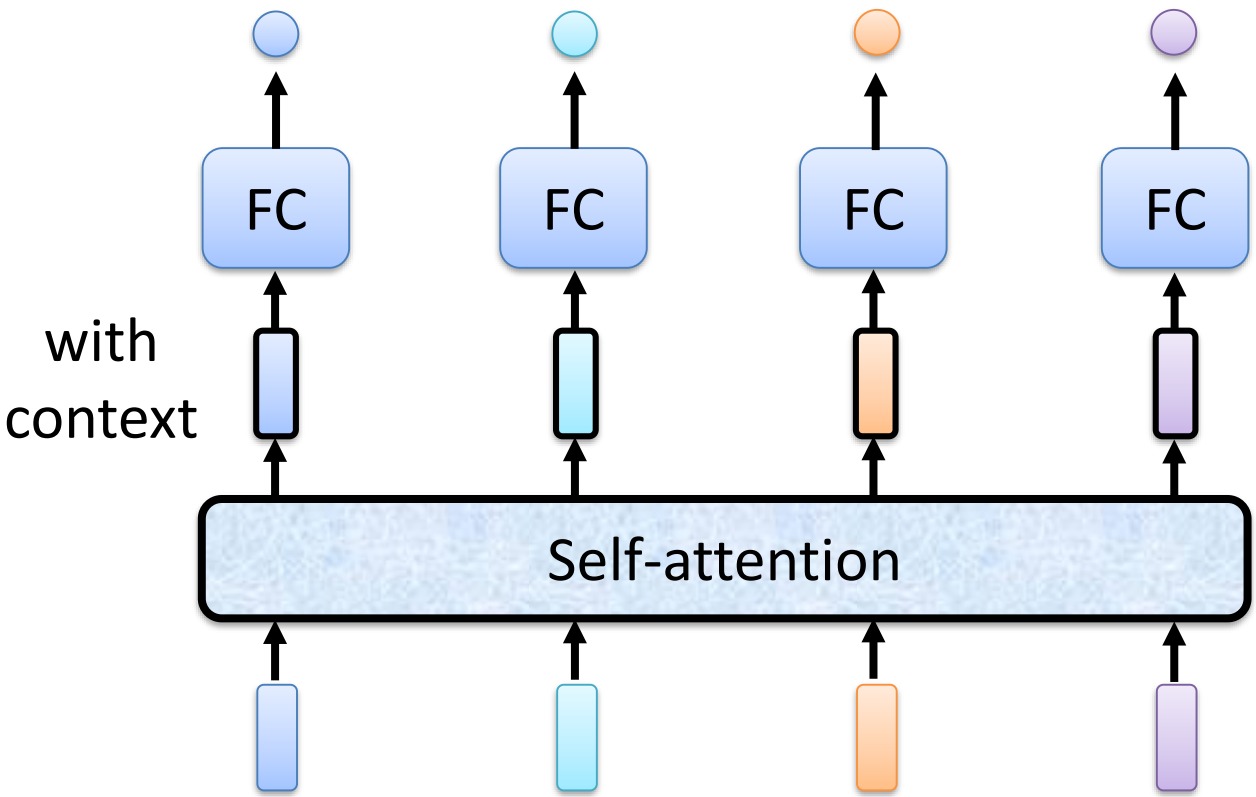
- 循环神经网络缺陷
  - 很难并行化
    - 方案1：使用CNN来替换RNN



缺点：只能捕获局部信息

引言

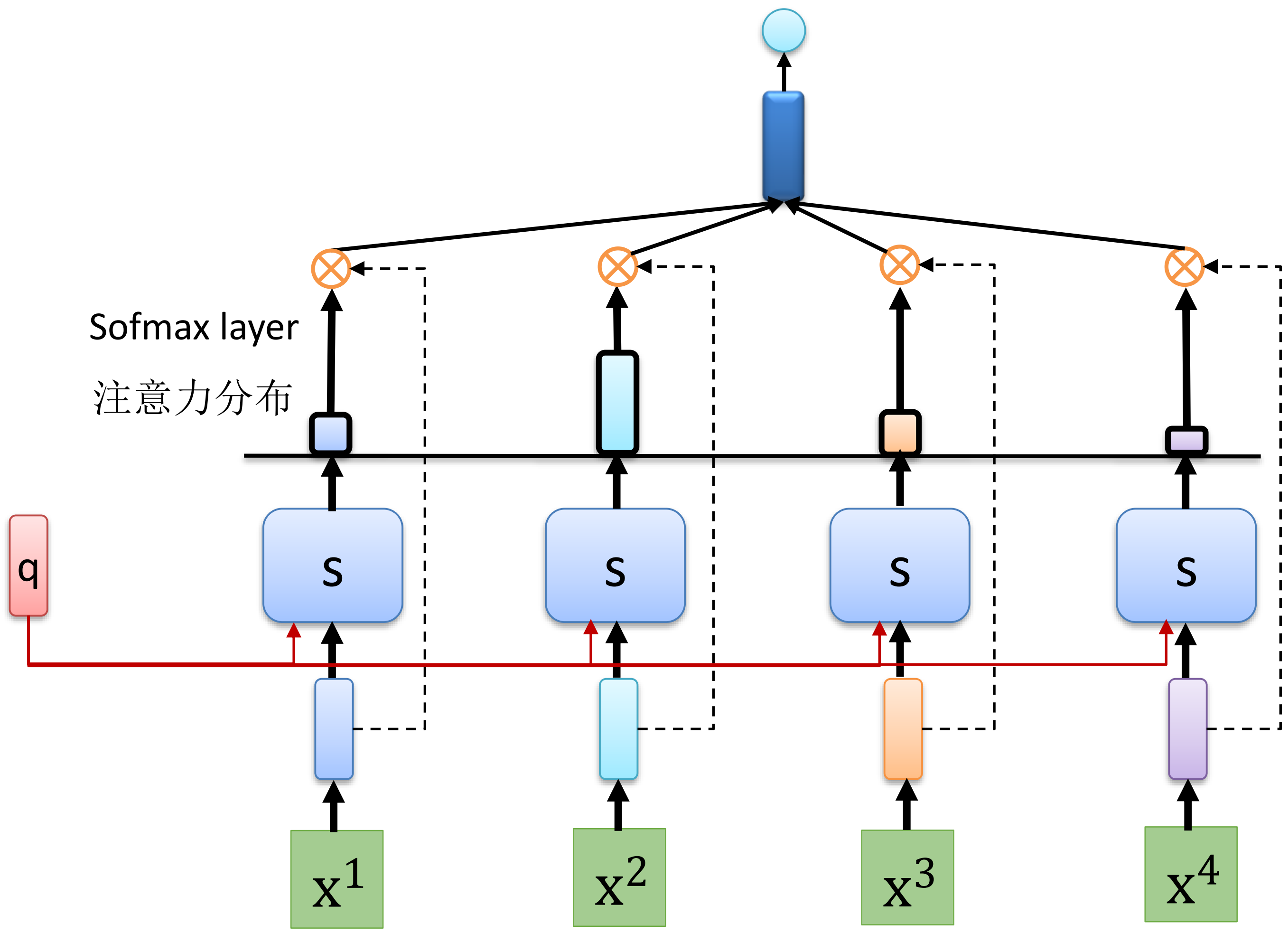
- 循环神经网络缺陷
  - 很难并行化
    - 方案2：自注意力机制（Self-Attention Mechanism）



本章内容

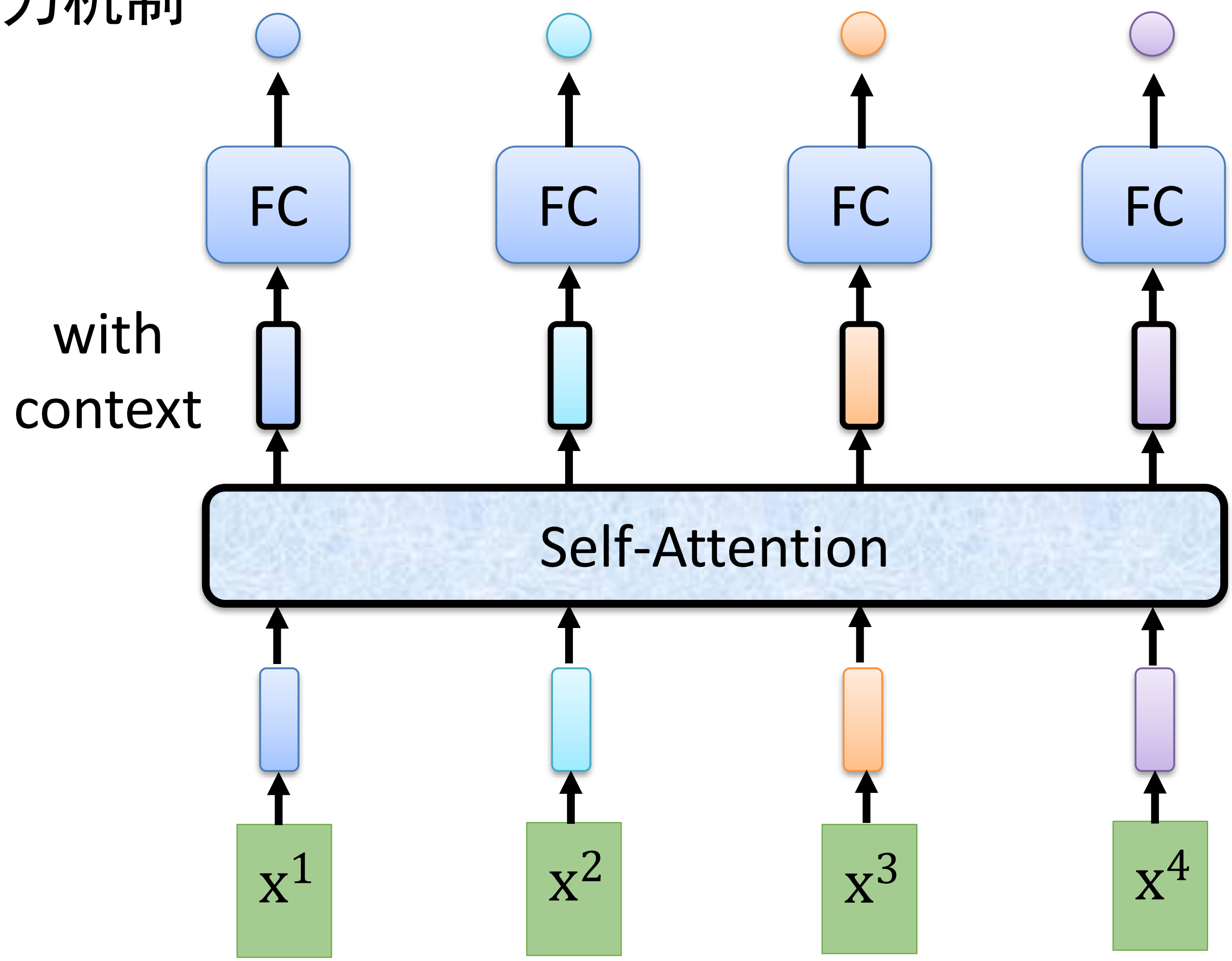
自注意力机制 (Self-Attention)

- 注意力机制回顾



自注意力机制 (Self-Attention)

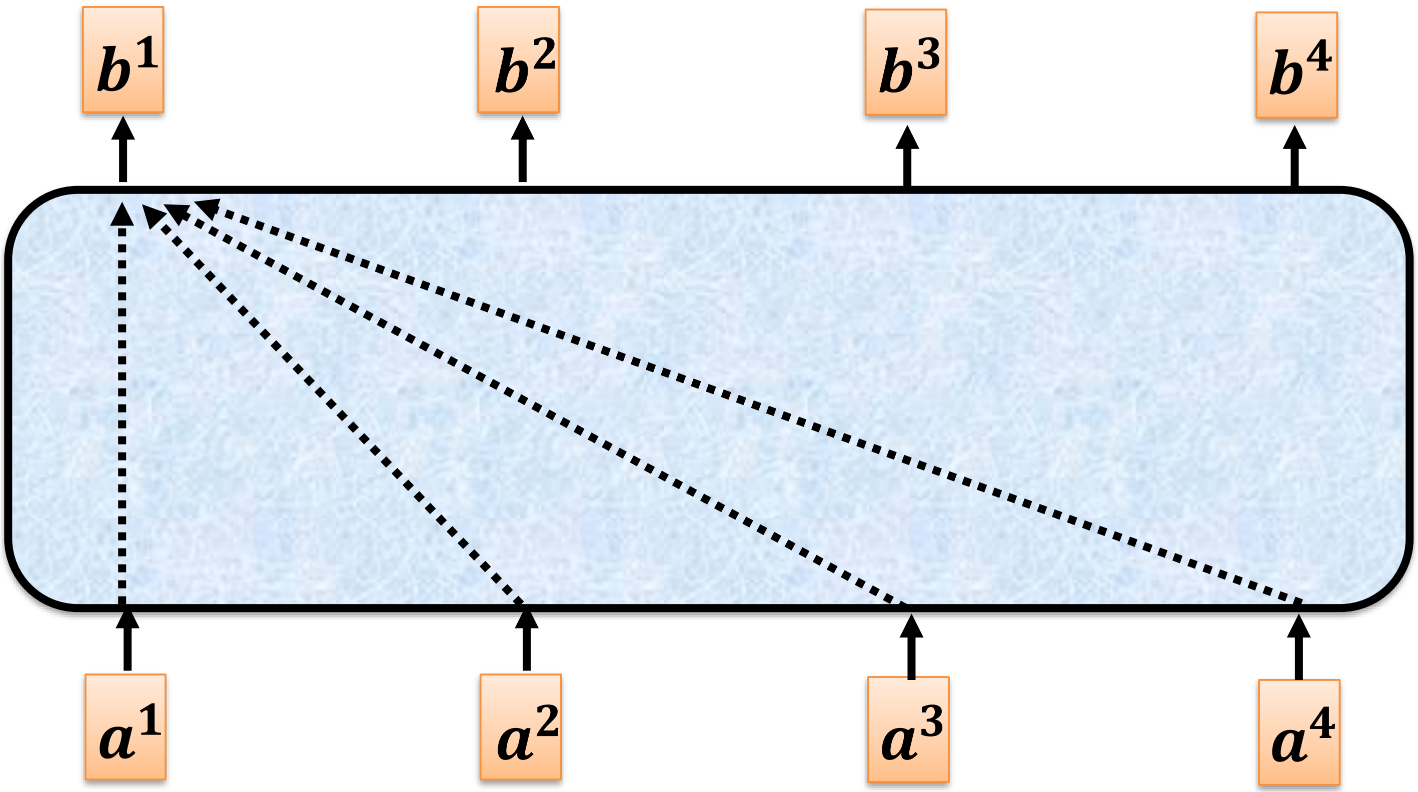
- 自注意力机制



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I.. Attention is all you need. NIPS 2017.

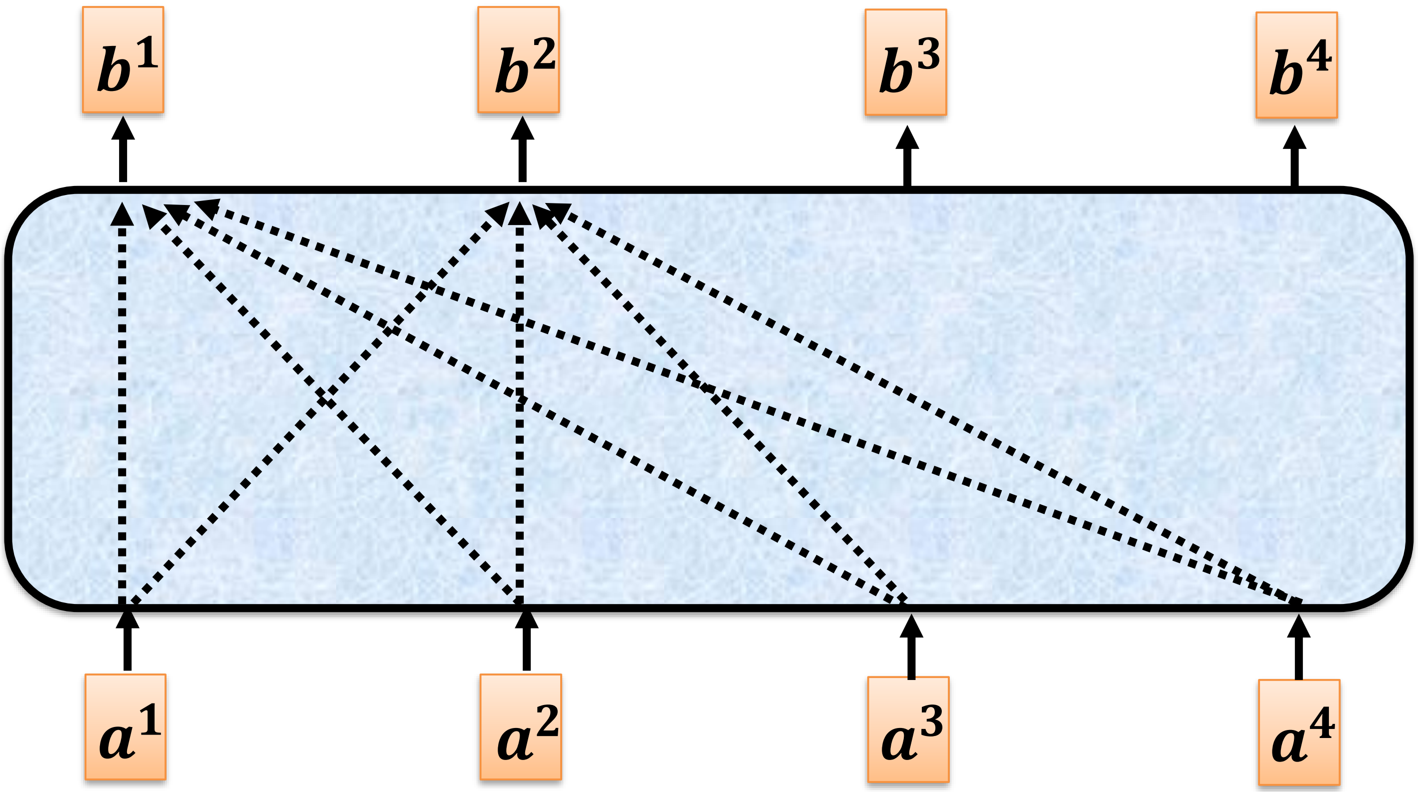
自注意力机制 (Self-Attention)

- 自注意力机制



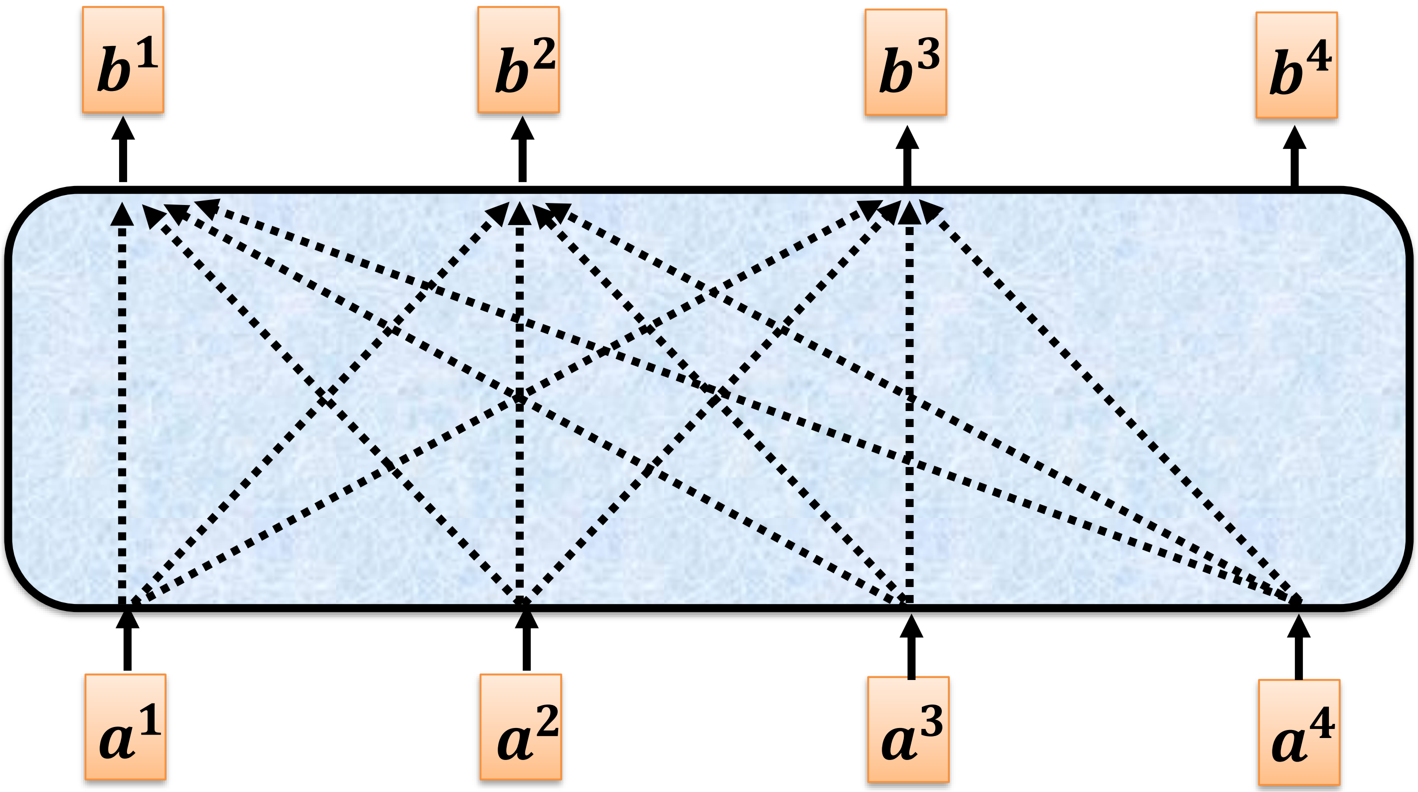
自注意力机制 (Self-Attention)

- 自注意力机制



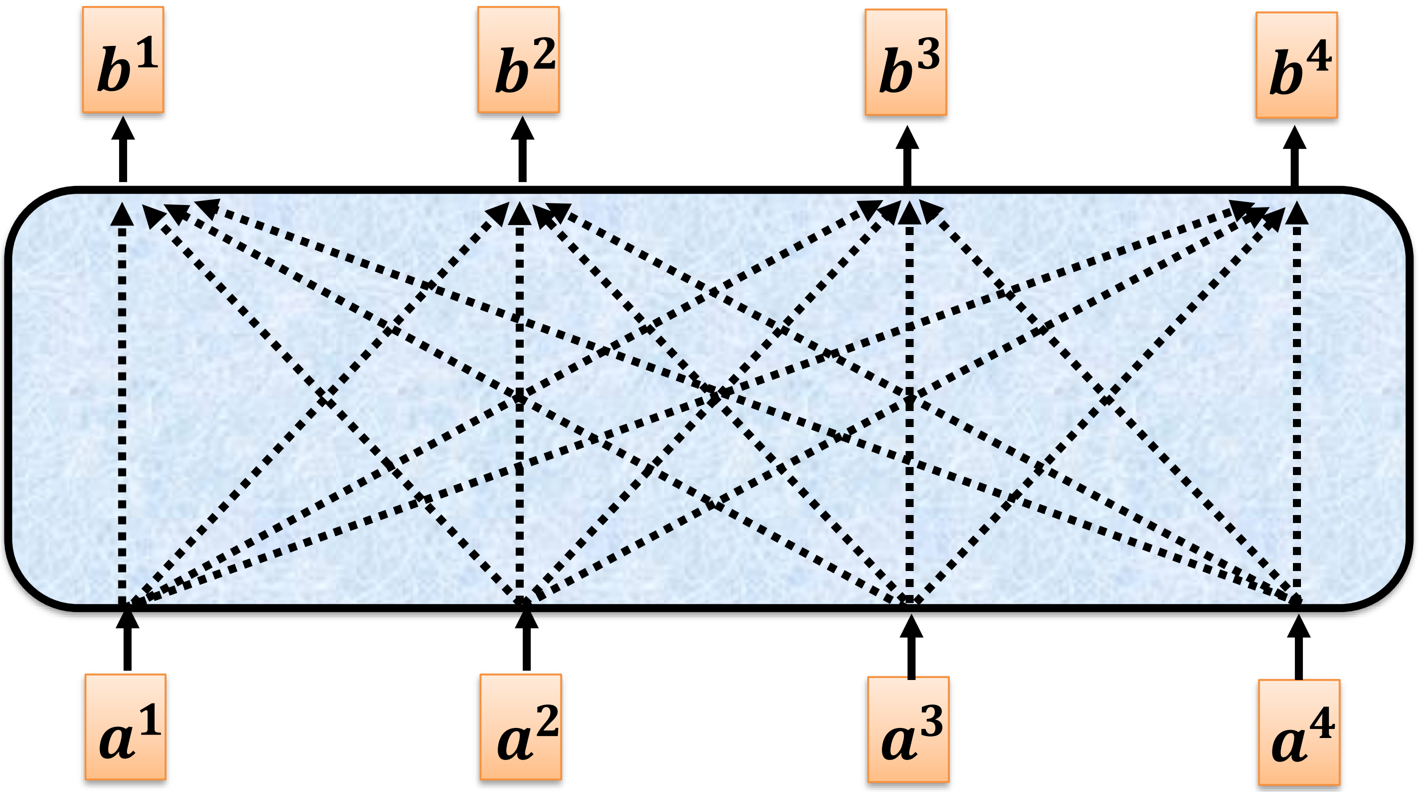
自注意力机制 (Self-Attention)

- 自注意力机制



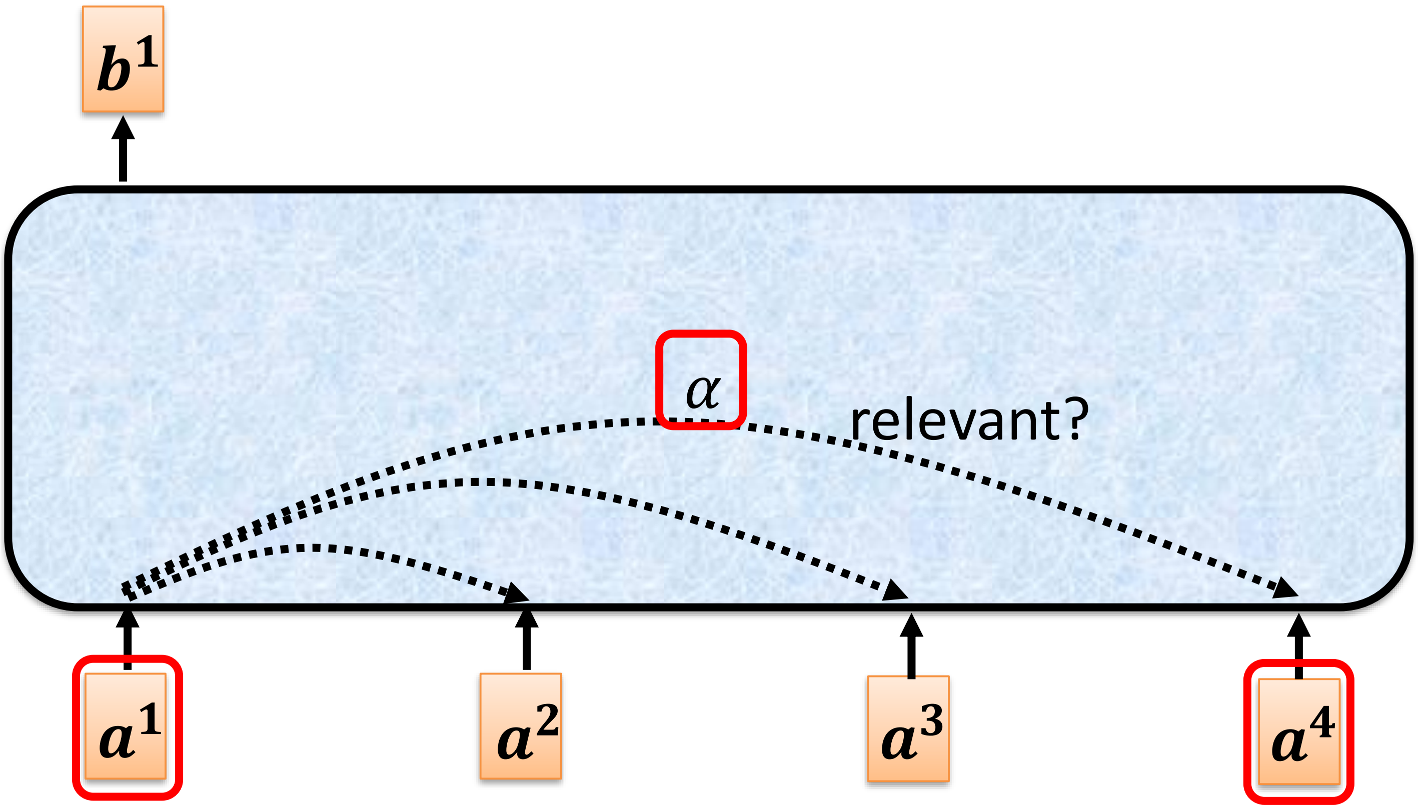
自注意力机制 (Self-Attention)

- 自注意力机制



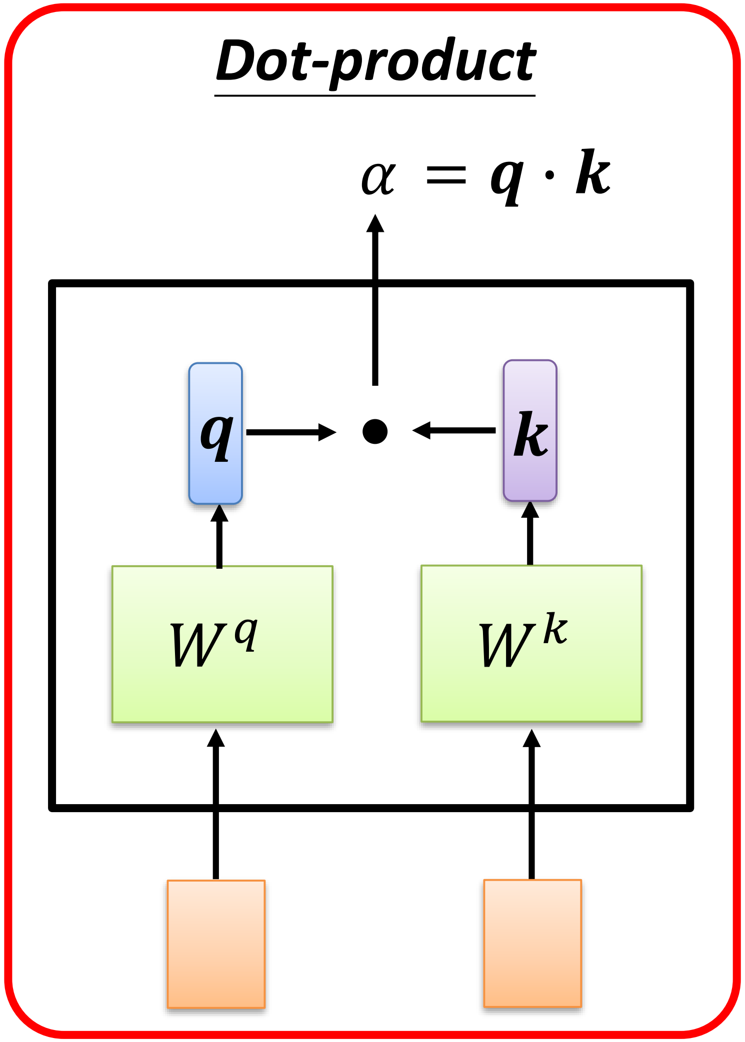
自注意力机制 (Self-Attention)

- 自注意力机制



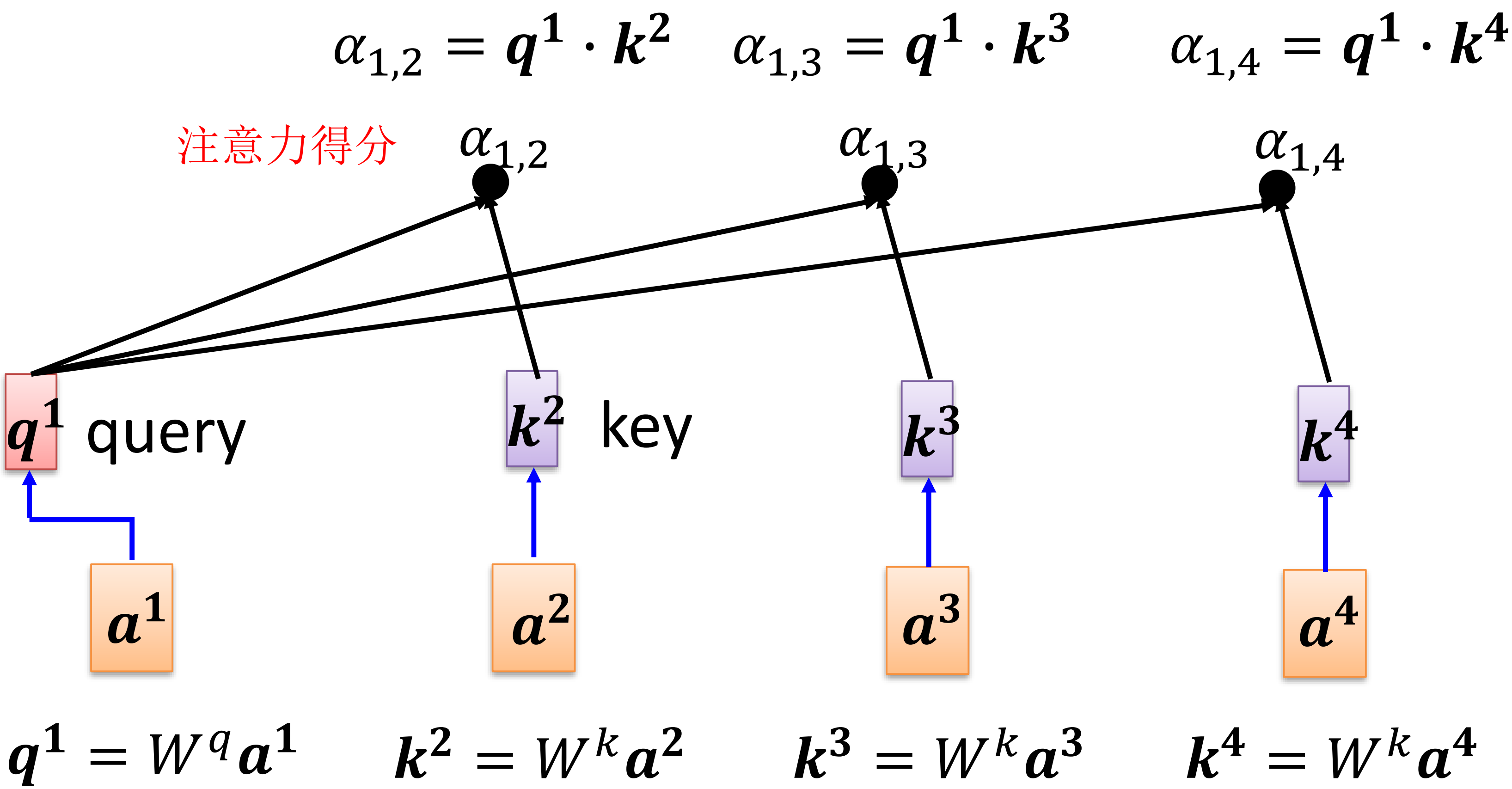
自注意力机制 (Self-Attention)

- 自注意力机制



自注意力机制 (Self-Attention)

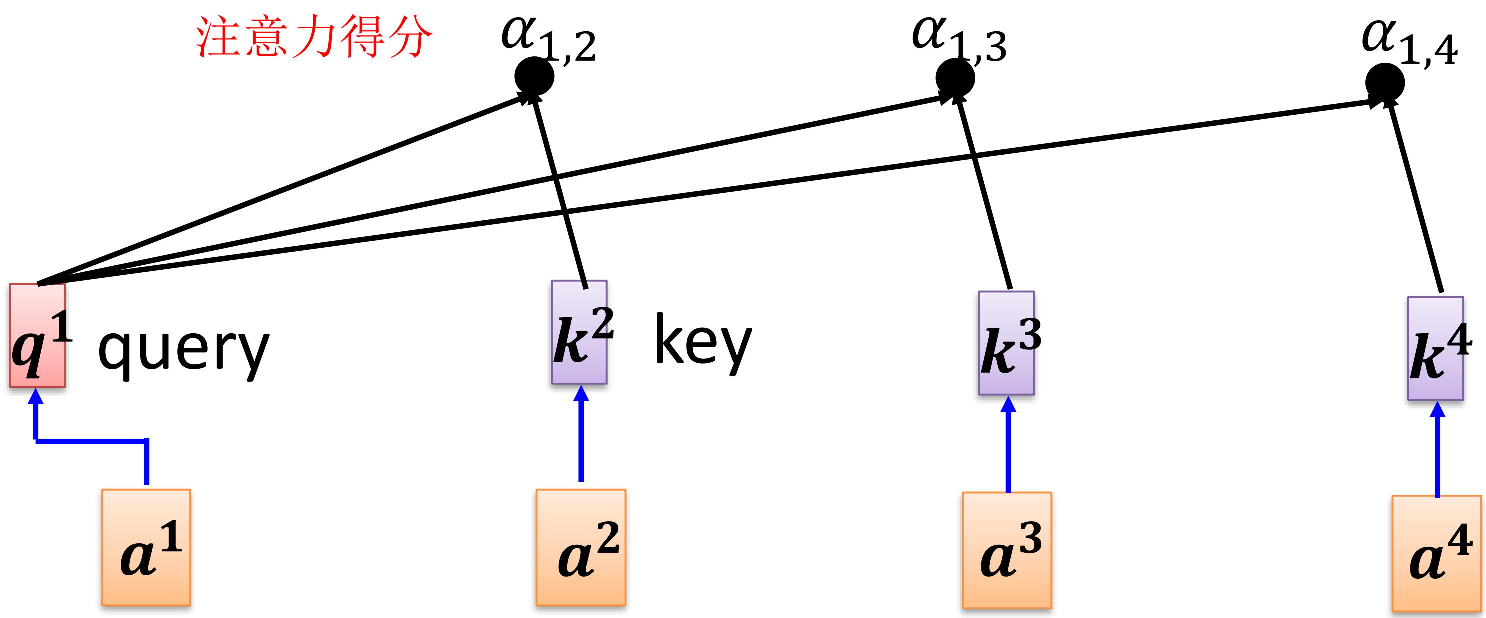
- 自注意力机制



自注意力机制 (Self-Attention)

- 自注意力机制

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

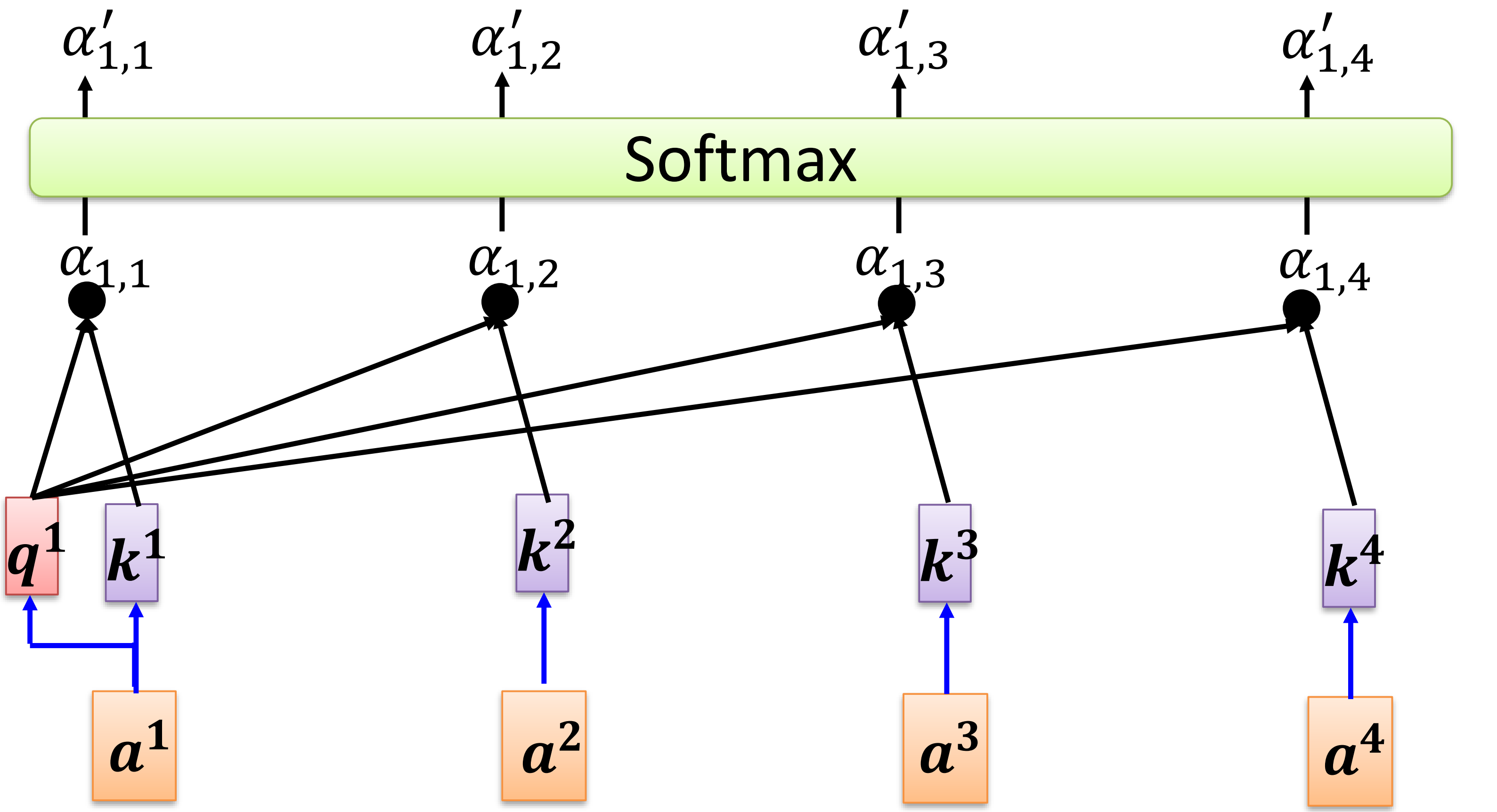


$\boldsymbol{q}^1 = W^q \boldsymbol{a}^1 \quad \boldsymbol{k}^2 = W^k \boldsymbol{a}^2 \quad \boldsymbol{k}^3 = W^k \boldsymbol{a}^3 \quad \boldsymbol{k}^4 = W^k \boldsymbol{a}^4$

$\boldsymbol{k}^1 = W^k \boldsymbol{a}^1$

自注意力机制 (Self-Attention)

- 自注意力机制



$\boldsymbol{q}^1 = W^q \boldsymbol{a}^1$
$\boldsymbol{k}^1 = W^k \boldsymbol{a}^1$

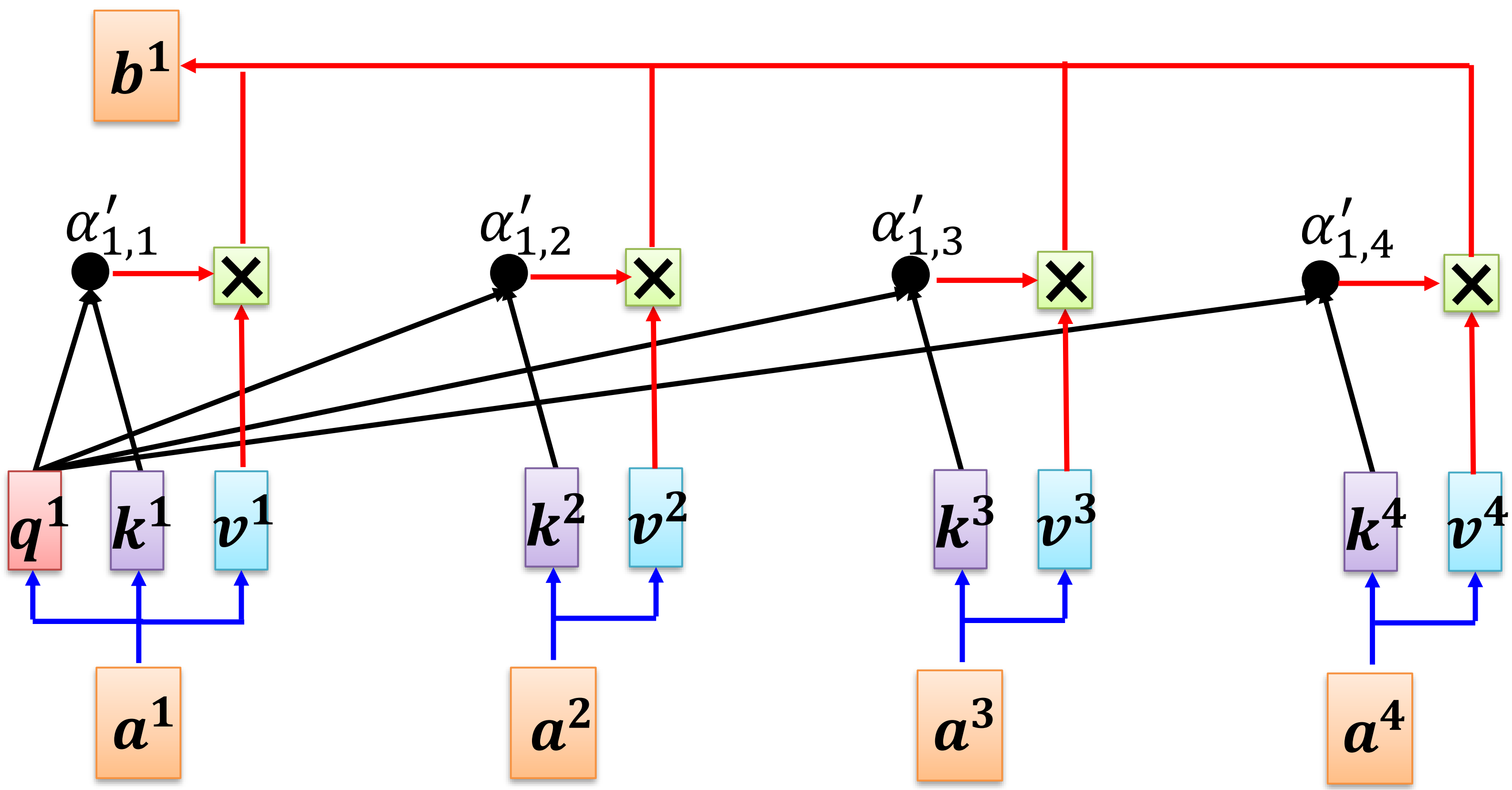
$\boldsymbol{k}^2 = W^k \boldsymbol{a}^2$

$\boldsymbol{k}^3 = W^k \boldsymbol{a}^3$

$\boldsymbol{k}^4 = W^k \boldsymbol{a}^4$

自注意力机制 (Self-Attention)

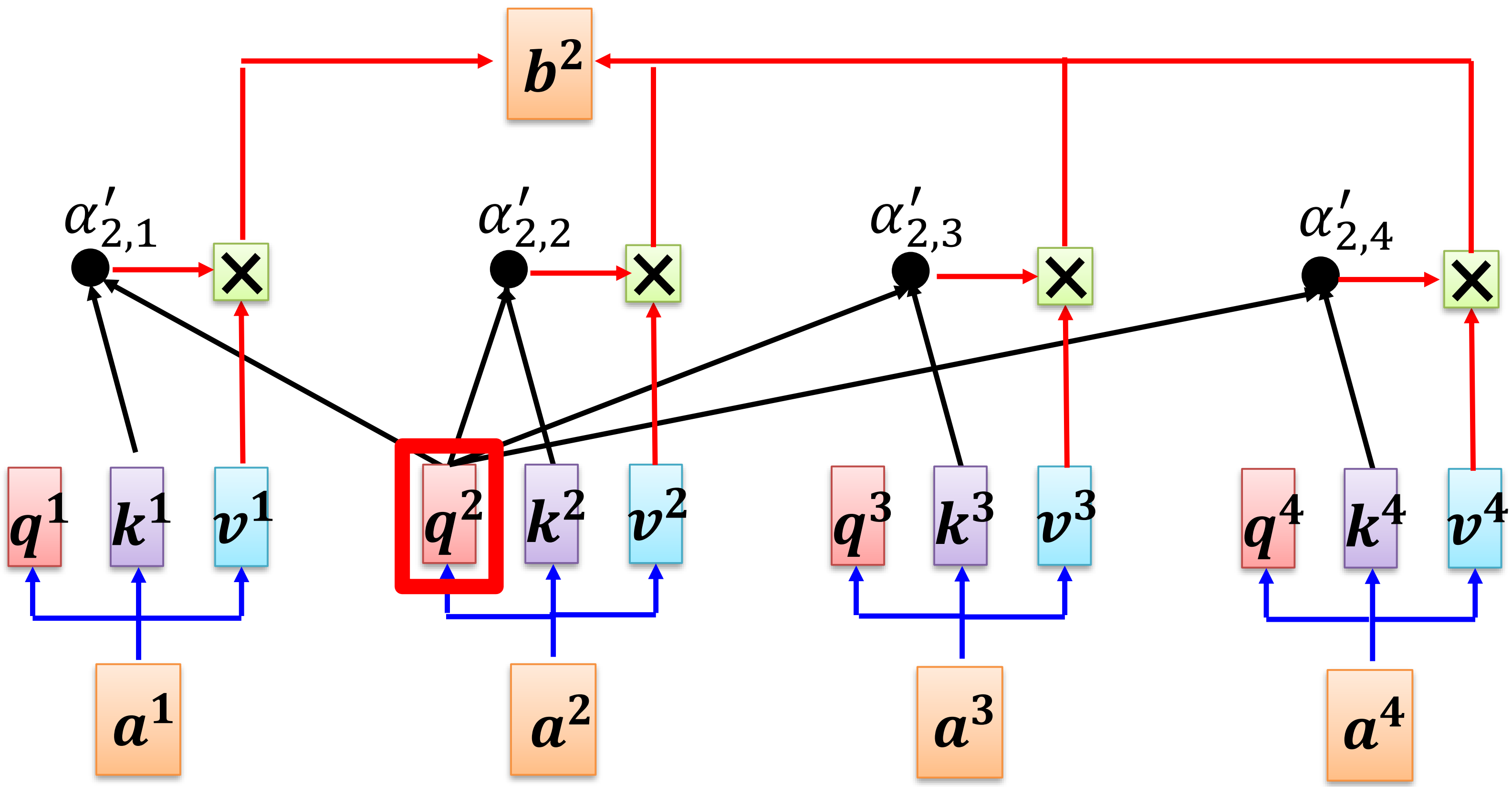
- 自注意力机制



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

自注意力机制 (Self-Attention)

- 自注意力机制



$\boldsymbol{k}^1 = W^k \boldsymbol{a}^1$

$\boldsymbol{q}^2 = W^q \boldsymbol{a}^2$

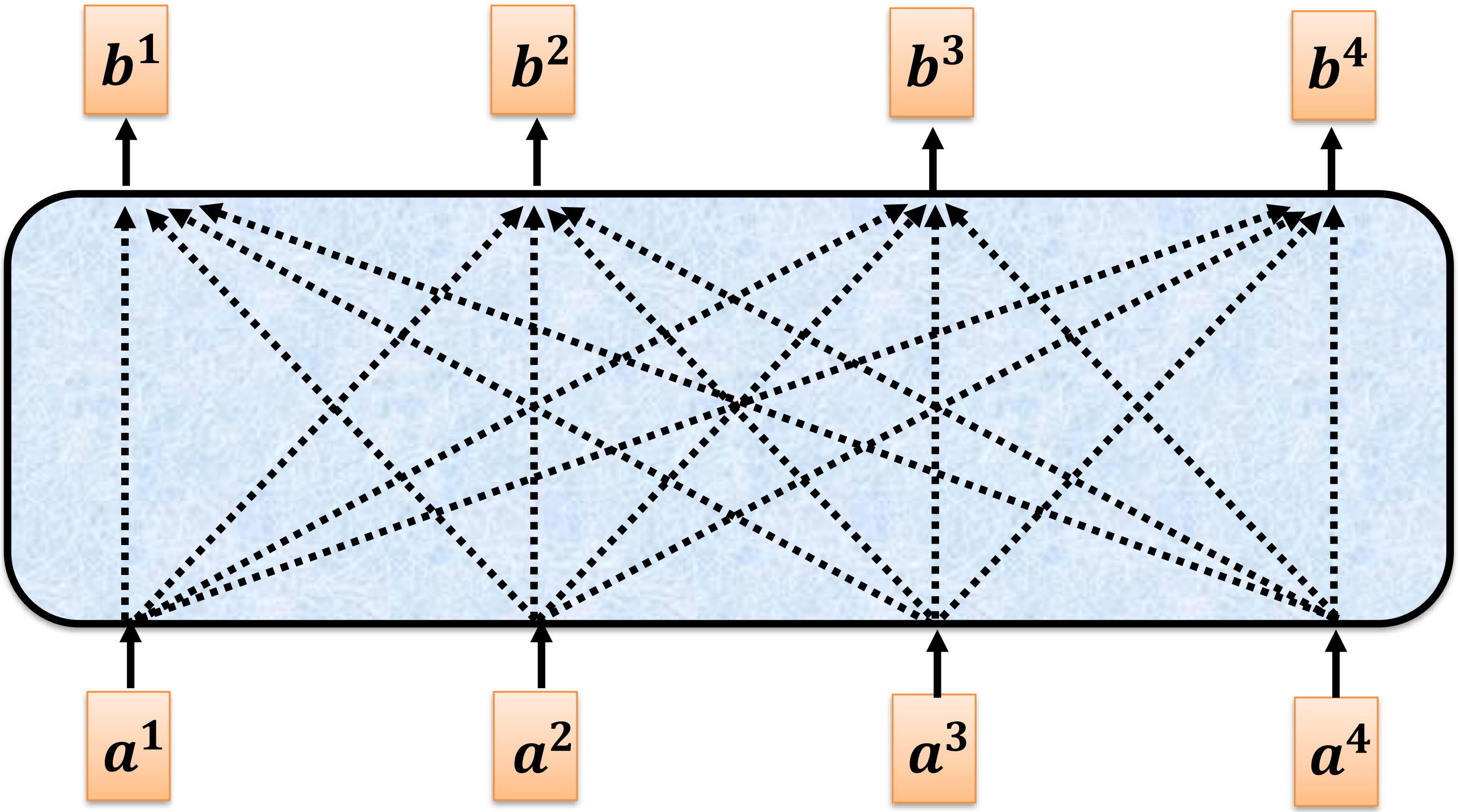
$\boldsymbol{k}^2 = W^k \boldsymbol{a}^2$

$\boldsymbol{k}^3 = W^k \boldsymbol{a}^3$

$\boldsymbol{k}^4 = W^k \boldsymbol{a}^4$

自注意力机制 (Self-Attention)

- 自注意力机制



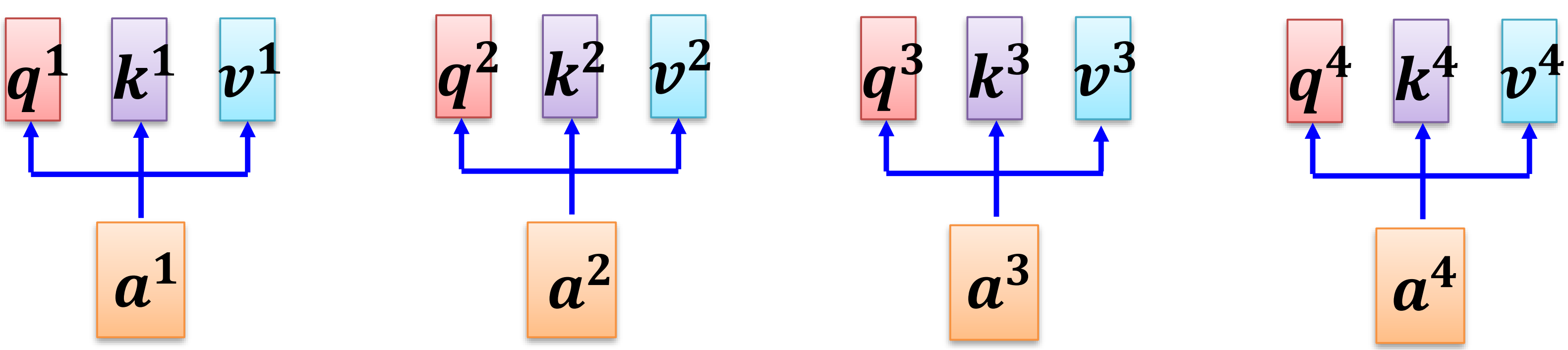
自注意力机制 (Self-Attention)

- 自注意力机制

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

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

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



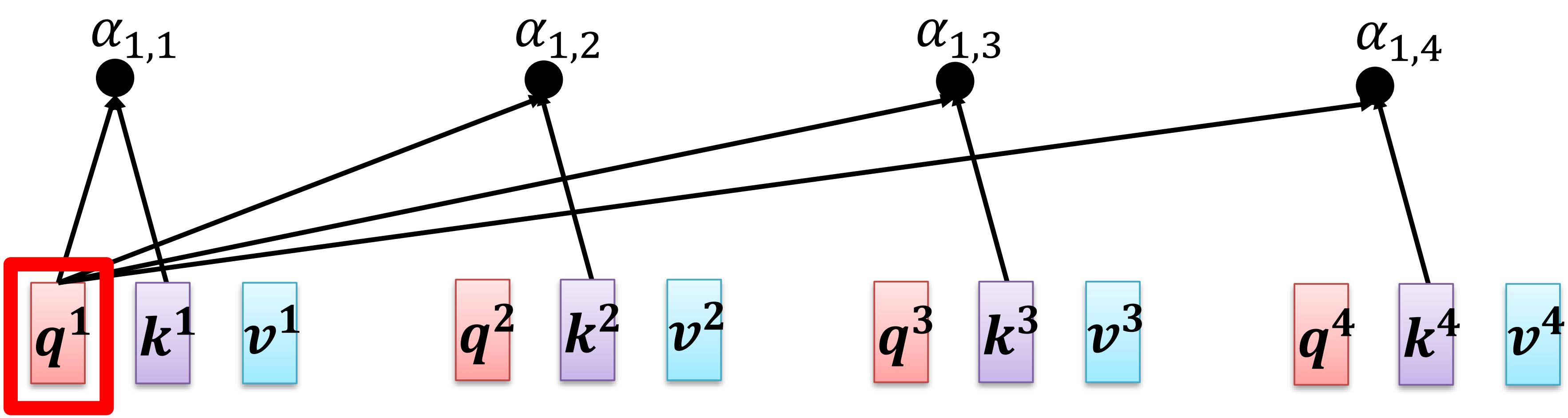
自注意力机制 (Self-Attention)

- 自注意力机制

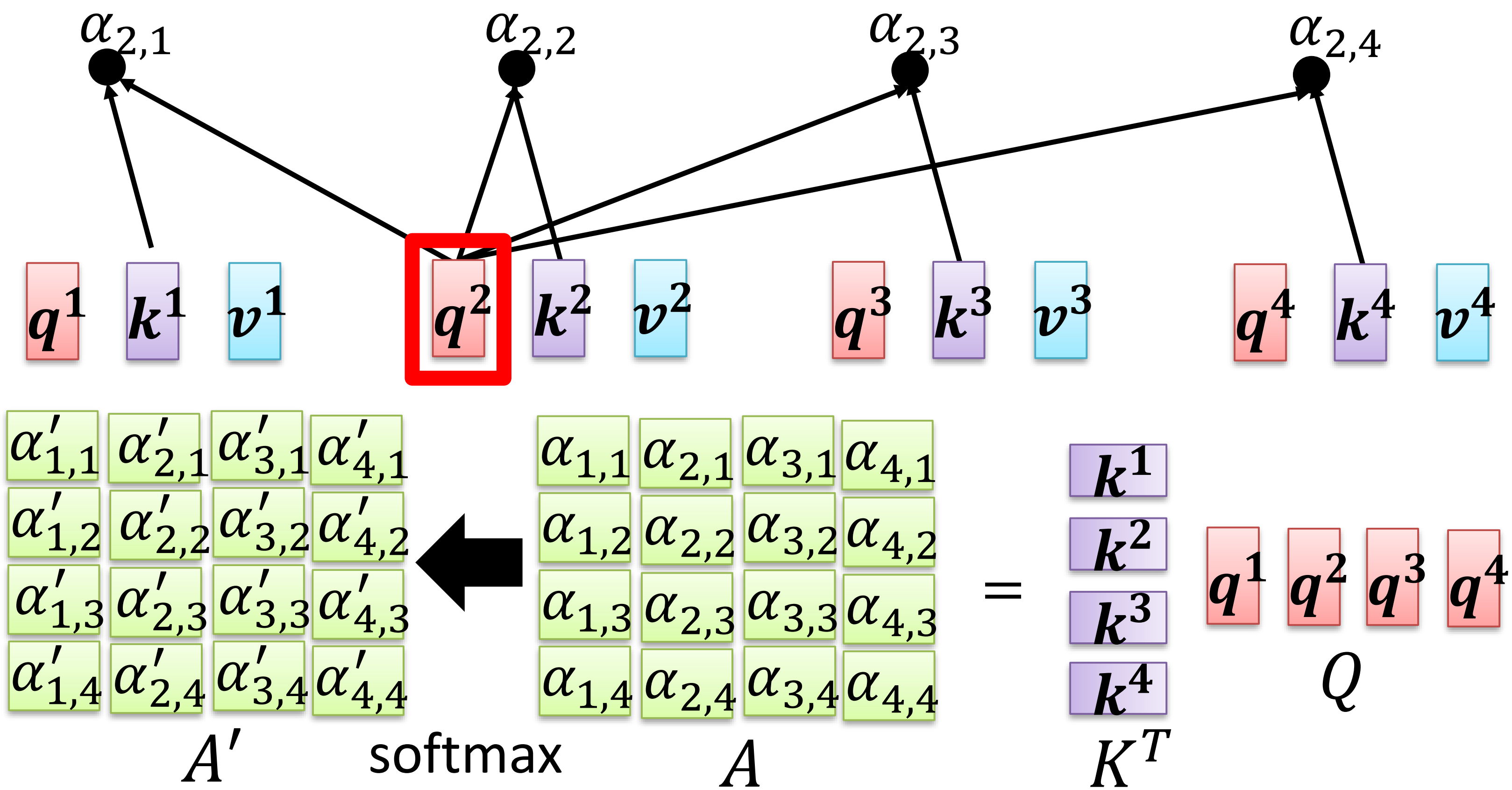
$$\alpha_{1,1} = k^1 \; q^1 \qquad \alpha_{1,2} = k^2 \; q^1$$

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

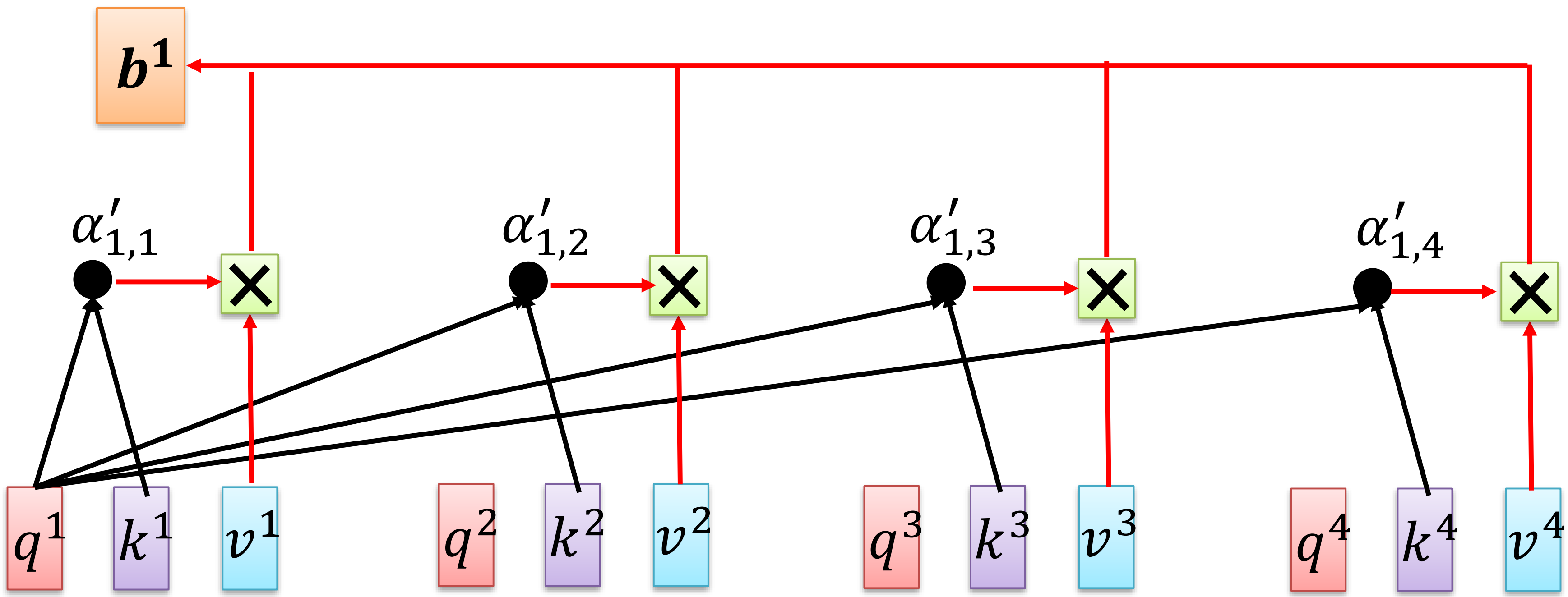
$$\begin{bmatrix}\alpha_{1,1}\\ \alpha_{1,2}\\ \alpha_{1,3}\\ \alpha_{1,4}\end{bmatrix} = \begin{bmatrix}k^1\\ k^2\\ k^3\\ k^4\end{bmatrix} q^1$$



自注意力机制 (Self-Attention)

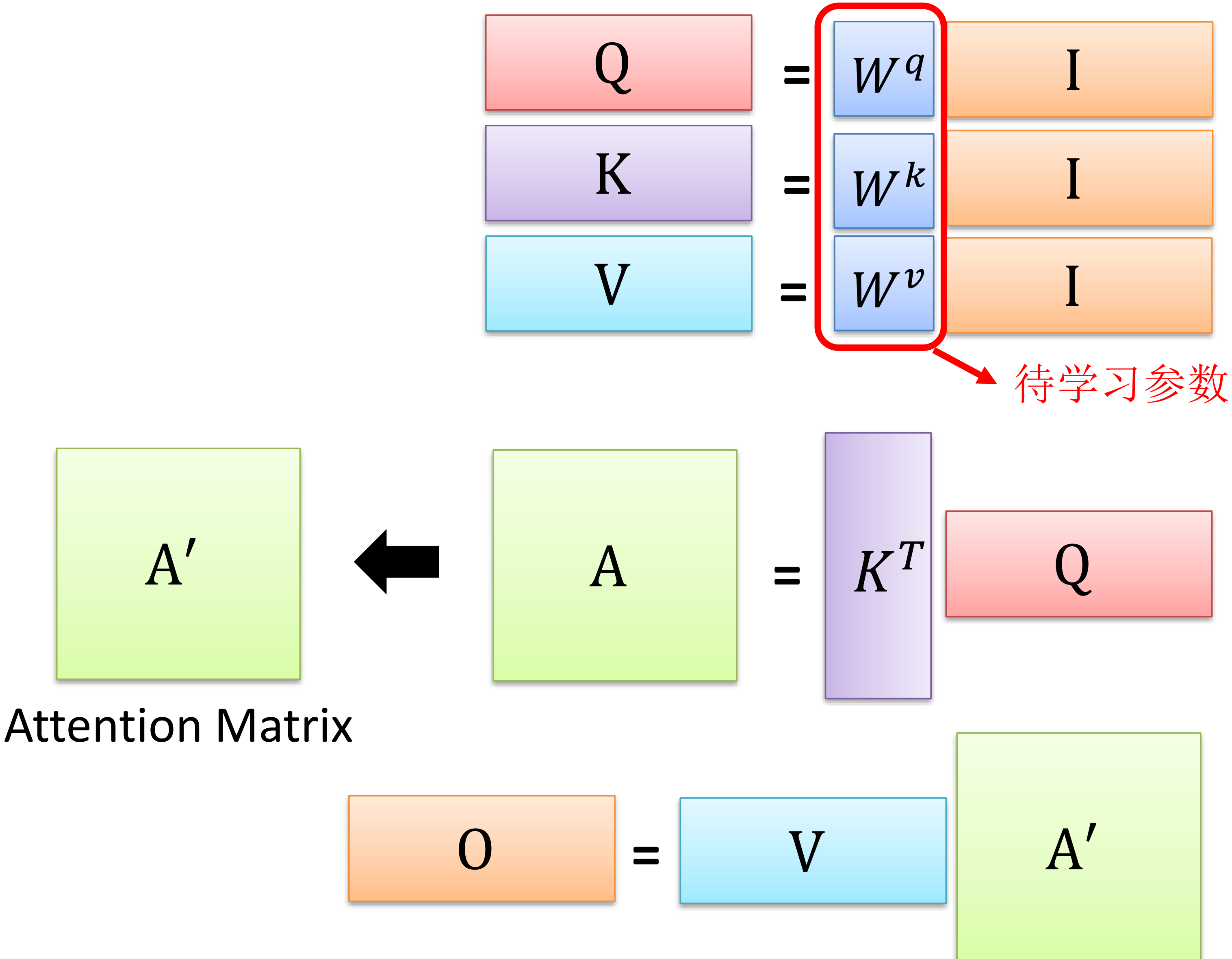


自注意力机制 (Self-Attention)



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

自注意力机制 (Self-Attention)



自注意力机制 (Self-Attention)

- 课堂练习

$$x^1 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad x^2 = \begin{bmatrix} 0 \\ 2 \\ 0 \\ 2 \end{bmatrix} \quad x^3 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

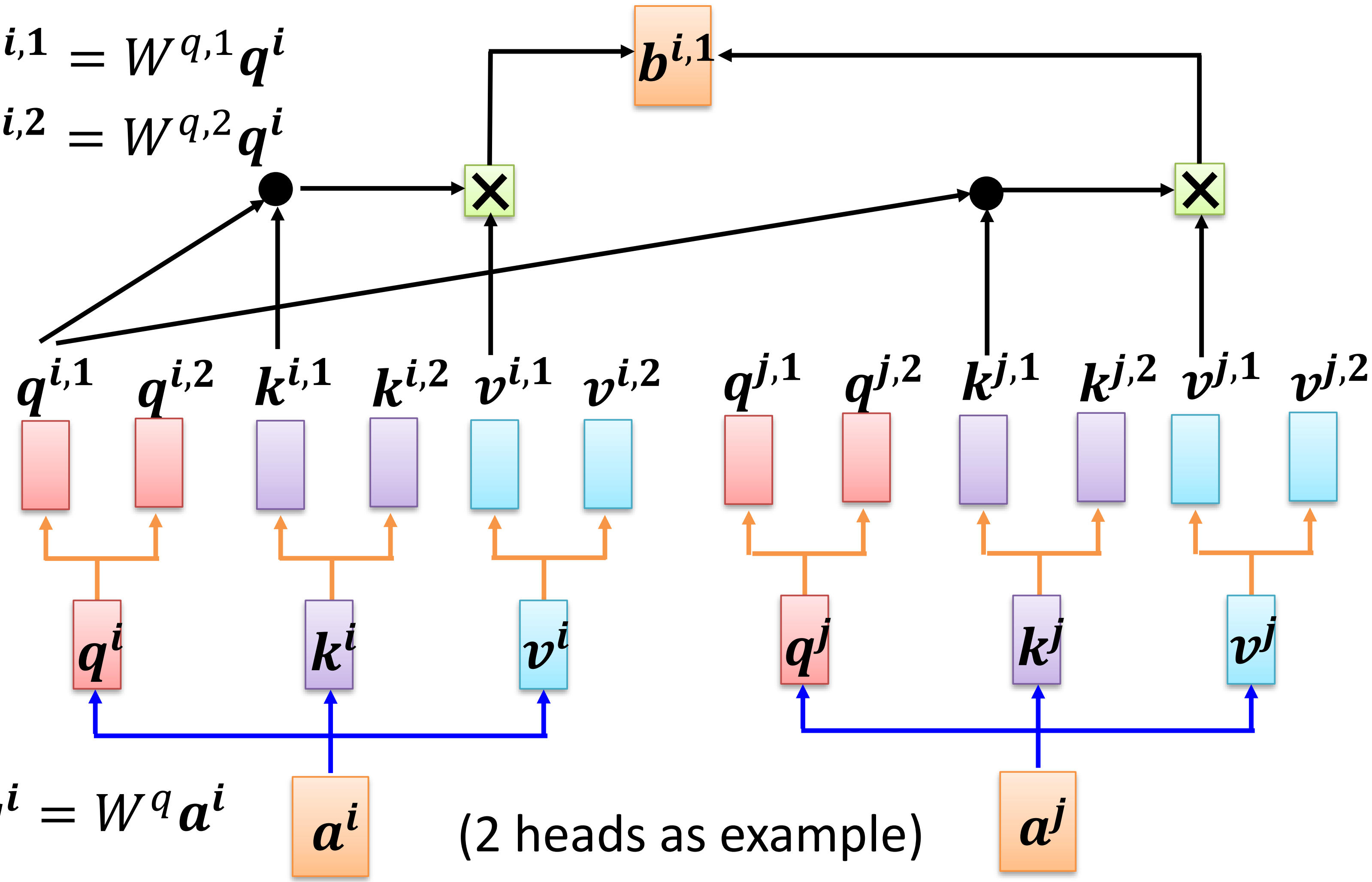
$$\mathrm{I} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix}$$

$$W^q = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \quad W^k = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad W^v = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 2 & 3 & 0 & 1 \\ 0 & 0 & 3 & 0 \end{bmatrix}$$

自注意力机制 (Self-Attention)

- 多头自注意力机制 (Multi-Head Self-Attention)

$$q^{i,1} = W^{q,1} q^i$$
$$q^{i,2} = W^{q,2} q^i$$



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

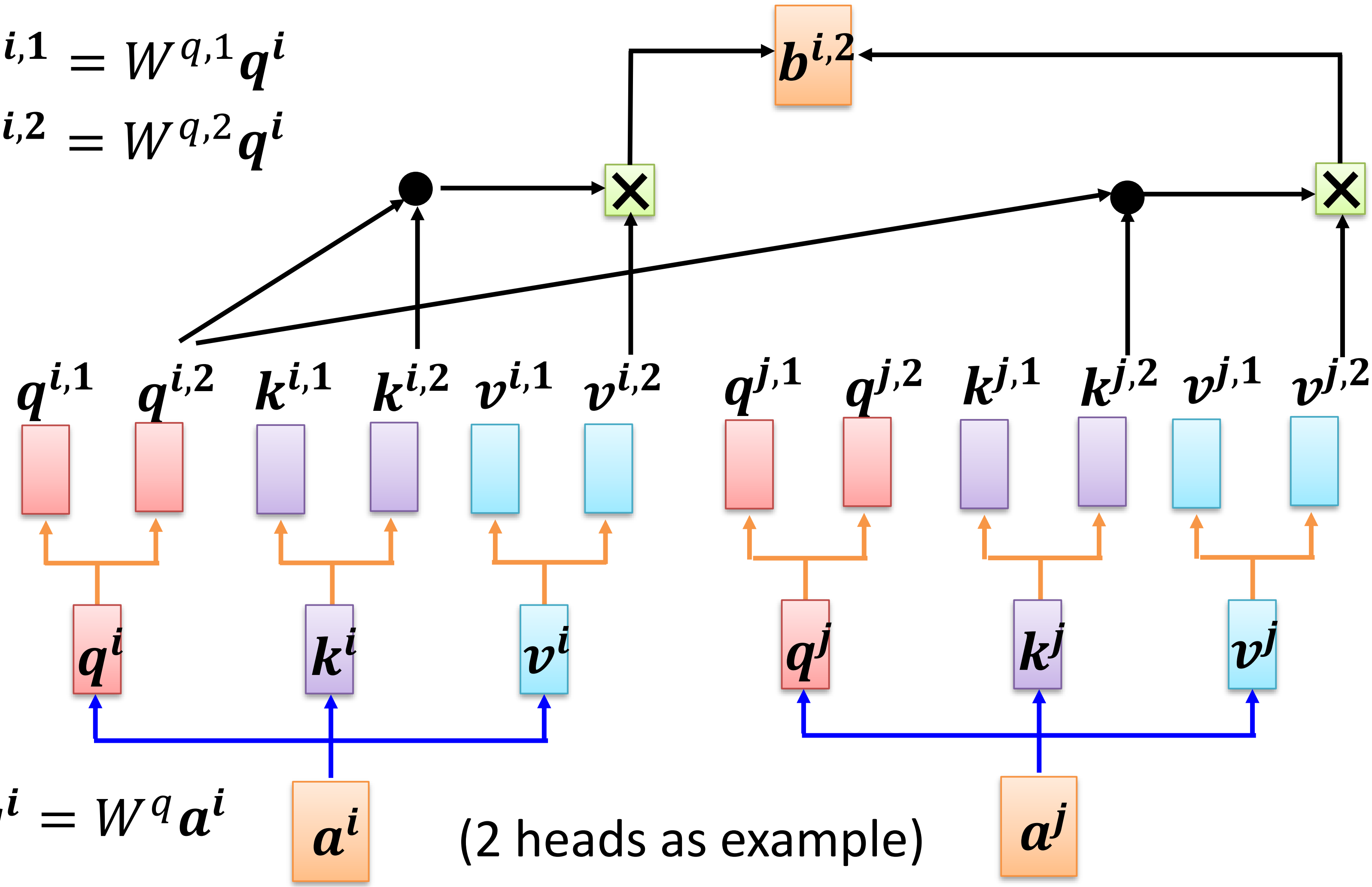
(2 heads as example)

自注意力机制 (Self-Attention)

- 多头自注意力机制 (Multi-Head Self-Attention)

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

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

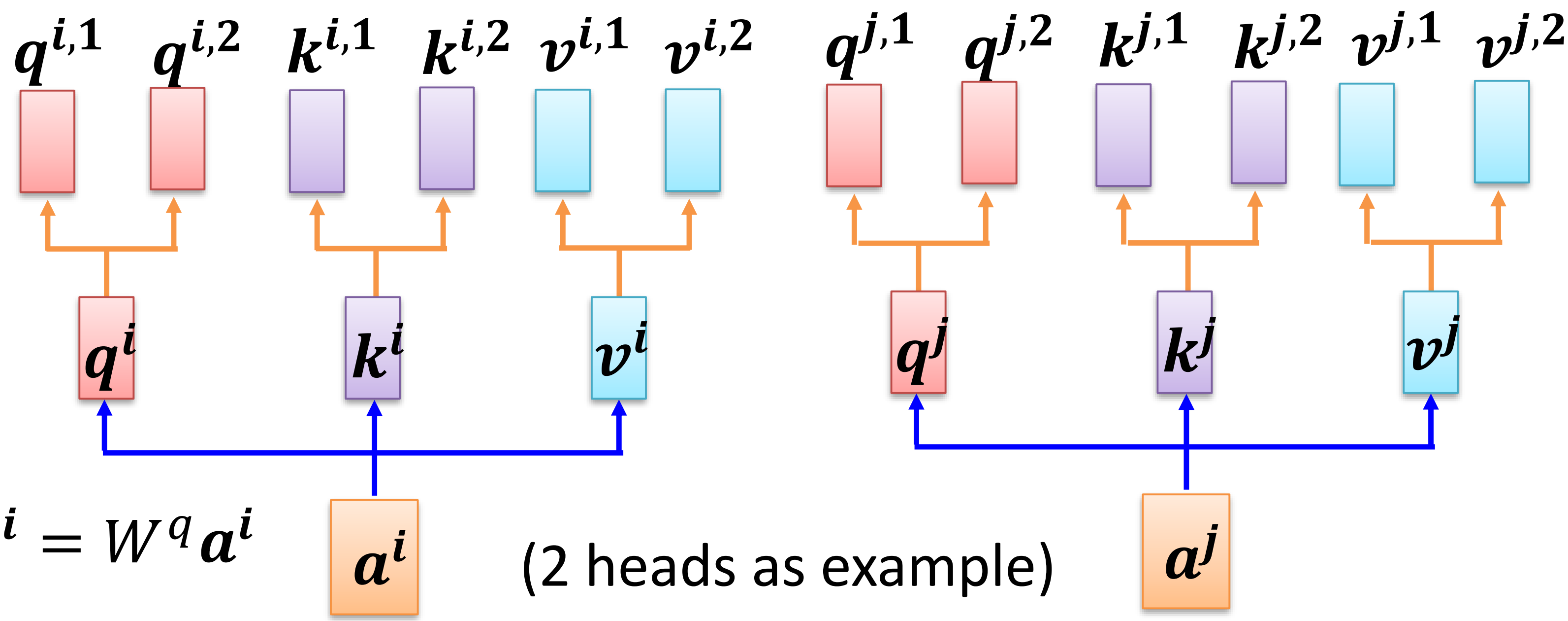


$$q^i = W^q a^i$$

自注意力机制 (Self-Attention)

- 多头自注意力机制 (Multi-Head Self-Attention)

$$\boldsymbol{b}^i = W^O \begin{bmatrix} \boldsymbol{b}^{i,1} \\ \boldsymbol{b}^{i,2} \end{bmatrix}$$

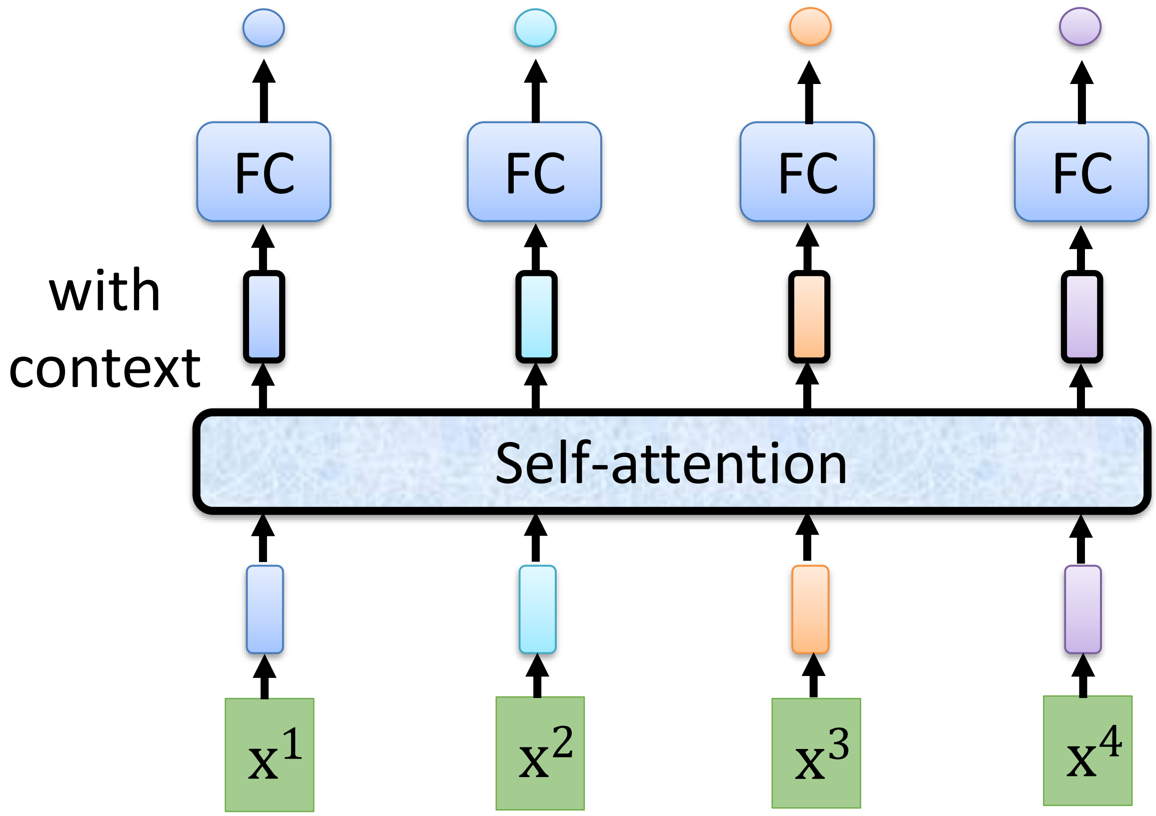


$\boldsymbol{q}^i = W^q \boldsymbol{a}^i$

(2 heads as example)

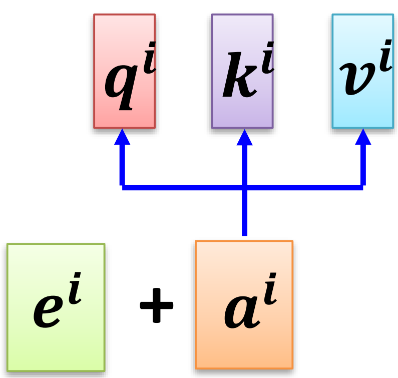
自注意力机制 (Self-Attention)

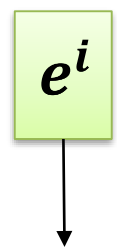
- 自注意力机制缺陷
  - 忽略了序列中的位置信息



自注意力机制 (Self-Attention)

- 位置编码 (Positional Encoding)
  - 为每个位置引入一个位置编码 e^i
    - **人工构造**
    - **参数学习**

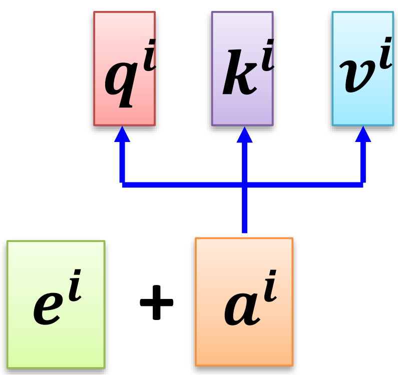


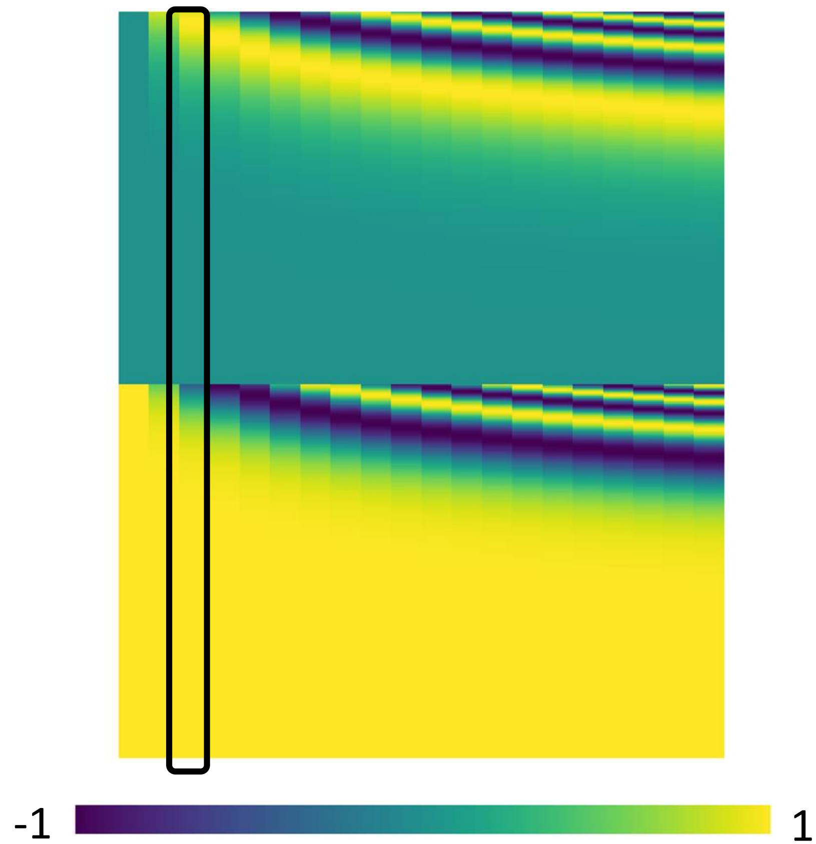


$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

自注意力机制 (Self-Attention)

- 位置编码 (Positional Encoding)
  - 为每个位置引入一个位置编码 e^i
    - **人工构造**
    - **参数学习**

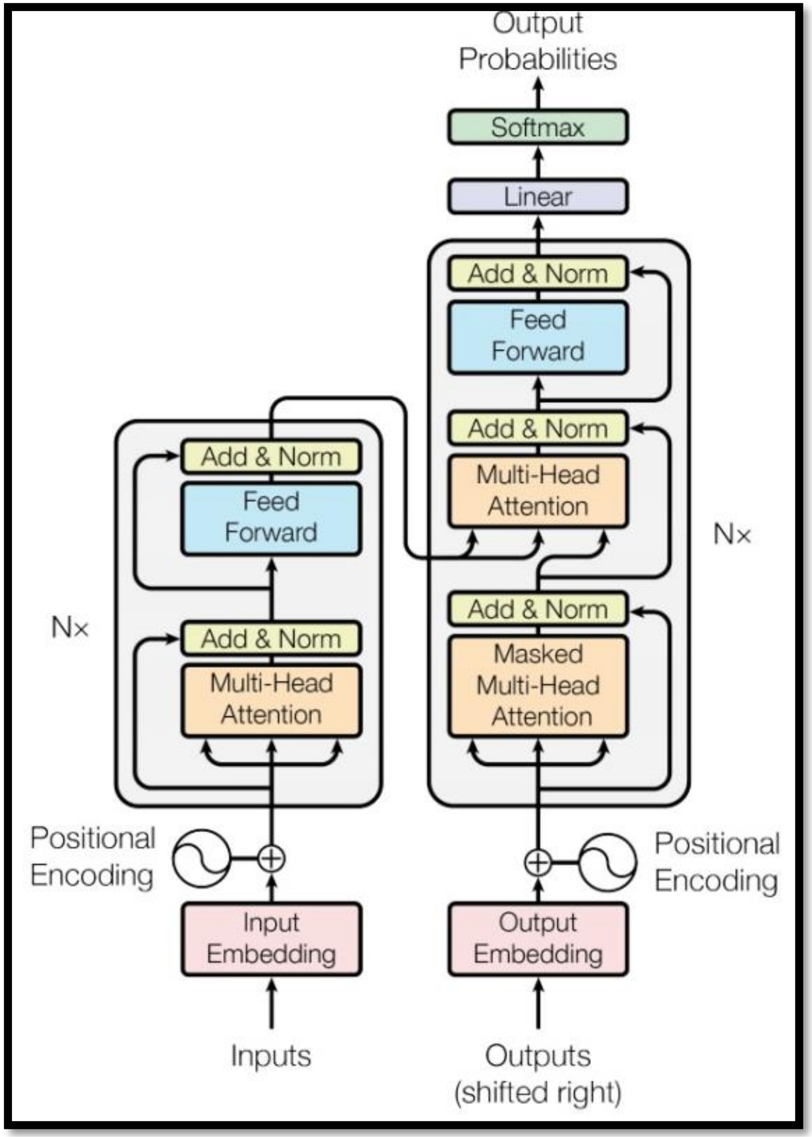




本章内容

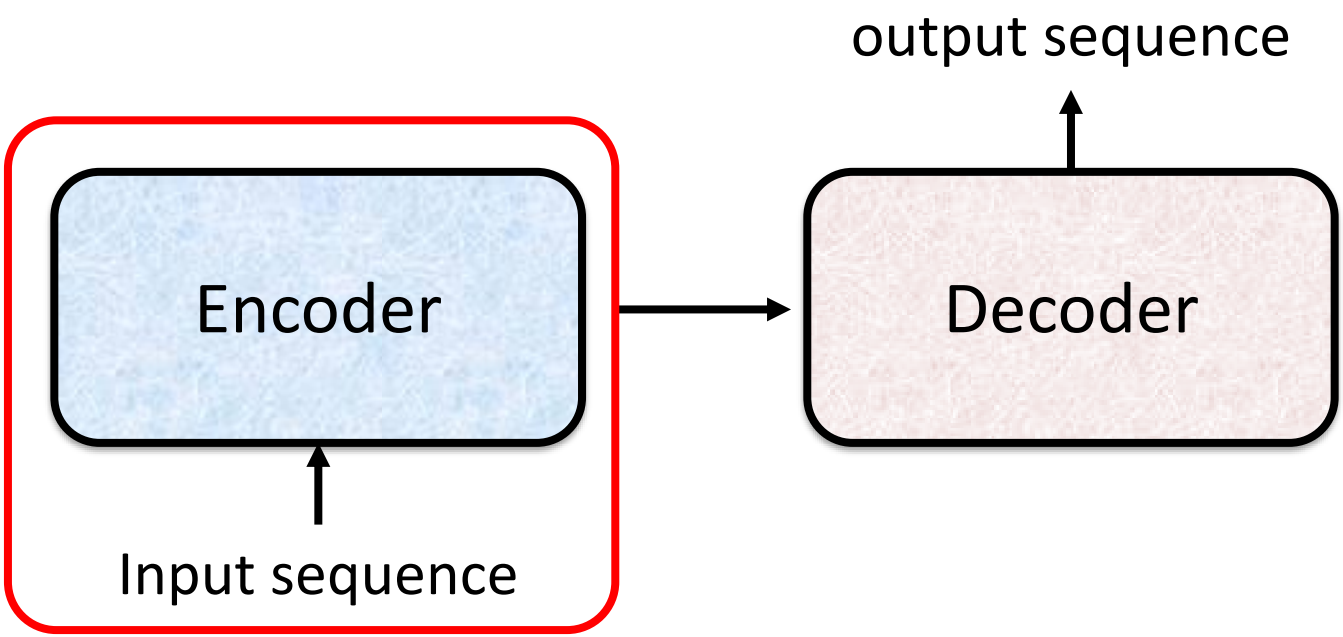
Transformer

- 整体架构
  - Encoder-Decoder



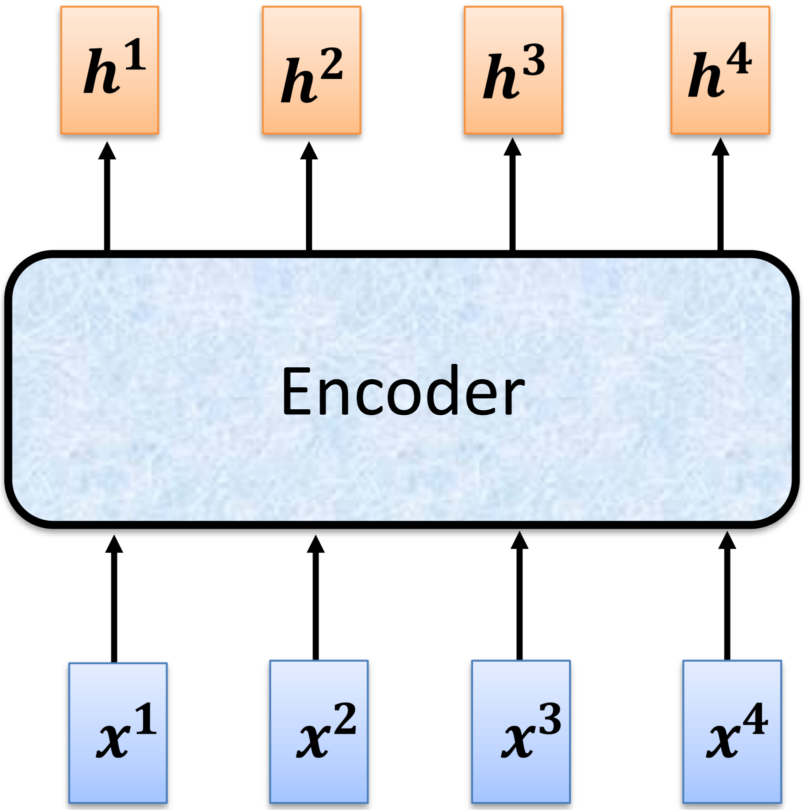
Transformer

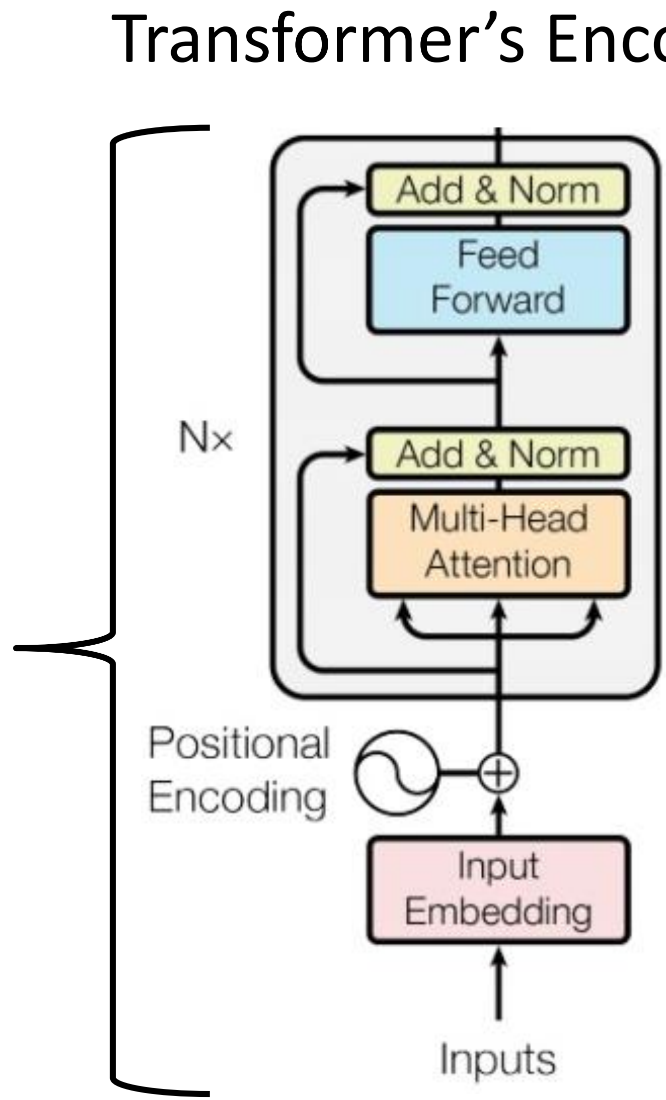
- 整体架构
  - Encoder-Decoder



Transformer

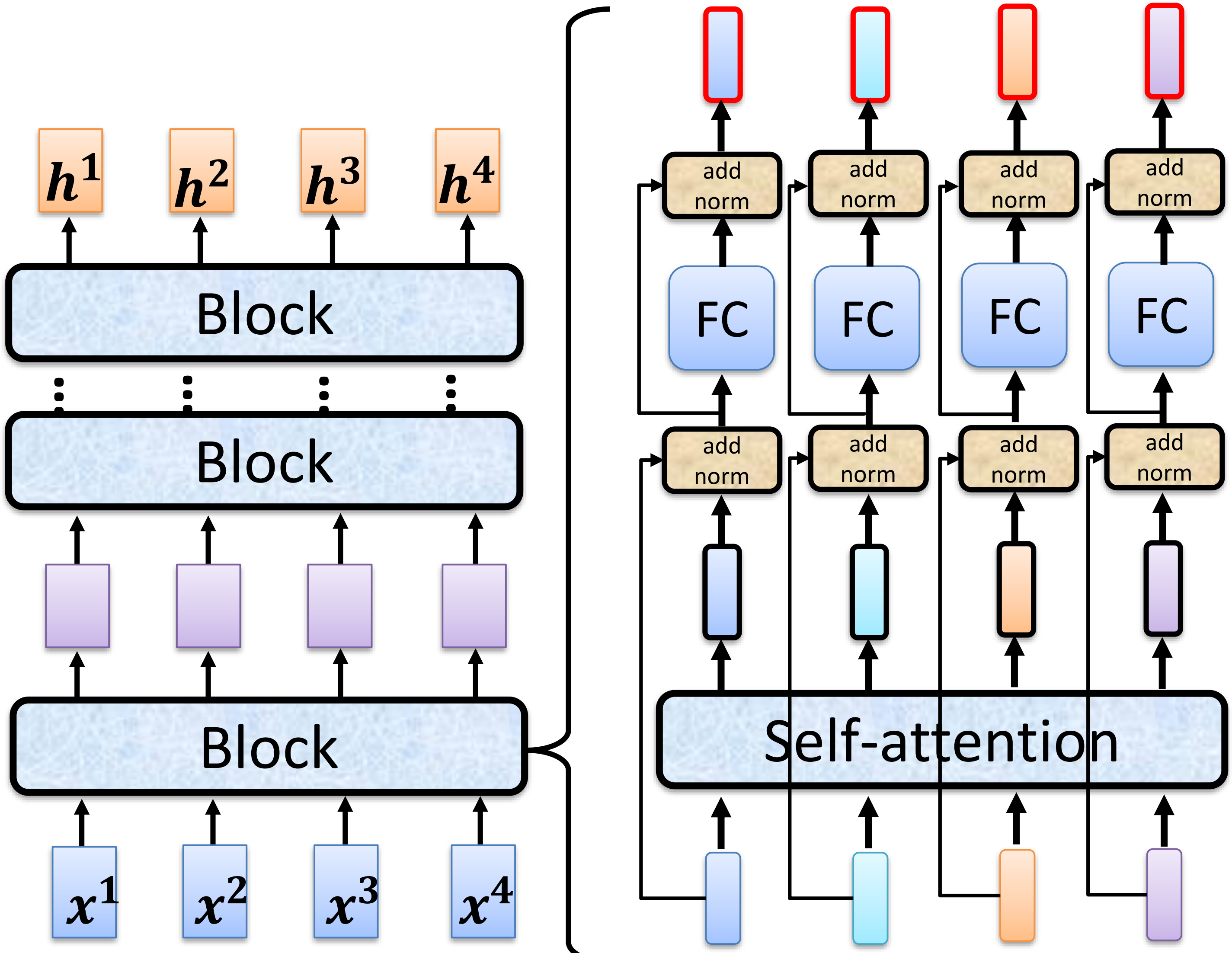
- 整体架构
  - Encoder





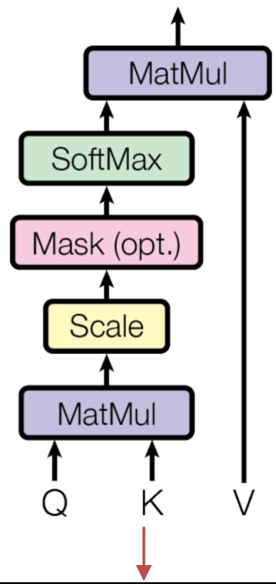
Transformer

- Encoder

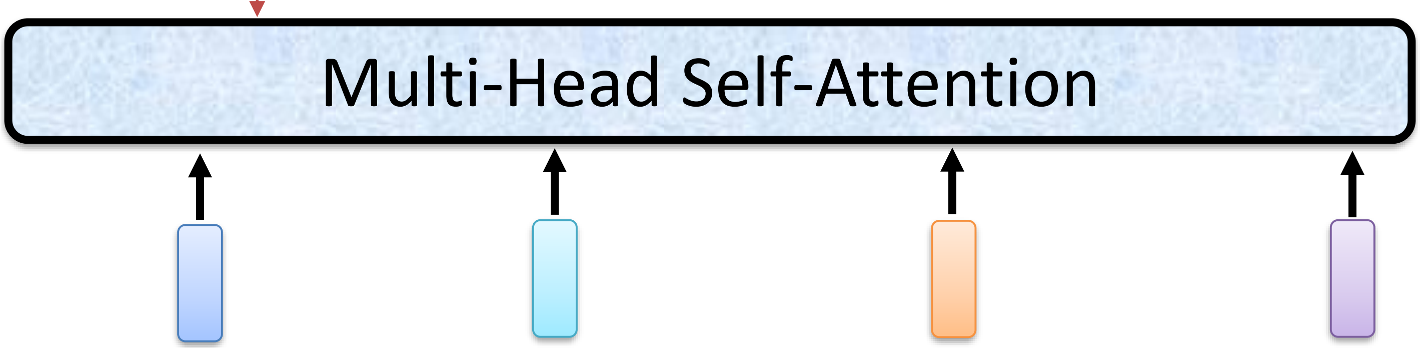


Transformer

- Encoder
  - Scaled Dot-Product Attention

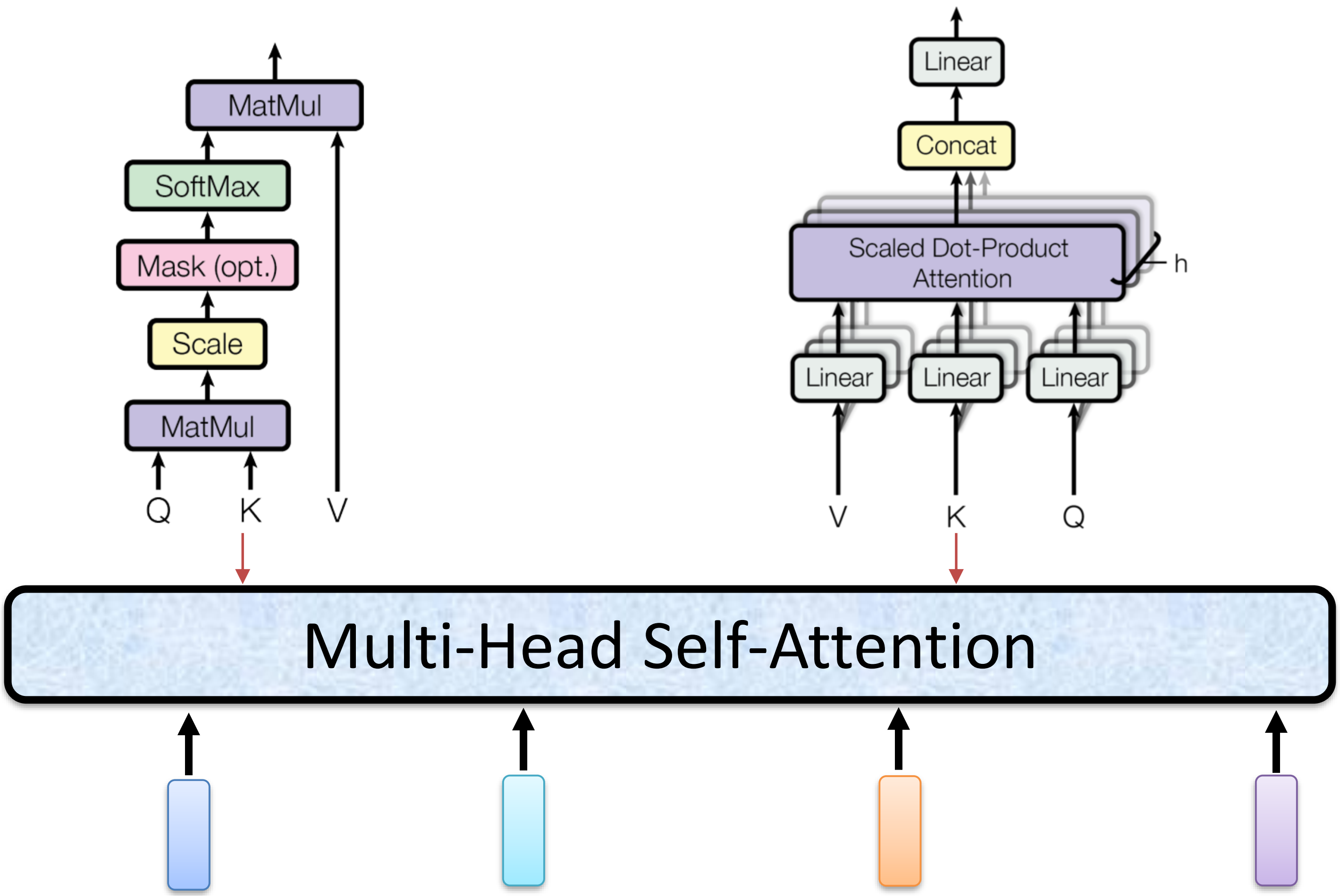


$$Attention(Q,K,V) = SoftMax(\frac{QK^T}{\sqrt{d}})V$$



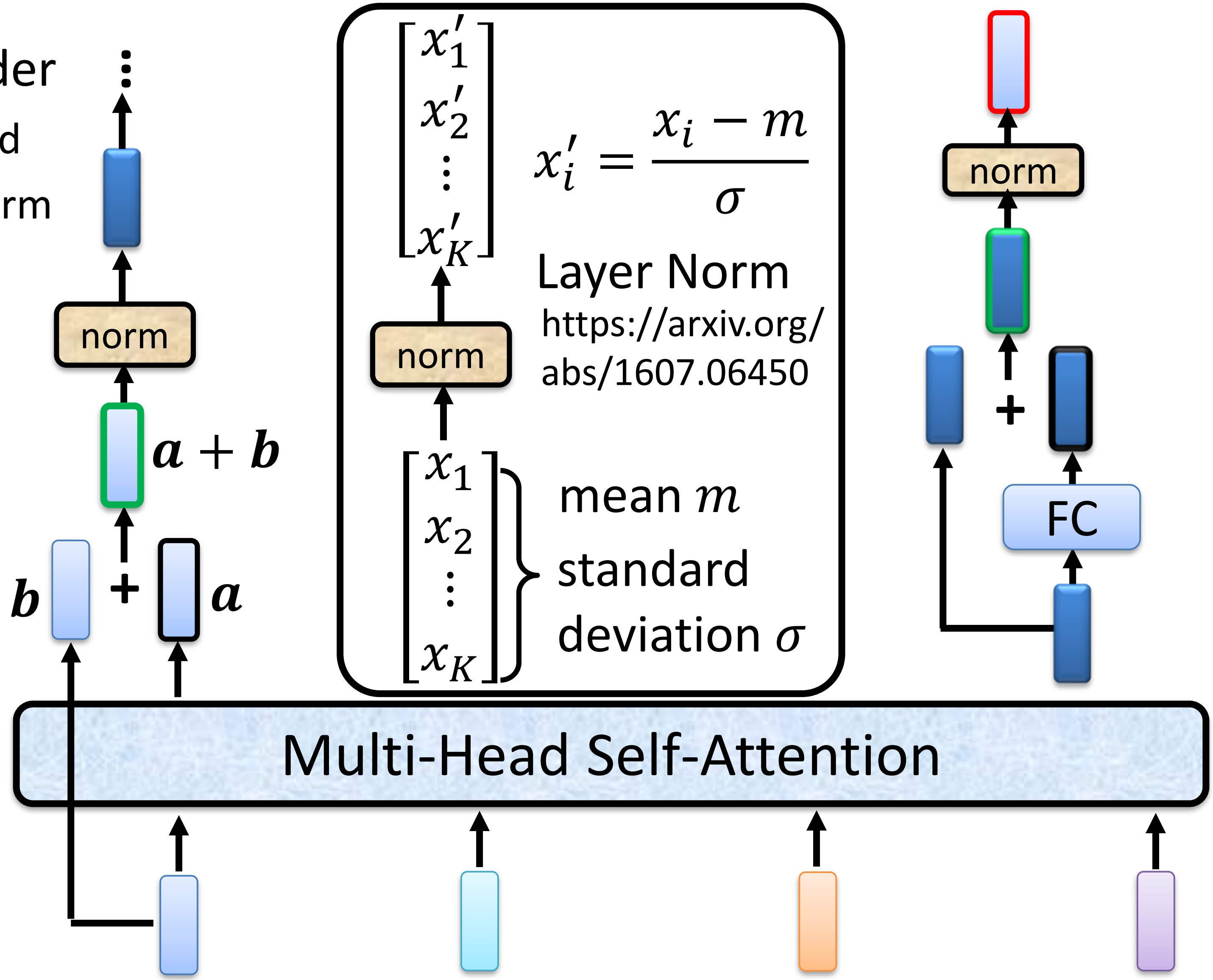
Transformer

- Encoder
  - Scaled Dot-Product Attention Multi-Head Attention



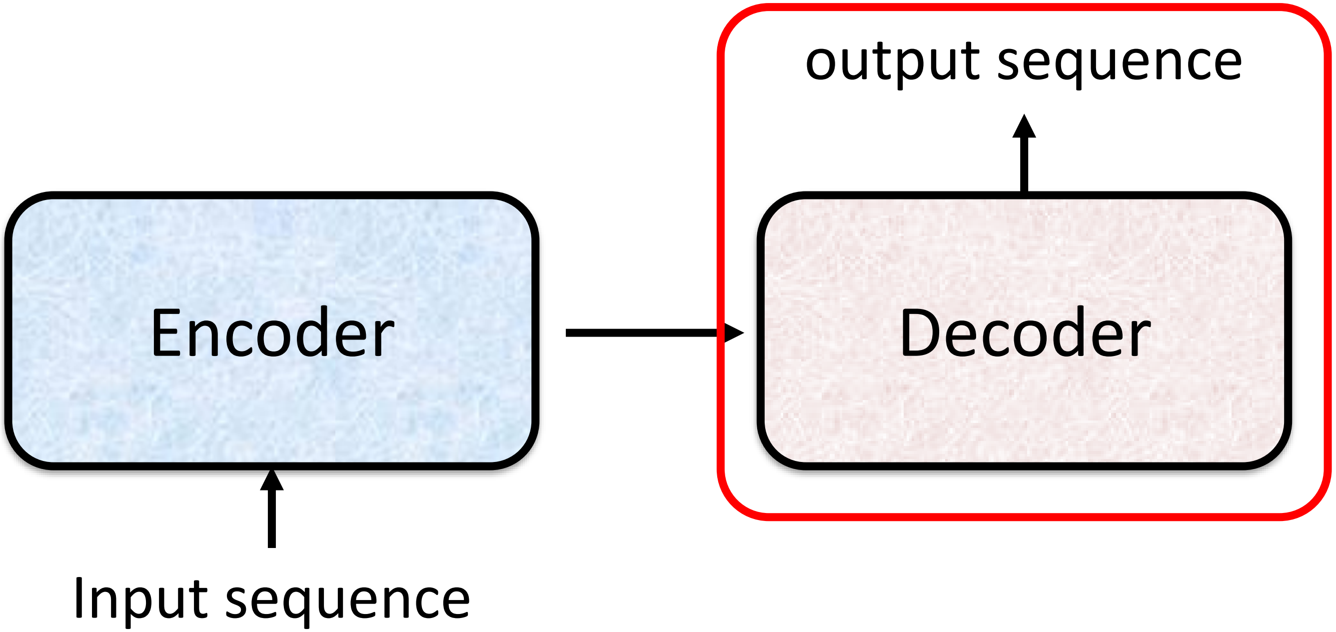
Transformer

- Encoder
  - Add
  - Norm



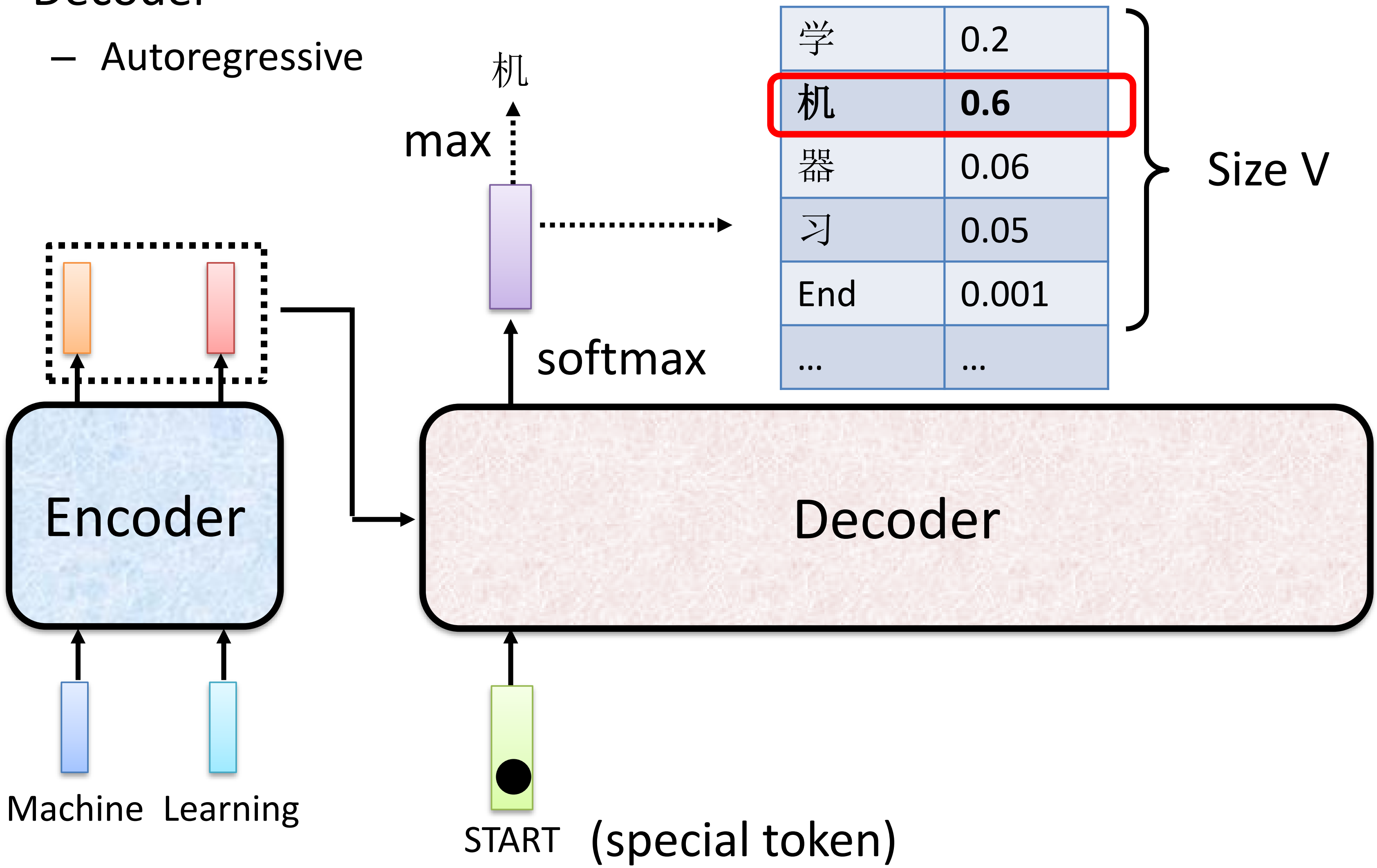
Transformer

- Decoder
  - Autoregressive



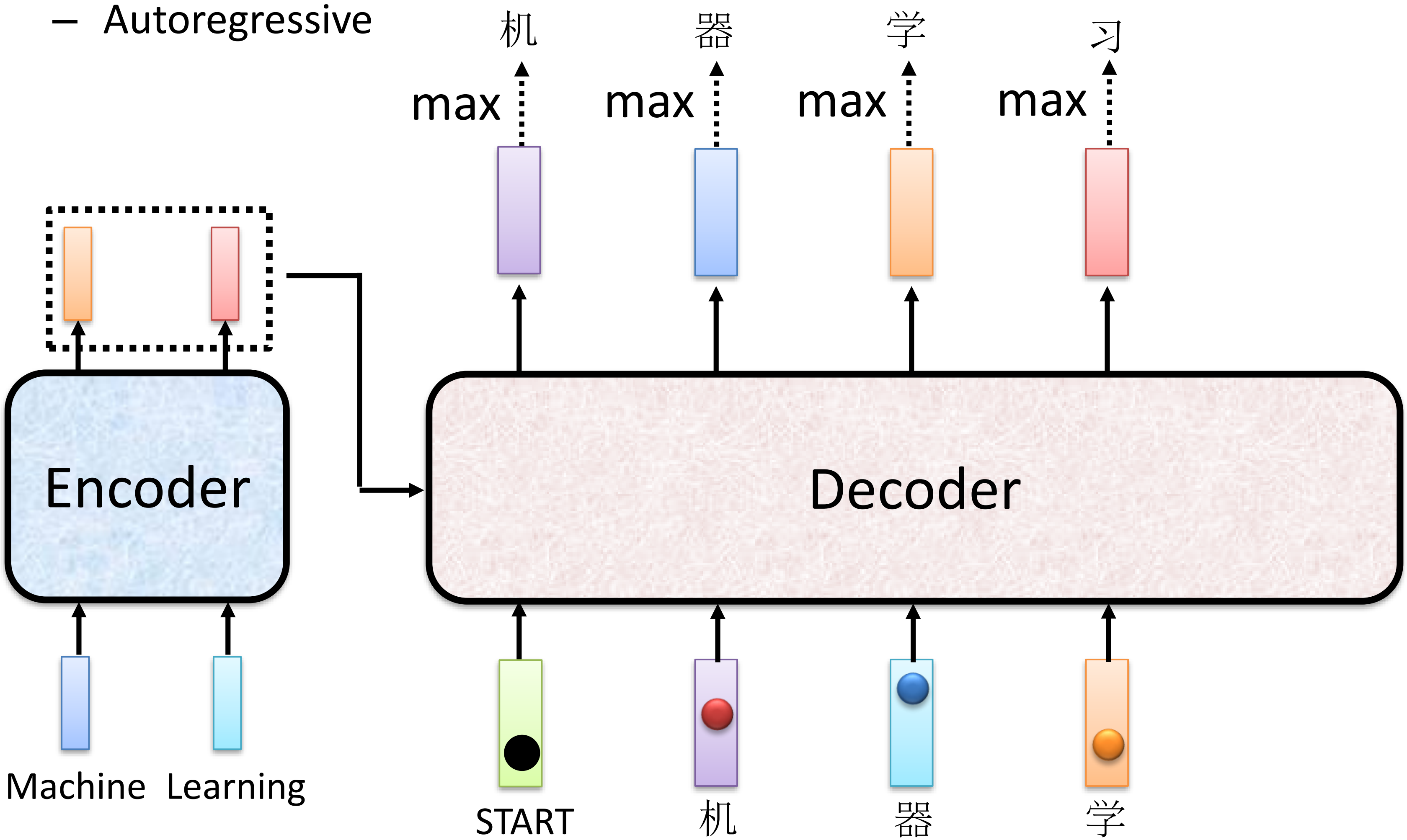
Transformer

- Decoder
  - Autoregressive



Transformer

- Decoder
  - Autoregressive

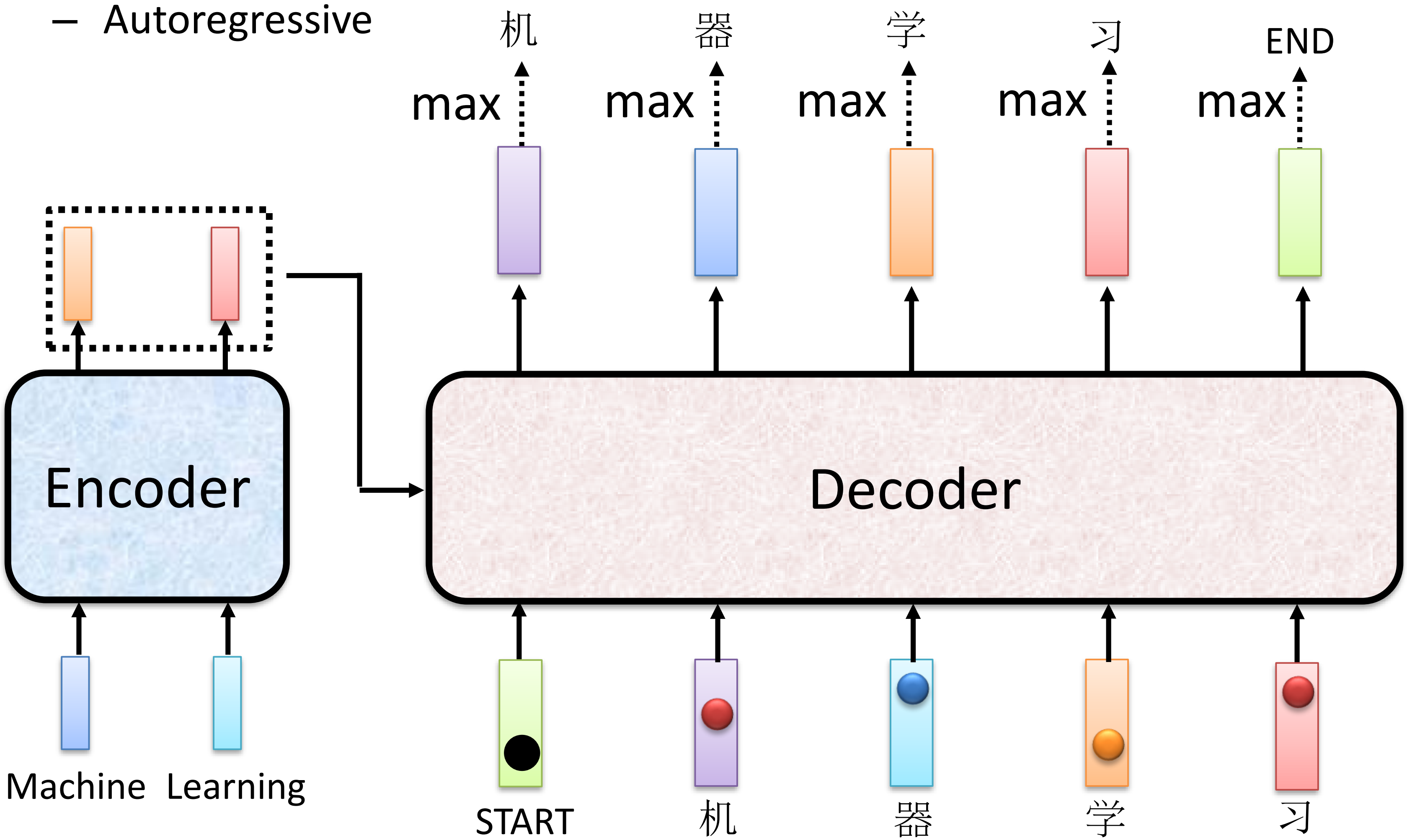


Transformer

学	0.001
机	**0.01**
器	0.001
习	0.001
End	0.9
...	...

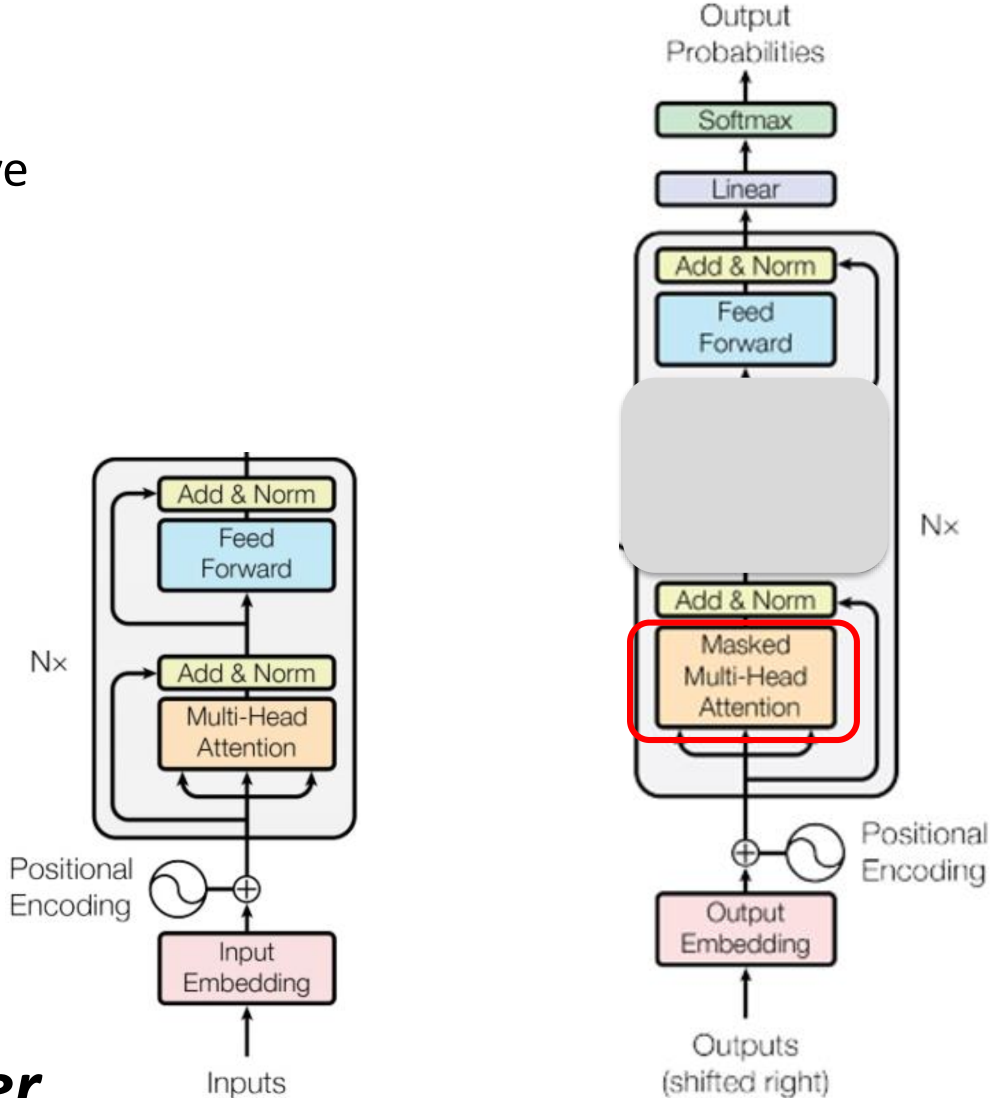
Size V

- Decoder
  - Autoregressive



Transformer

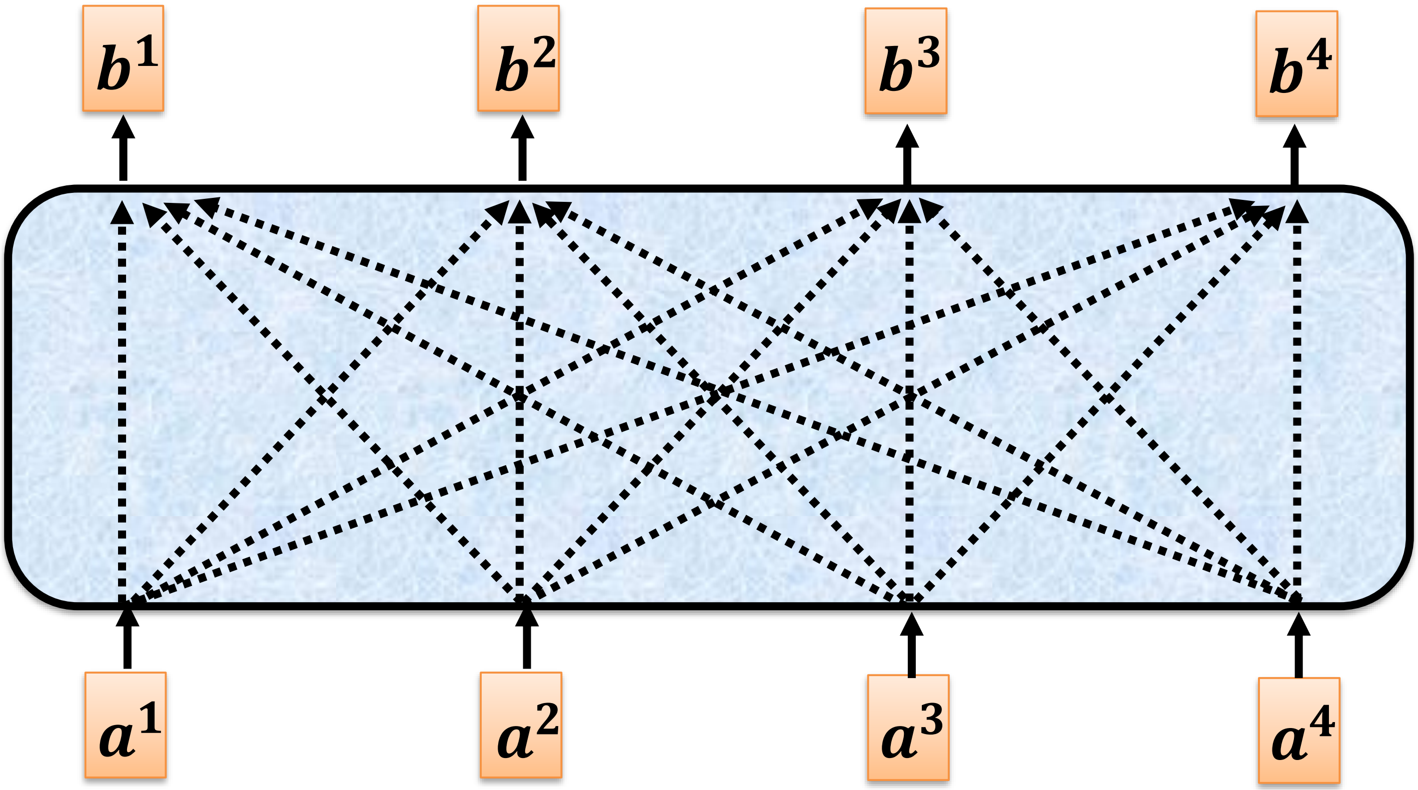
- Decoder
  - Autoregressive



Encoder ***Decoder***

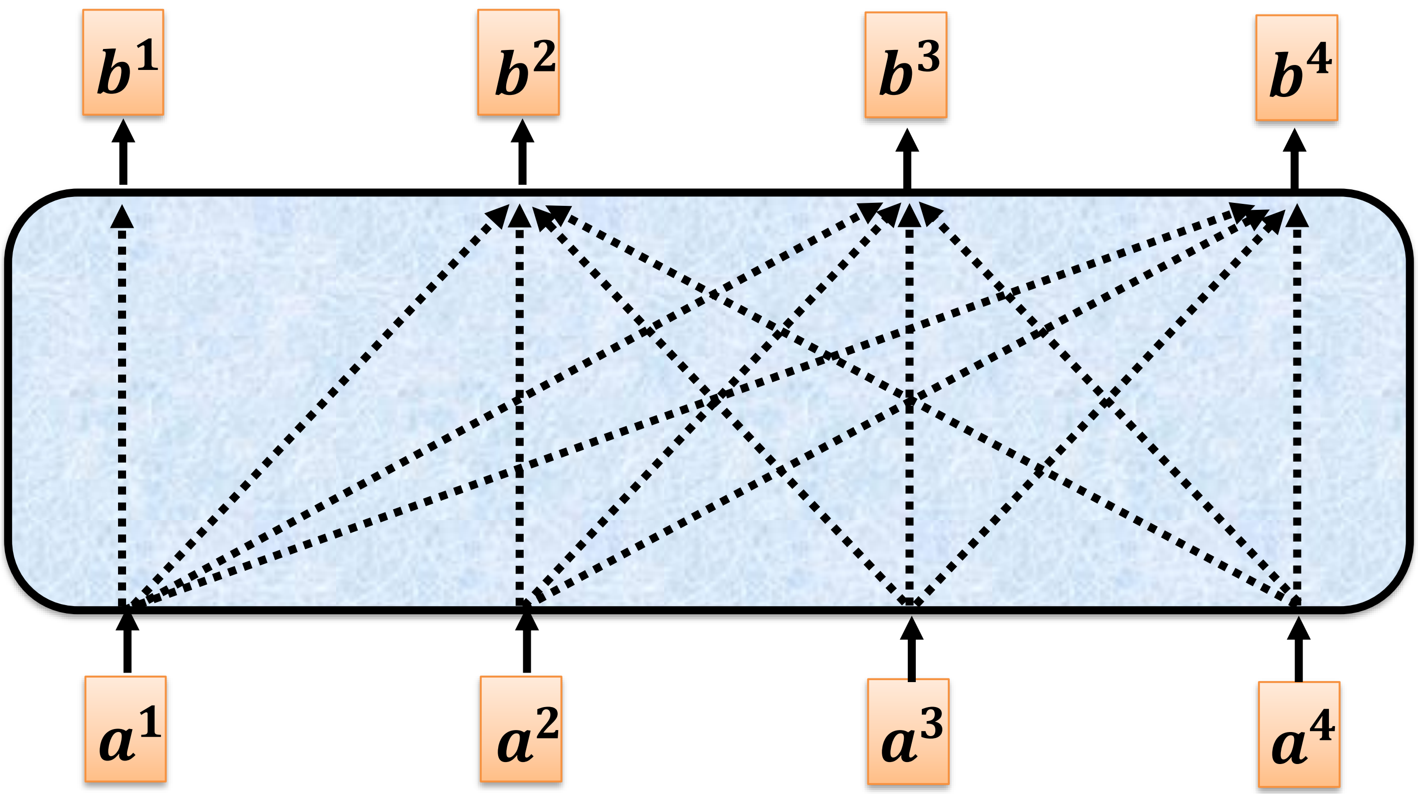
Transformer

- Decoder
  - Self-Attention



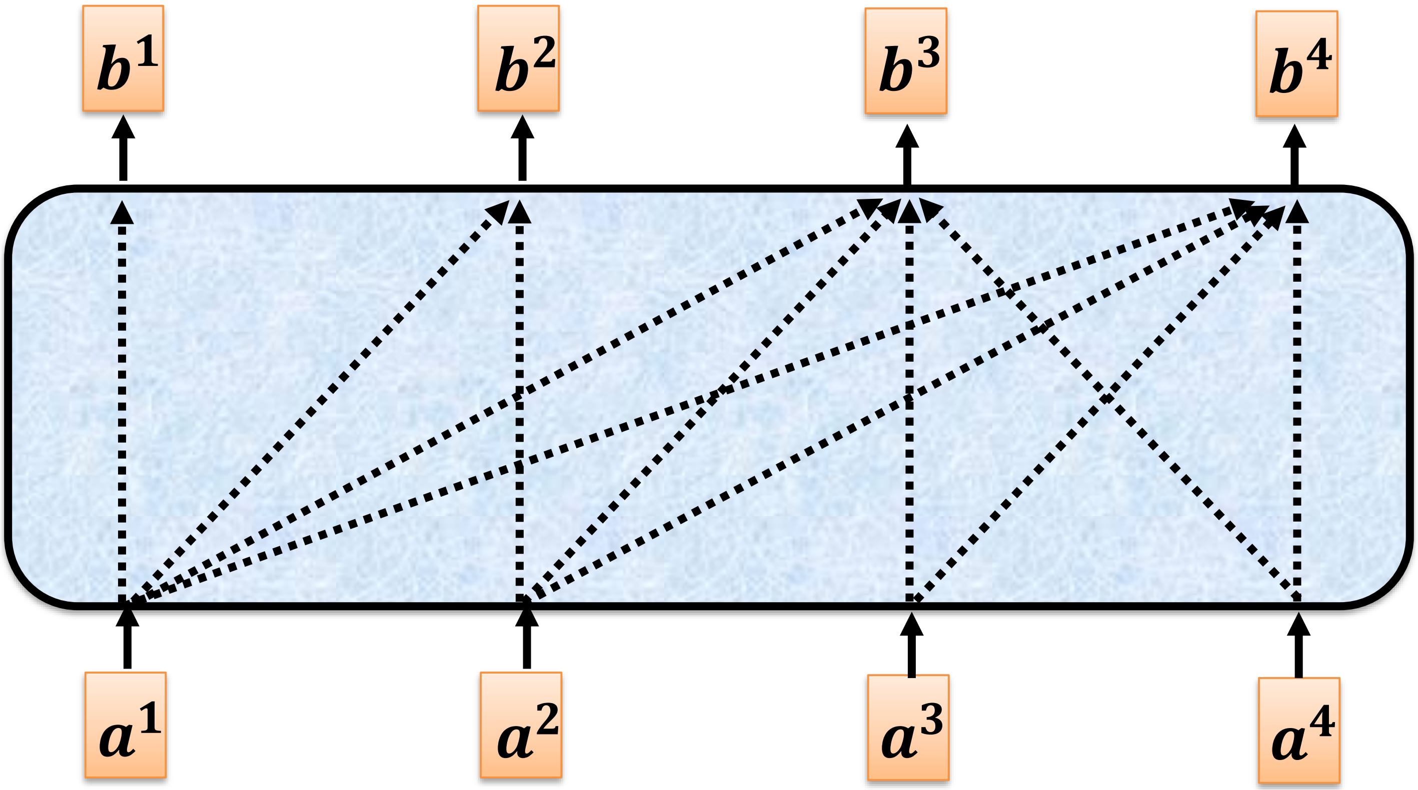
Transformer

- Decoder
  - Masked Self-Attention



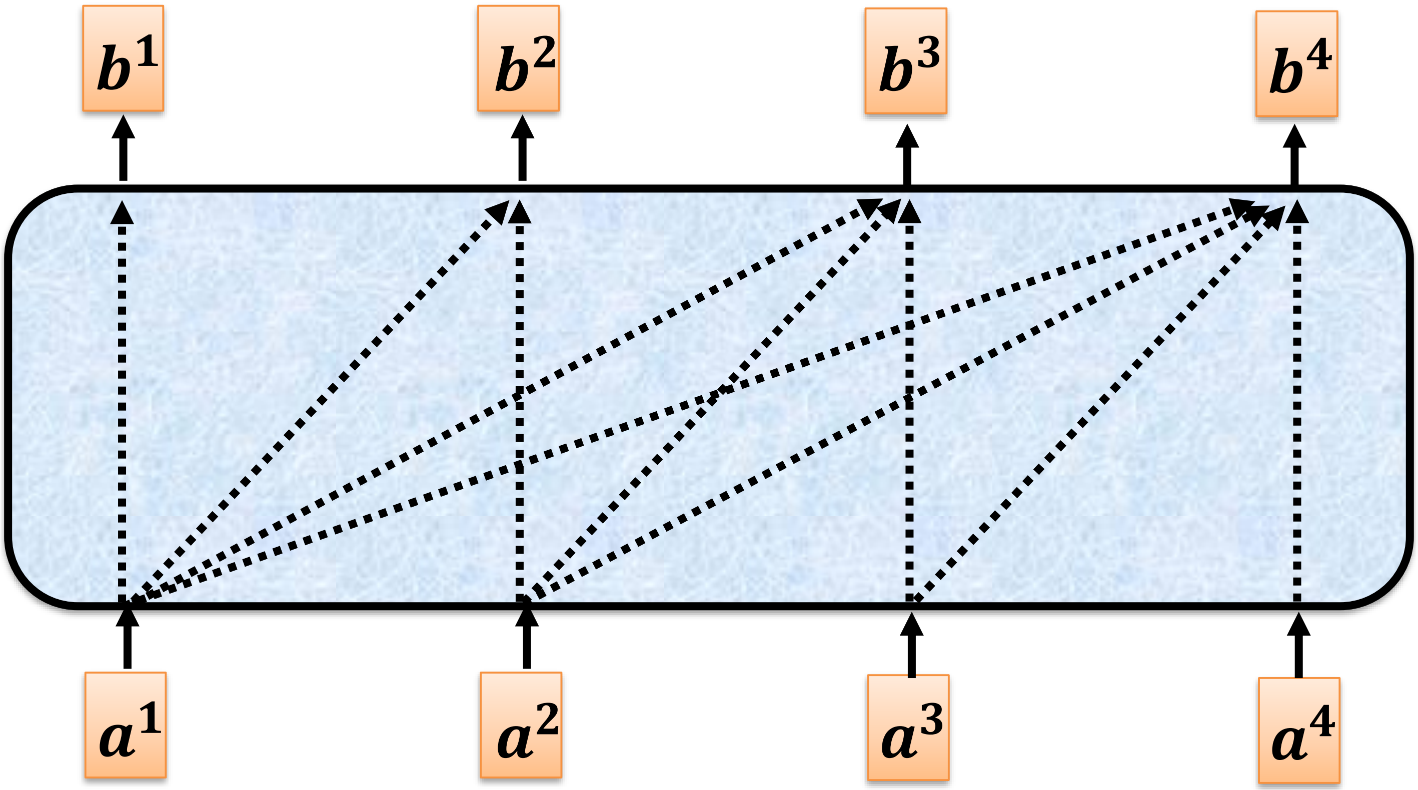
Transformer

- Decoder
  - Masked Self-Attention



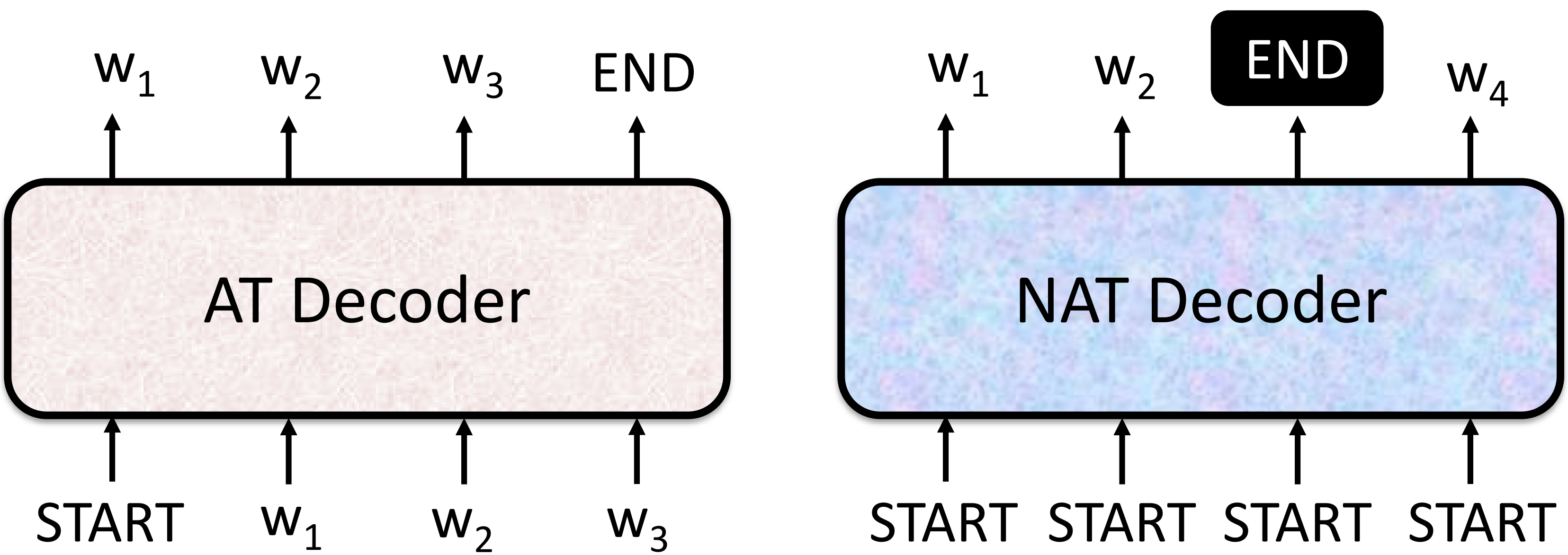
Transformer

- Decoder
  - Masked Self-Attention



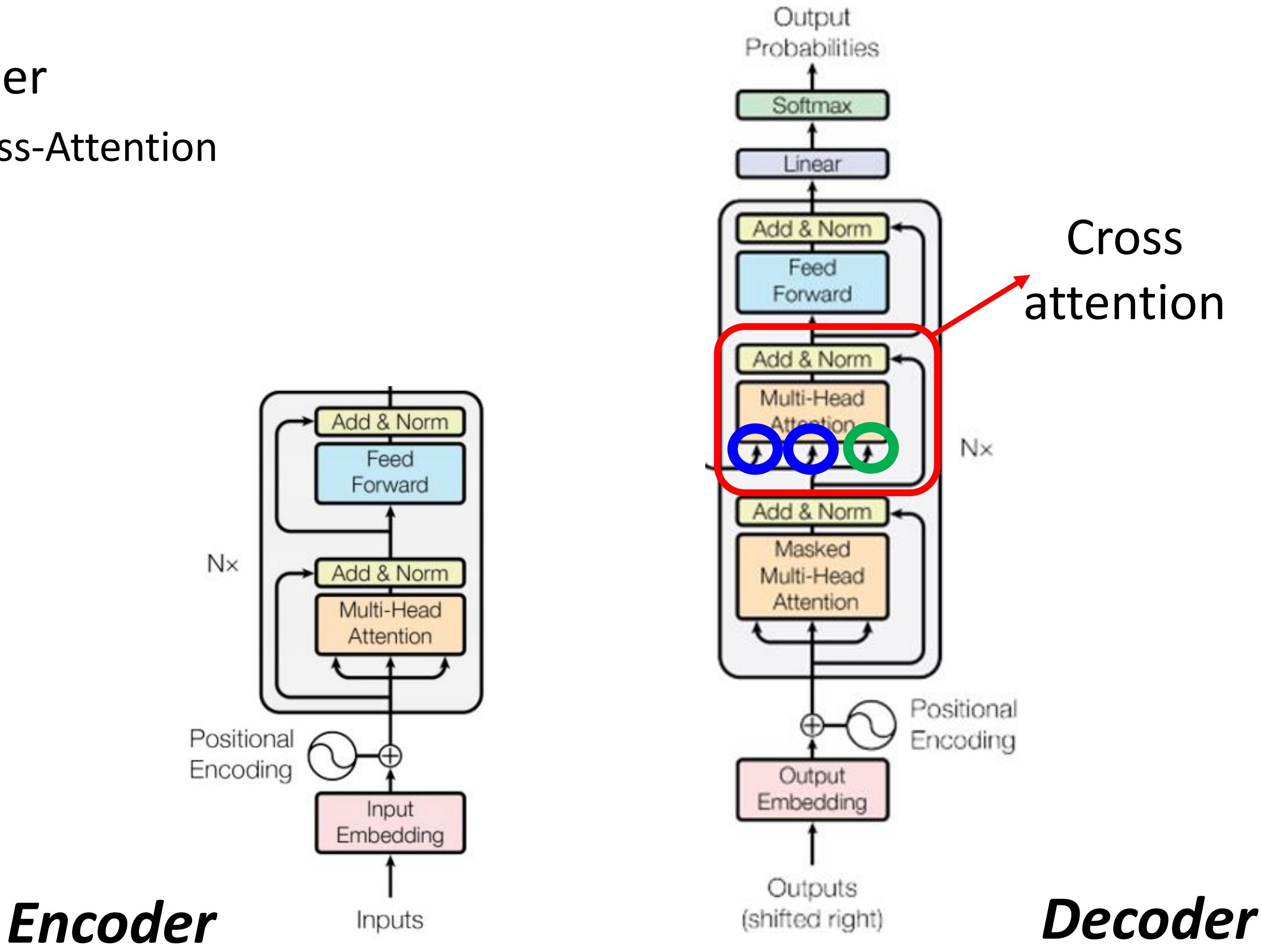
Transformer

- Decoder
  - Autoregressive (AT) -> Non-Autoregressive (NAT)



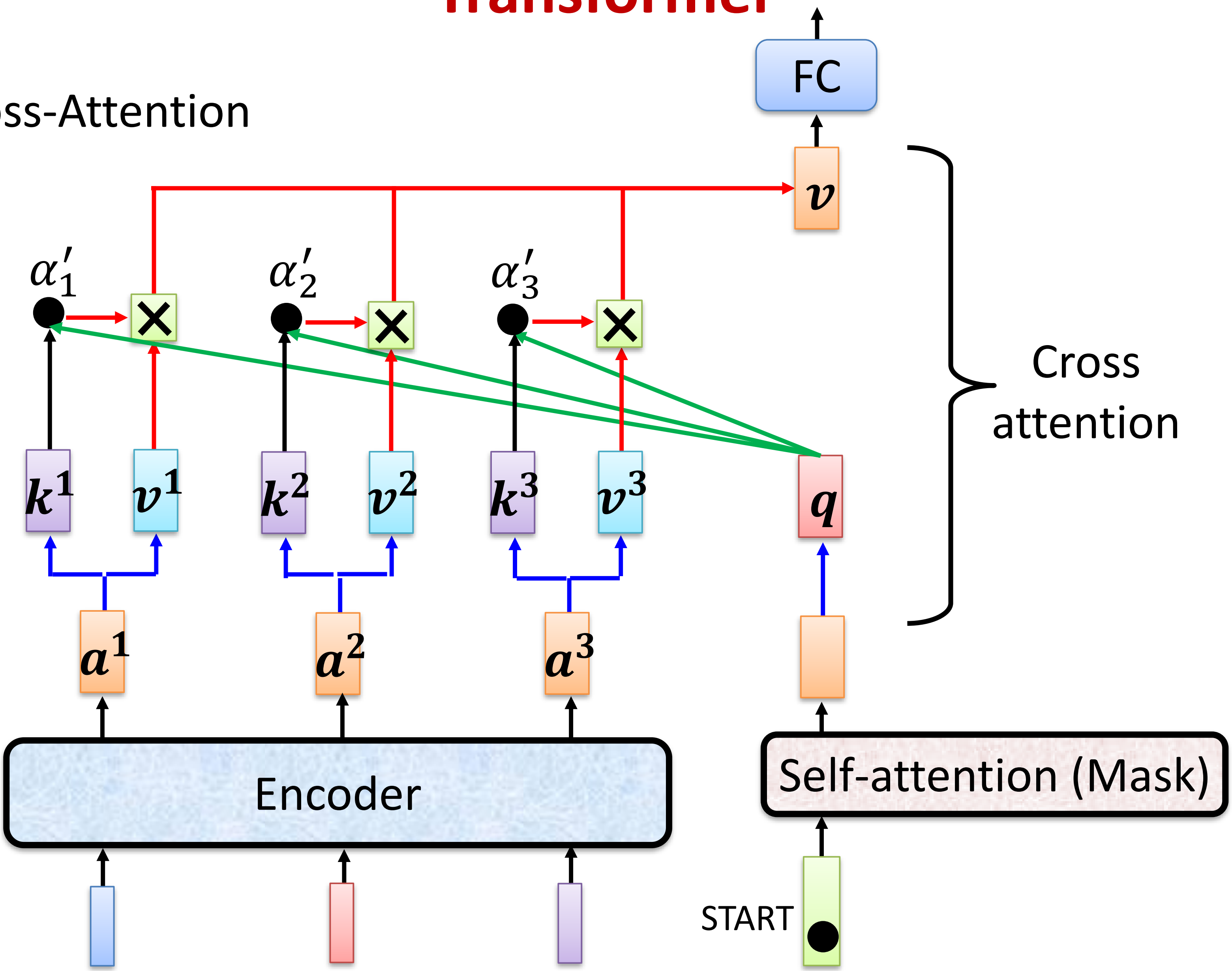
Transformer

- Decoder
  - Cross-Attention



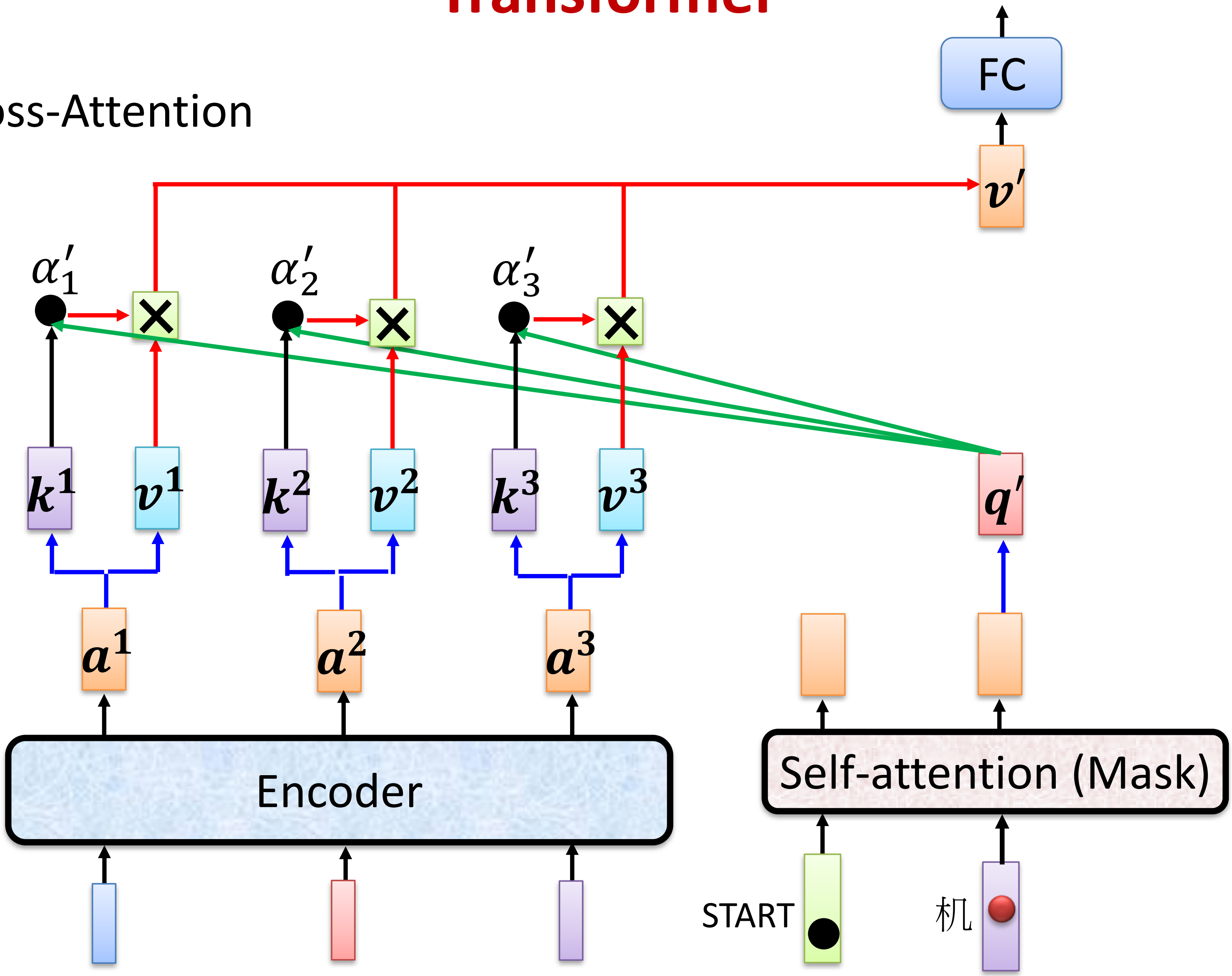
Transformer

- Cross-Attention



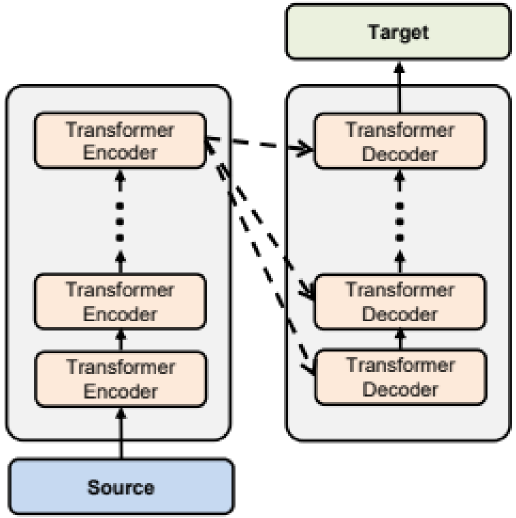
Transformer

- Cross-Attention

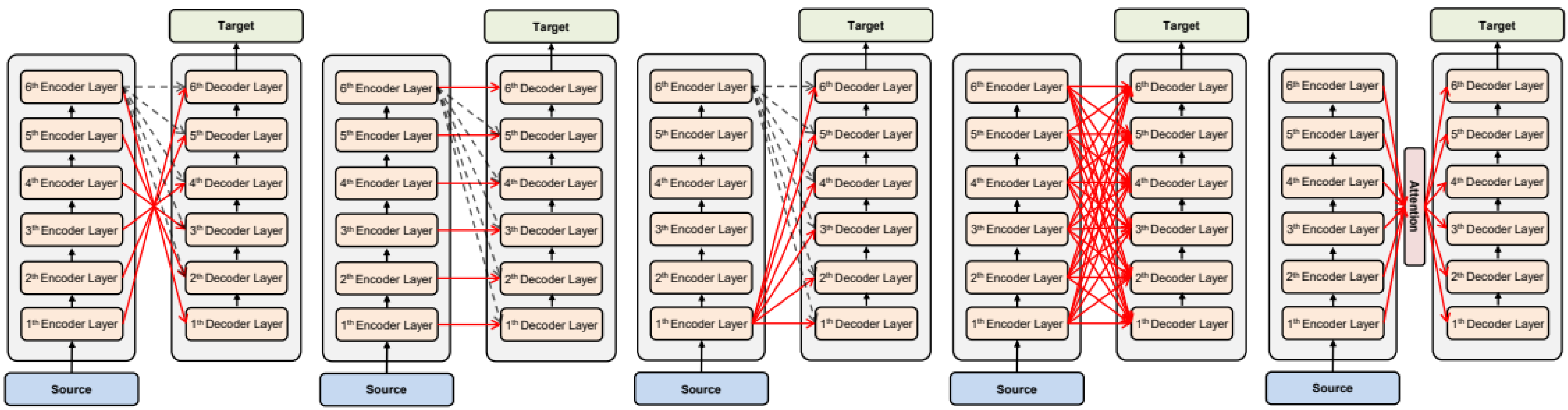


Transformer

- Cross-Attention
  - 图片来源: https://arxiv.org/abs/2005.08081



(a) Conventional Transformer



(a) Granularity Consistent Attention (b) Granularity Parallel Attention (c) Fine-Grained Attention (d) Full Matching Attention (e) Adaptive Matching Attention

课程作业

- 基于Transformer的机器翻译系统
  - **问题描述**
    - 利用Transformer，将输入的英文句翻译成中文
  - **数据集**
    - 输入： 一句英文 （e.g. tom is a student .）
    - 输出： 中文翻译 （e.g. 汤姆 是 个 学生 。）
    - 训练集：18000句
    - 验证集： 500句
    - 测试集： 2636句
  - **数据集下载地址**
    - 链接：https://pan.baidu.com/s/1Vb3PvFkfCvJ_JdapgEU_Hg
    - 提取码：h43i
  - **要求**
    - Tensorflow或者Pytorch实现

课程作业

- 基于Transformer的机器翻译系统
  - **评价指标**
    - BLEU score
      - BiLingual Evaluation Understudy, IBM
      - 将机器翻译产生的候选译文与人翻译的多个参考译文相比较，越接近，候选译文的正确率越高。
      - 统计同时出现在系统译文和参考译文中的n元词的个数，最后把匹配到的n元词的数目除以系统译文的n元词数目，得到评测结果。

课程作业

- 基于Transformer的机器翻译系统
  - **评价指标**

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

最大语法的阶数，实际取4。

出现在答案译文中的*n*元词语接续组占候选译文中*n*元词语接续组总数的比例。

长度过短句子的惩罚因子

$w_n = 1/N$

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}$$

***c*为候选译文中单词的个数，*r* 为答案译文中与*c* 最接近的译文单词个数。**

BLEU 分值范围：0 ~ 1，分值越高表示译文质量越好，分值越小，译文质量越差。

课程作业

- 基于Transformer的机器翻译系统
  - **参考资料**
    - 《神经网络与深度学习》 第15.4.2小节
    - http://nlp.seas.harvard.edu/2018/04/03/attention.html
    - Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I.. Attention is all you need. In NIPS 2017.
    - Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., & Chao, L. S. (2019, July). Learning Deep Transformer Models for Machine Translation. In ACL 2019.

欢迎提问！