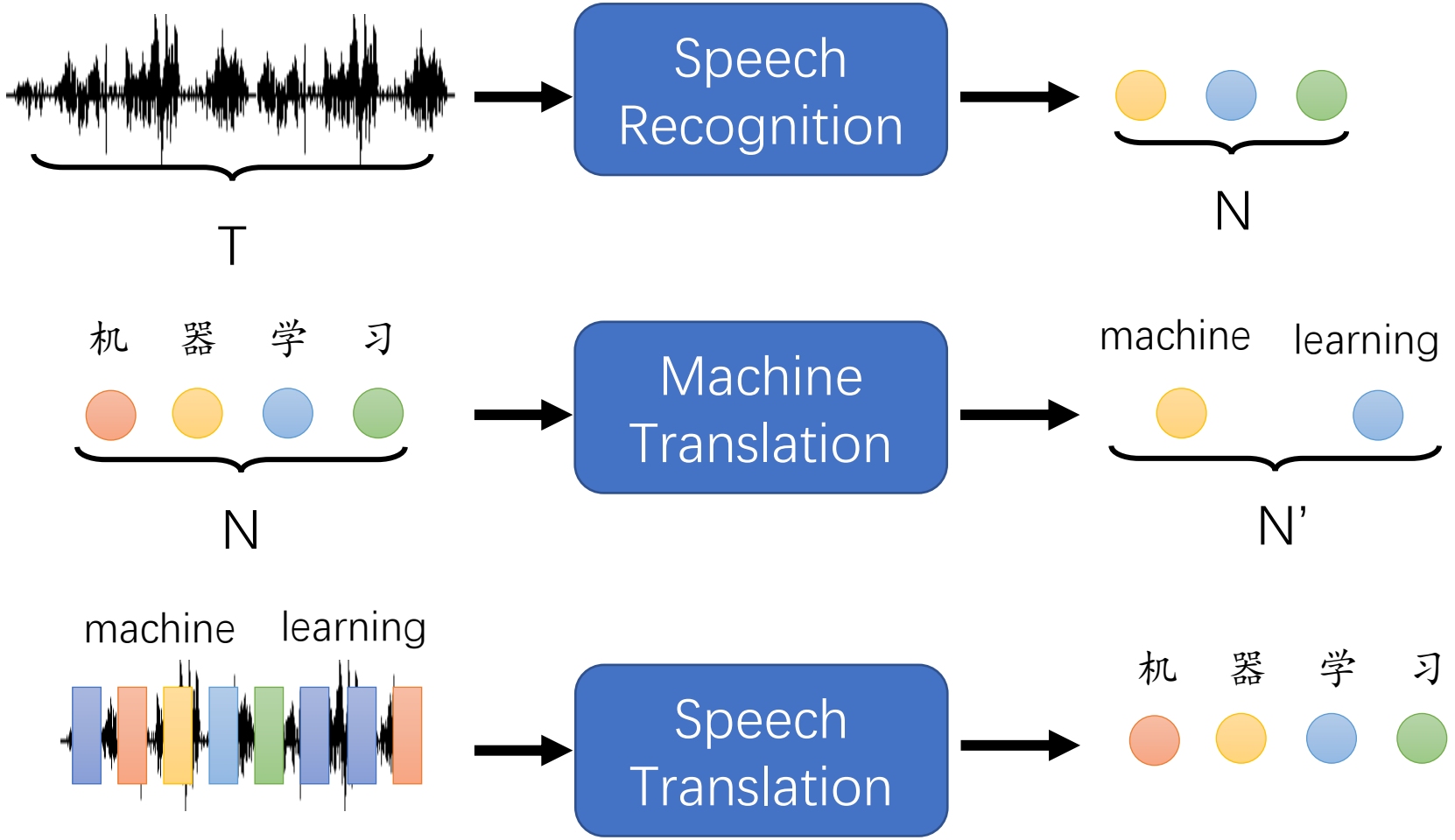
机器学习

苏州大学计算机科学与技术学院
自然语言处理实验室
主讲：周夏冰
邮箱：zhouxiabing@suda.edu.cn

Sequence-to-sequence (Seq2seq)



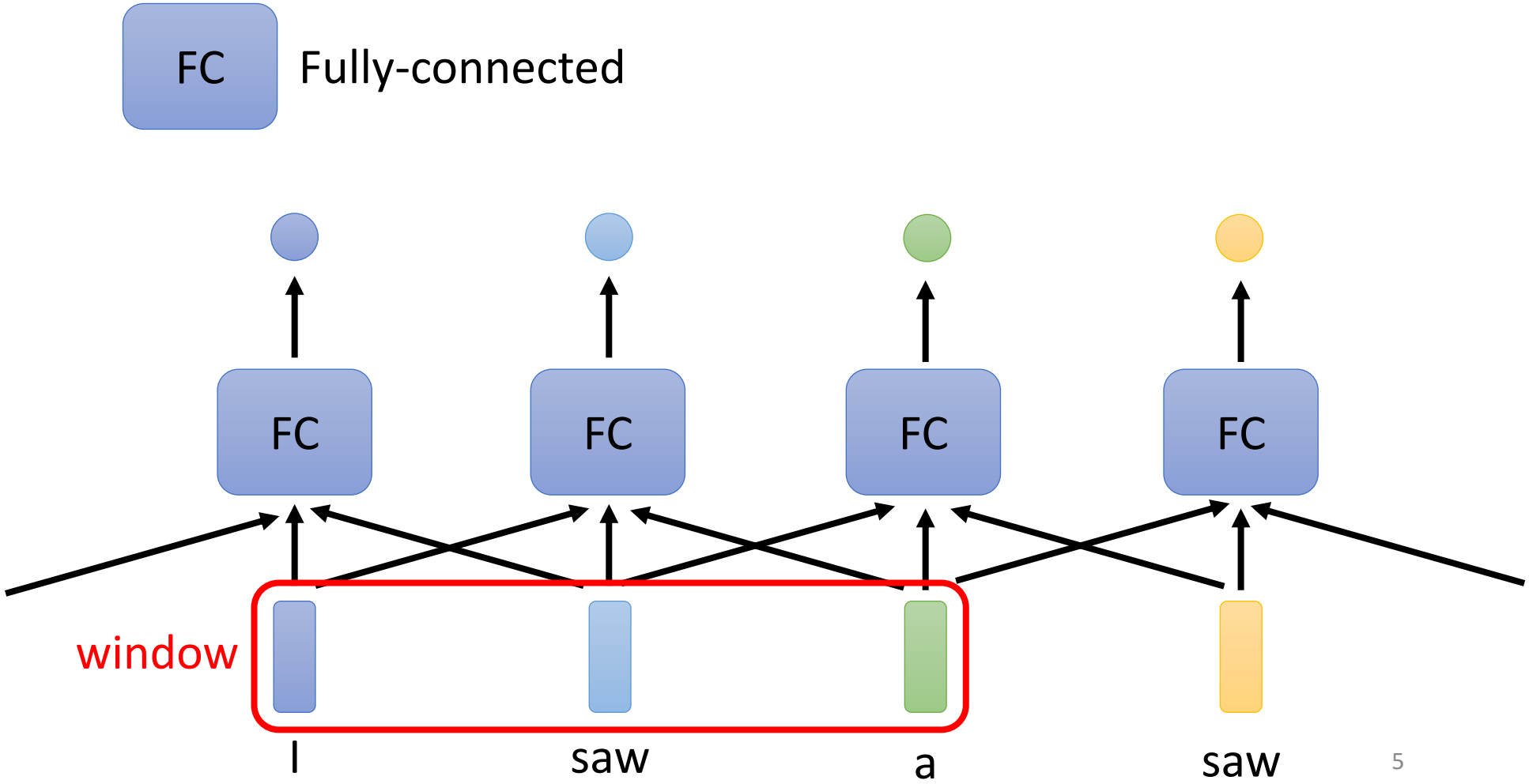
01

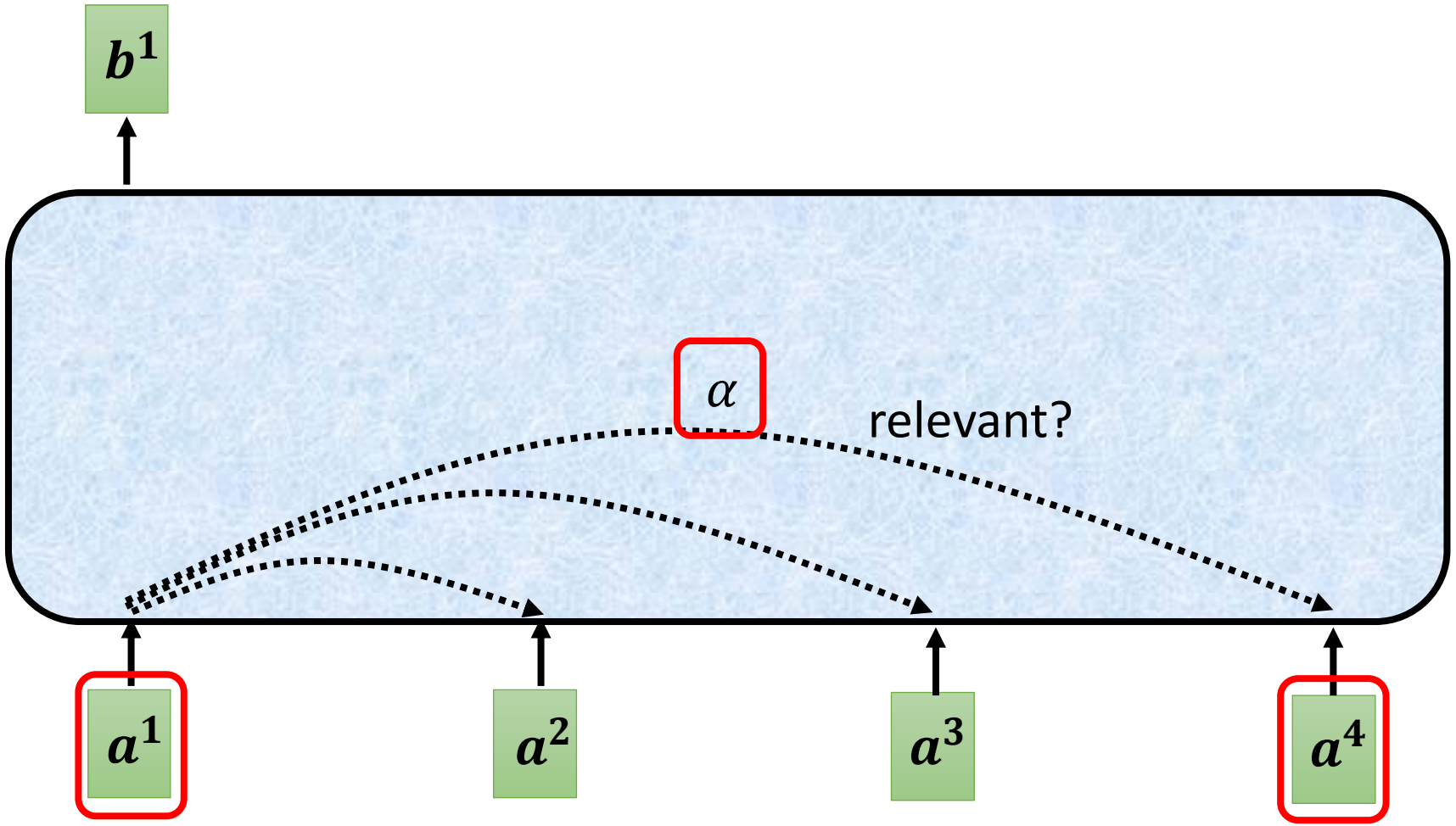
自注意力机制

注意力机制

- 注意力机制（Attention Mechanism）是一种模仿人类视觉和认知系统的方法
- 在处理输入数据时集中注意力于相关的部分

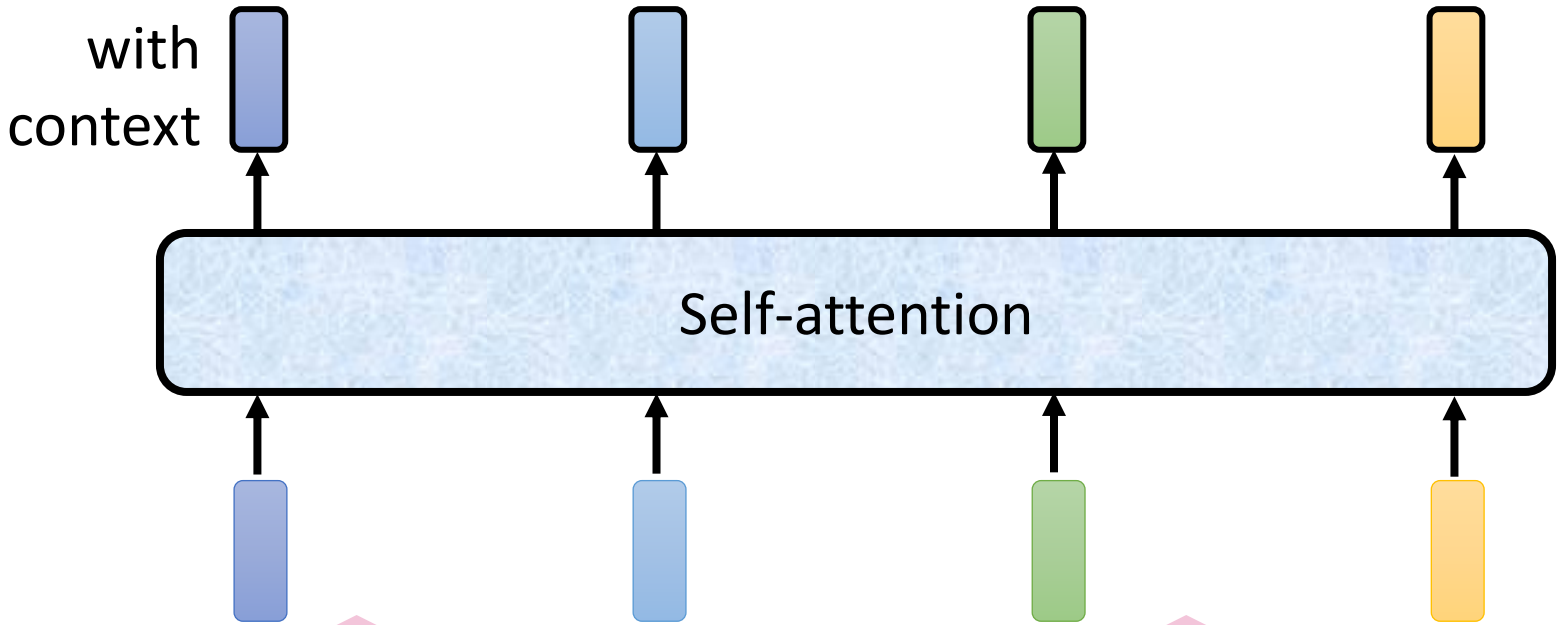
Sequence labeling



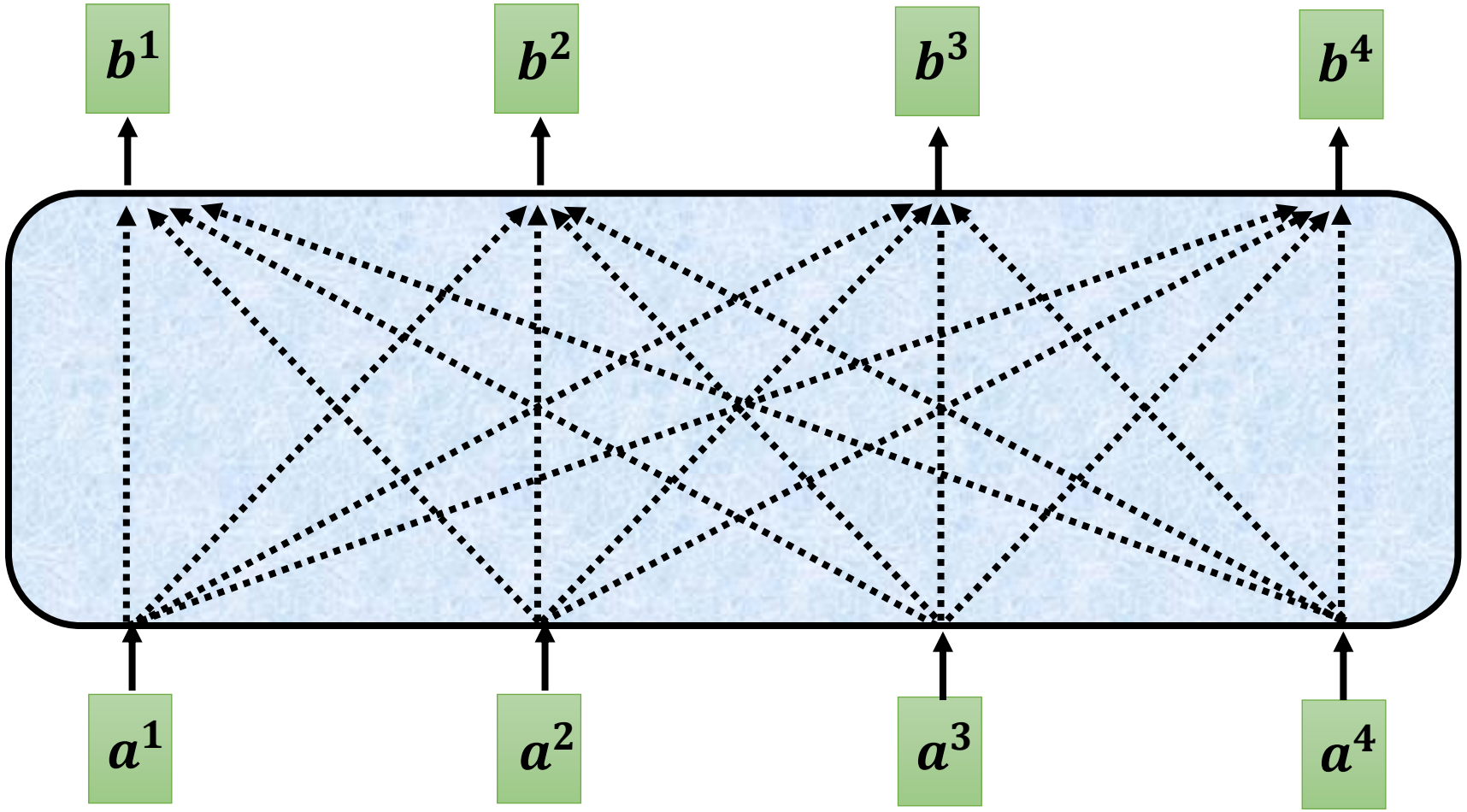


寻找输入中相关的信息

Self-attention

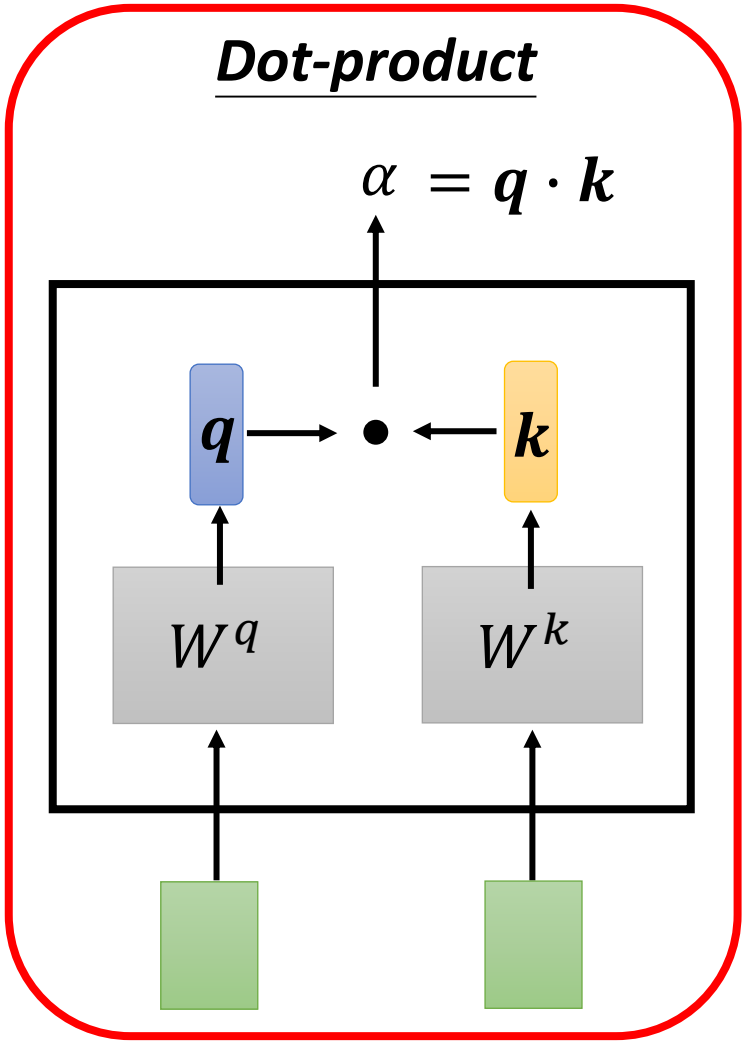


Self-attention



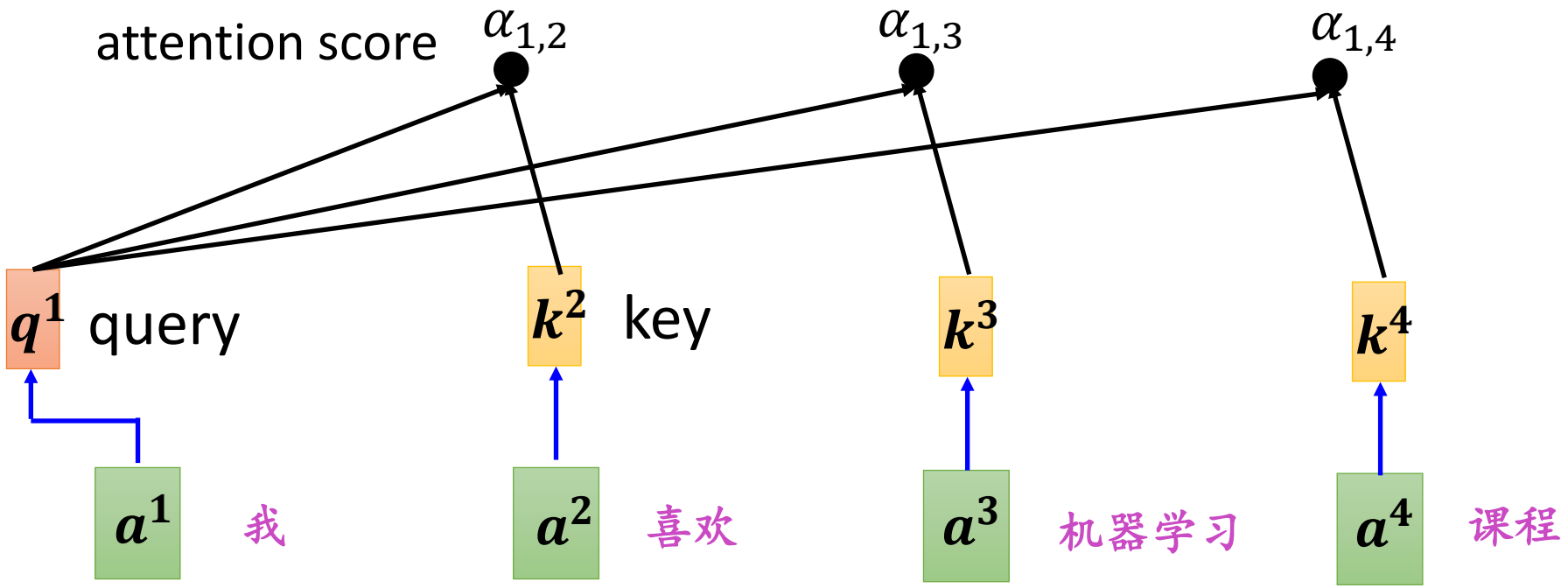
Self-attention

- Key, Query, Value
- Query: 查询向量
- Key: 键，用来计算和Query相关性
- Value: 值，用来获取信息的向量



Self-attention

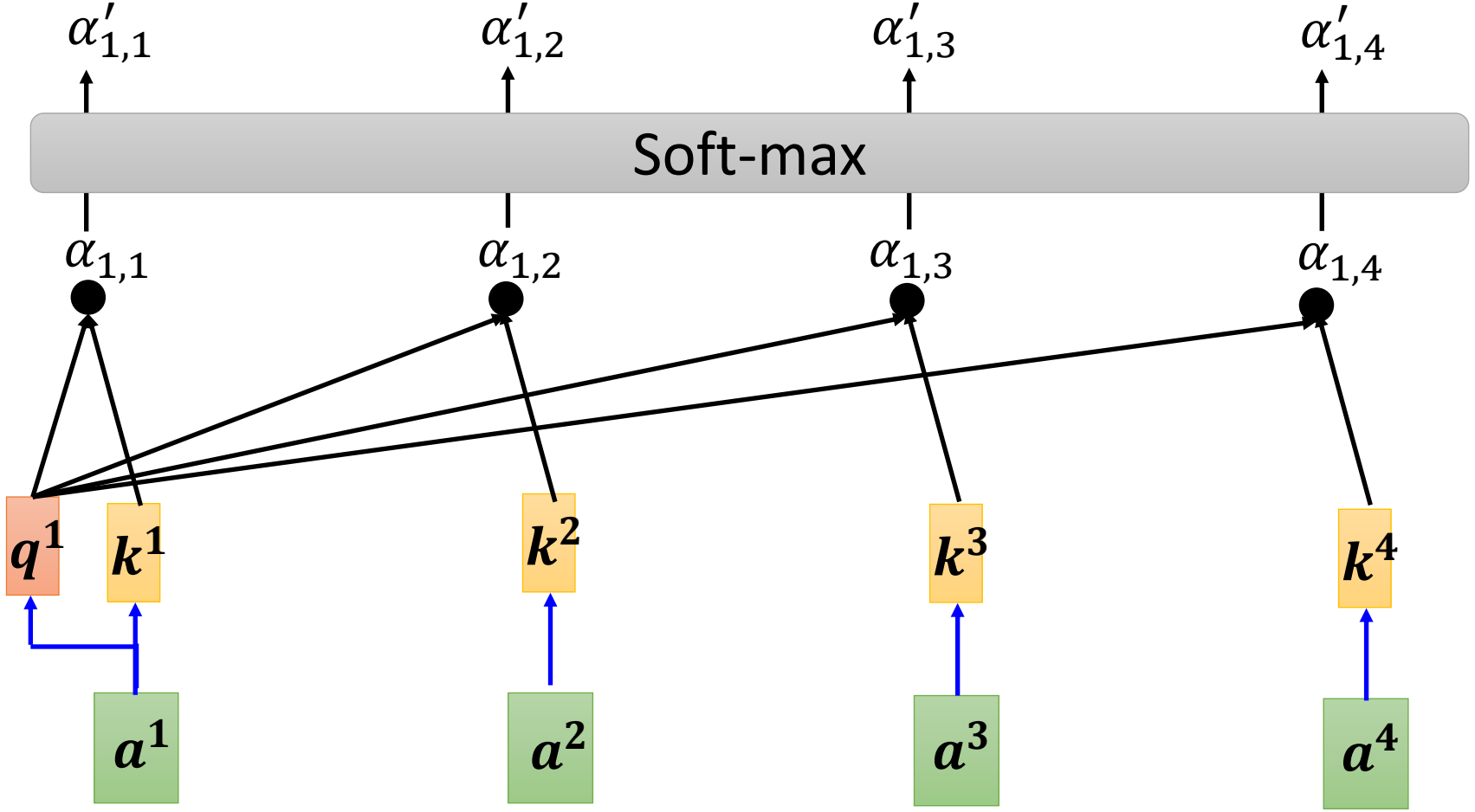
$$\alpha_{1,2} = \boldsymbol{q^1} \cdot \boldsymbol{k^2} \qquad \alpha_{1,3} = \boldsymbol{q^1} \cdot \boldsymbol{k^3} \qquad \alpha_{1,4} = \boldsymbol{q^1} \cdot \boldsymbol{k^4}$$



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

Self-attention

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

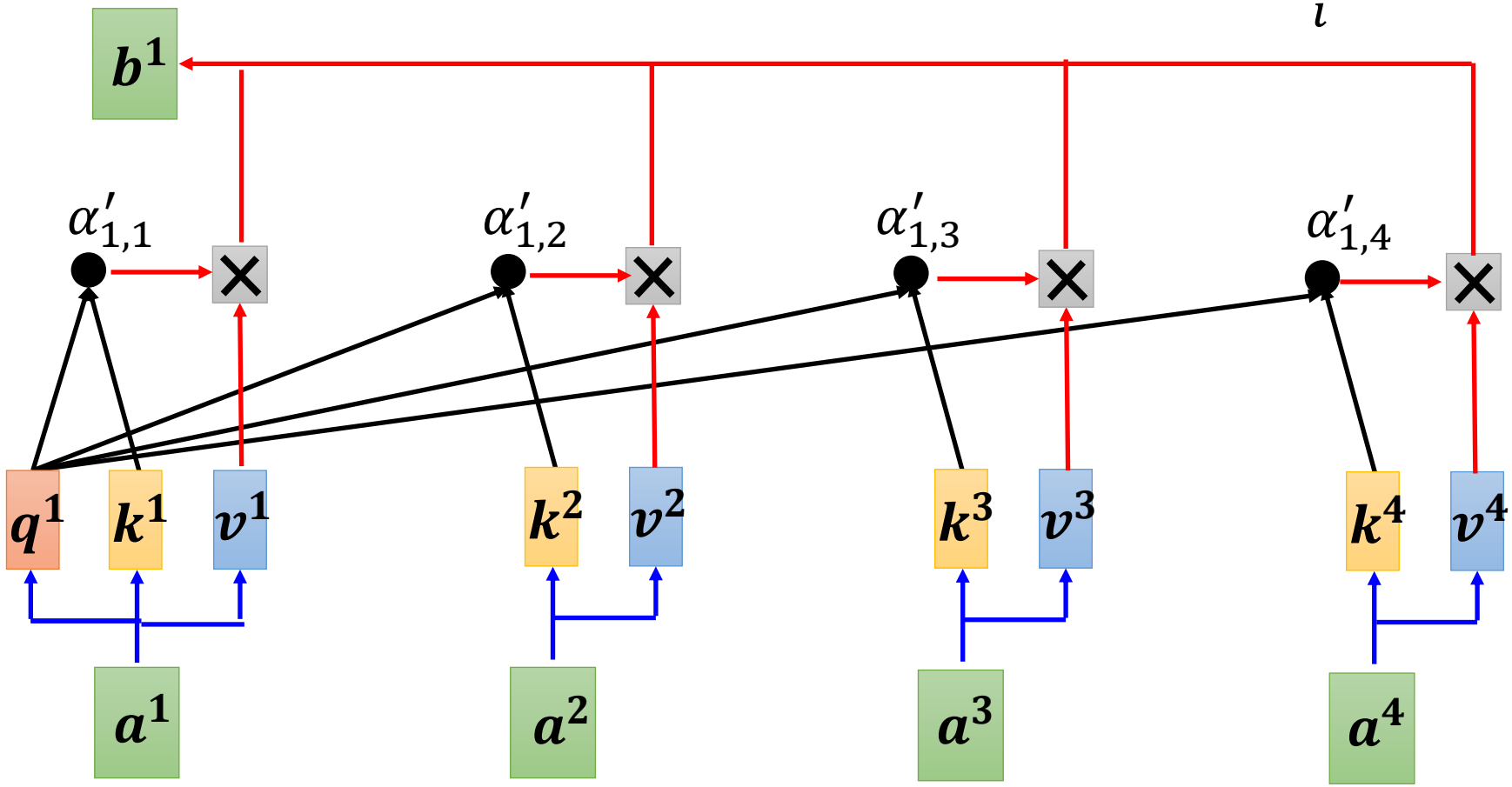


$q^1 = W^q a^1$ $k^2 = W^k a^2$ $k^3 = W^k a^3$ $k^4 = W^k a^4$

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

Self-attention

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

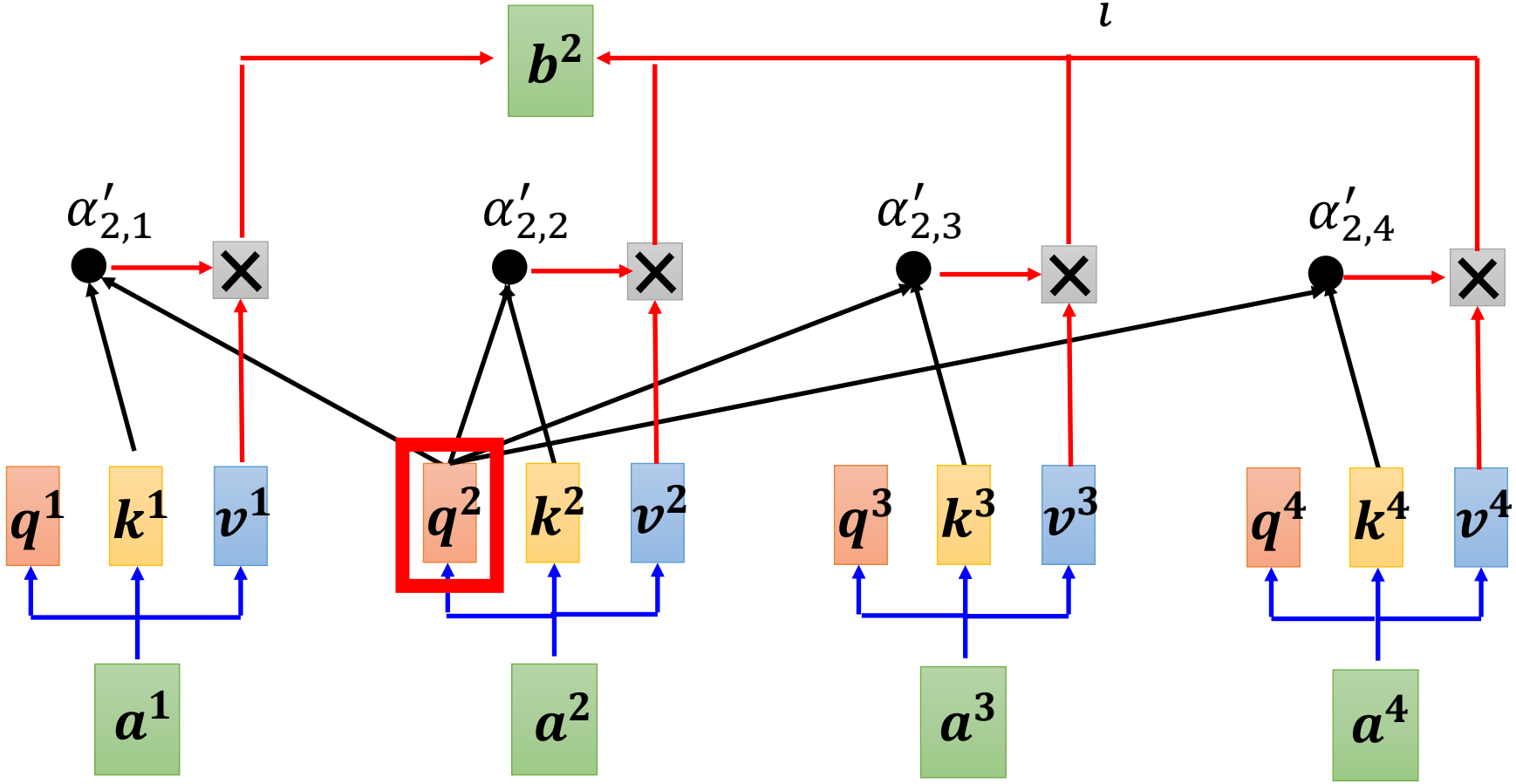


根据权重抽取信息

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

Self-attention

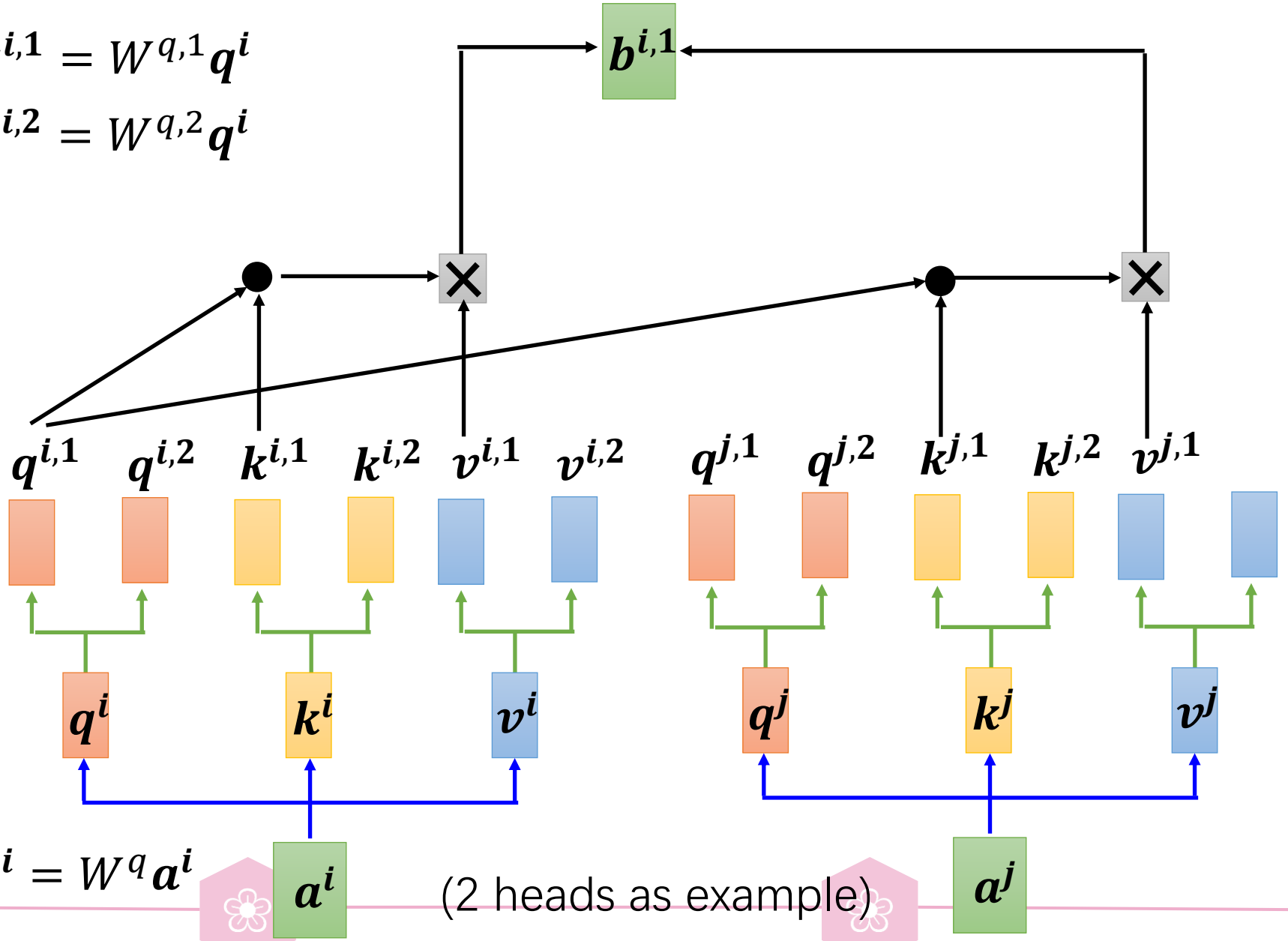
$$b^2 = \sum_i \alpha'_{2,i} v^i$$



Multi-head Self-attention

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

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



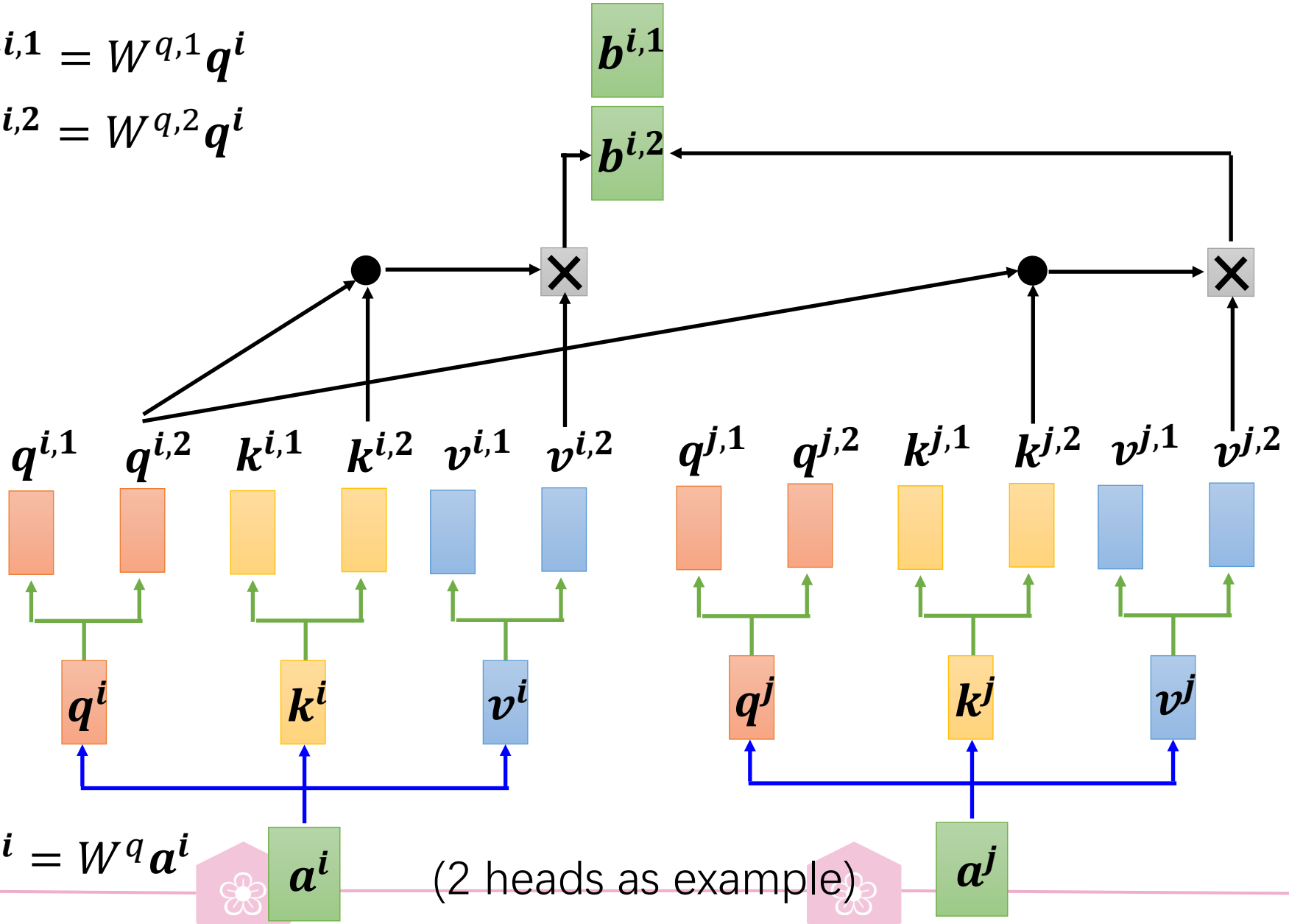
$$q^i = W^q a^i$$

(2 heads as example)

Multi-head Self-attention

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

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

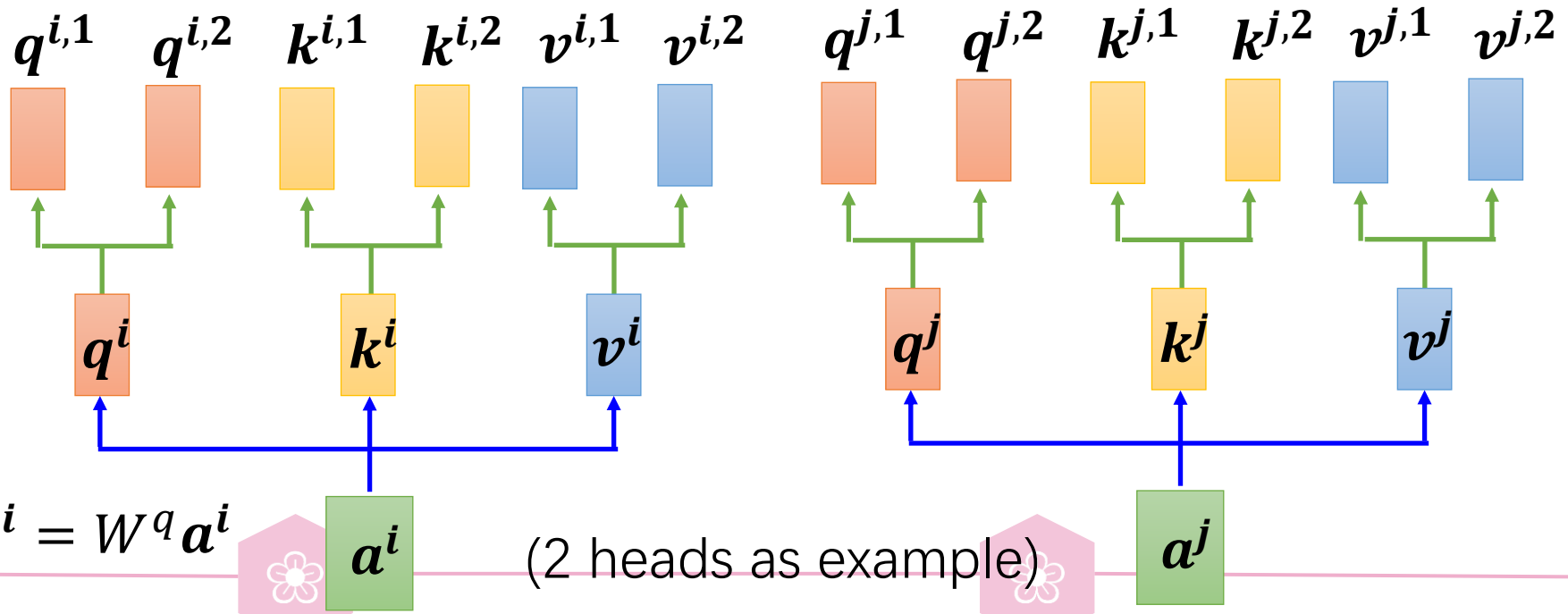


$$q^i = W^q a^i$$

(2 heads as example)

Multi-head Self-attention

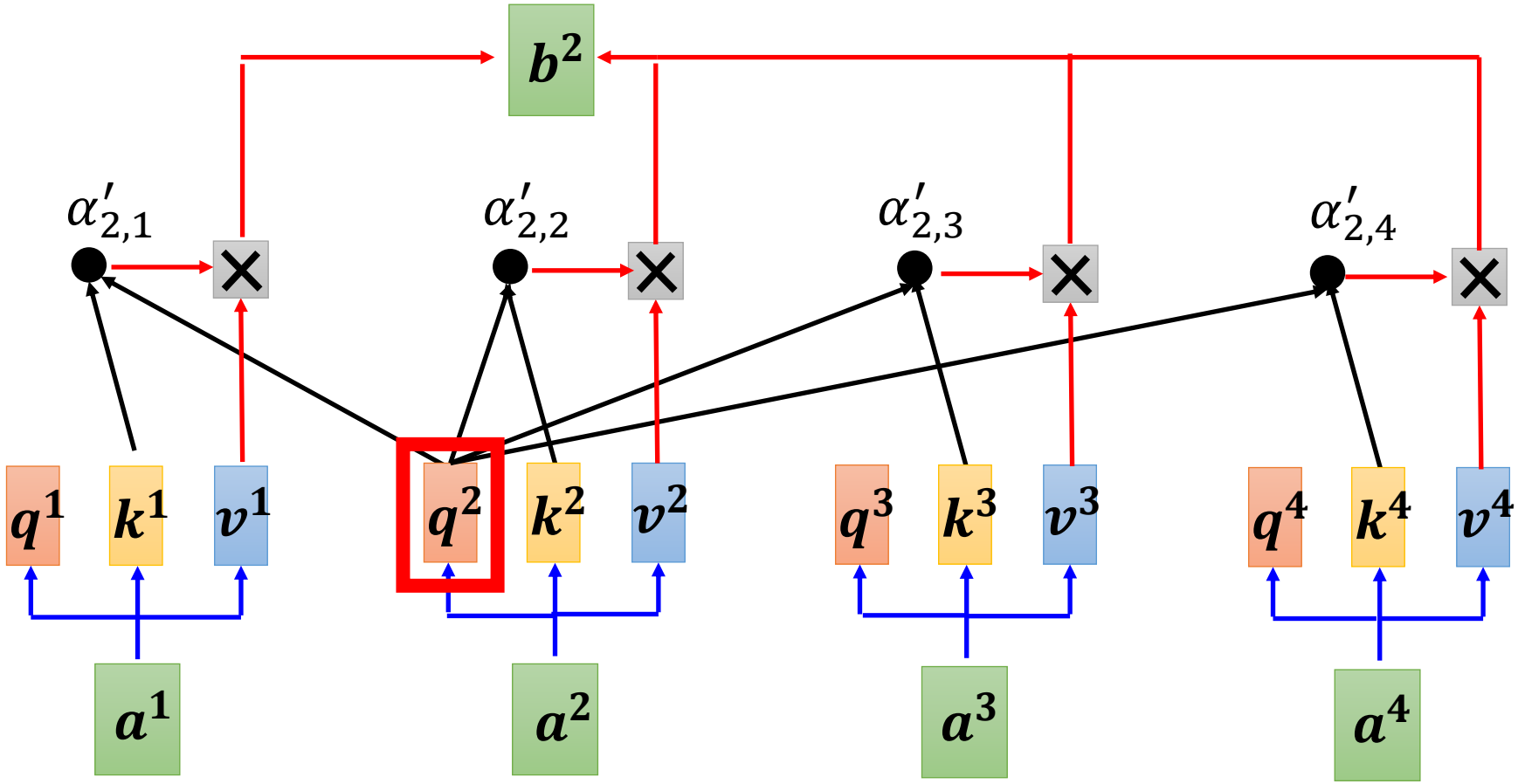
$$b^i = W^O \begin{bmatrix} b^{i,1} \\ b^{i,2} \end{bmatrix}$$



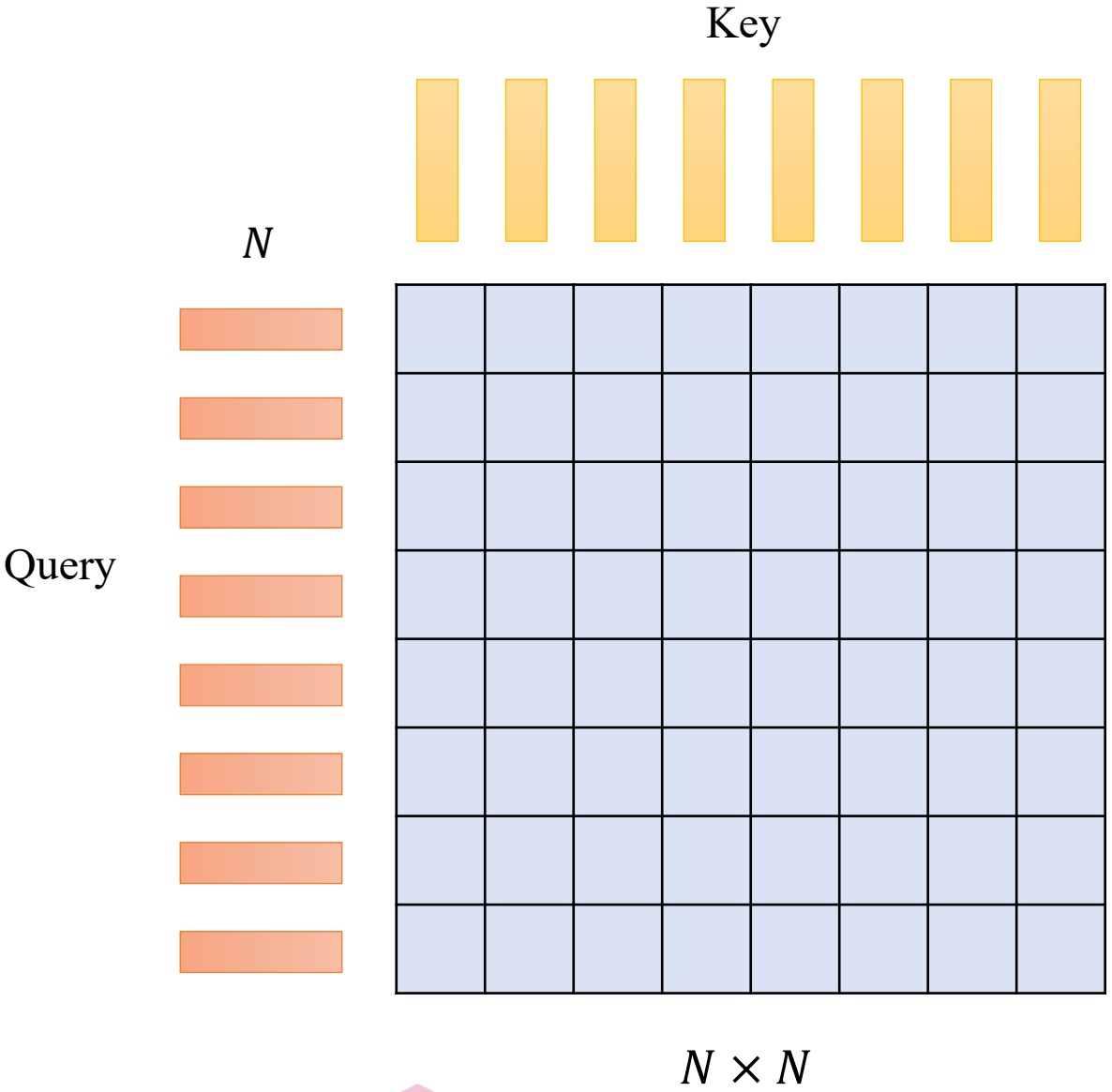
$q^i = W^q a^i$

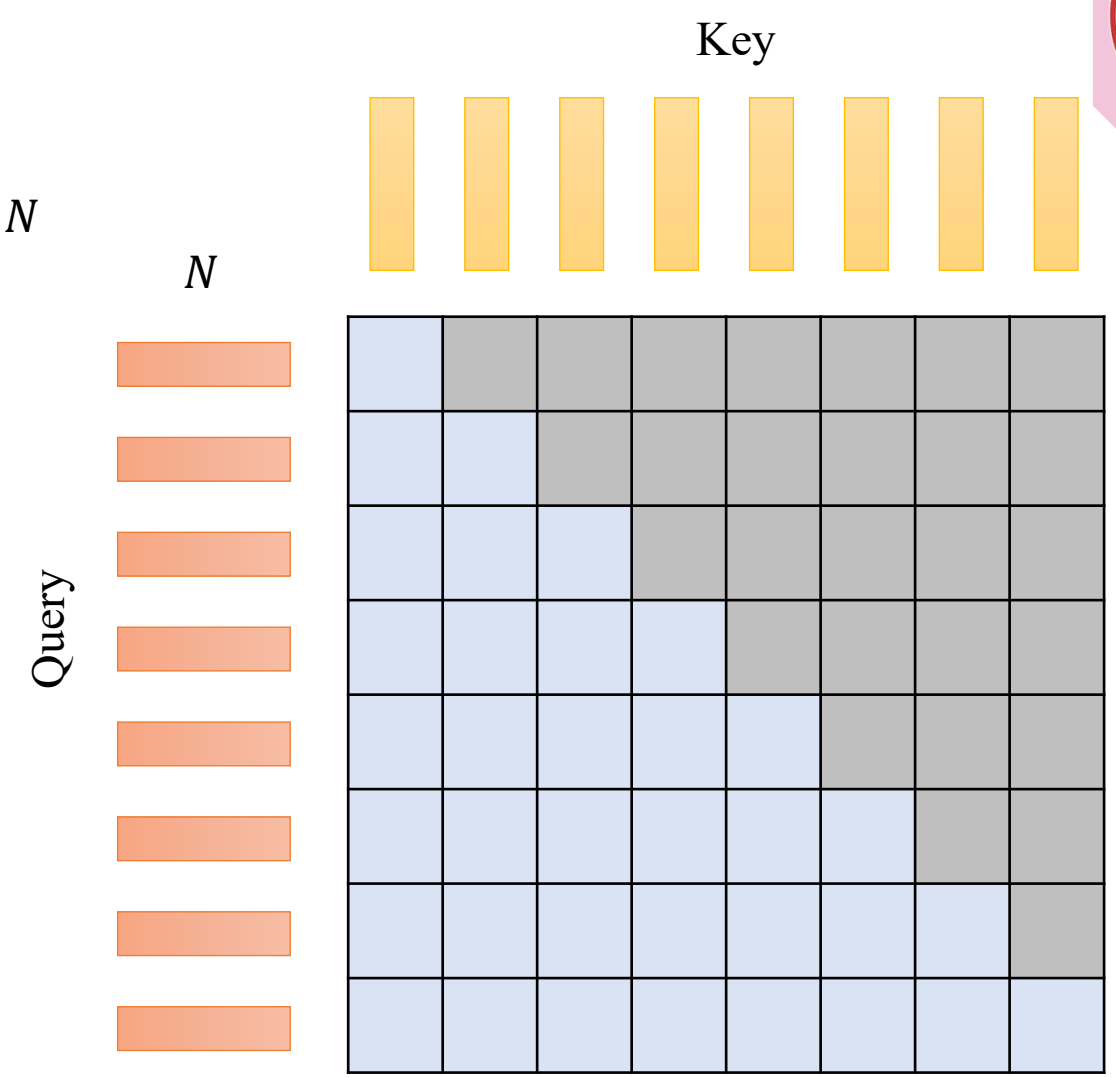
(2 heads as example)

Self-attention → Masked Self-attention



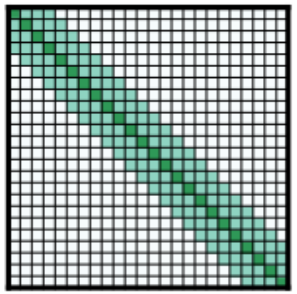
Attention Matrix



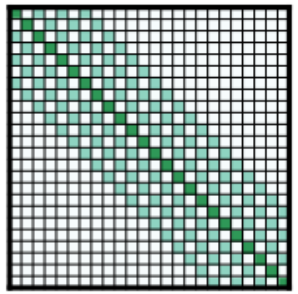


- Longformer

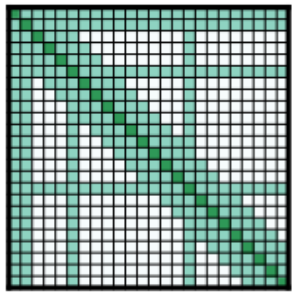
https://arxiv.org/abs/2004.05150



(b) Sliding window attention



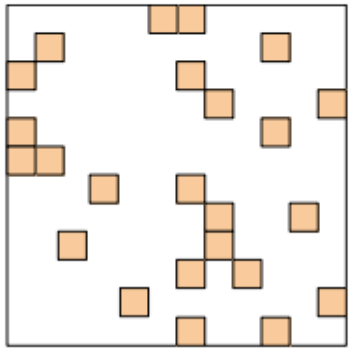
(c) Dilated sliding window



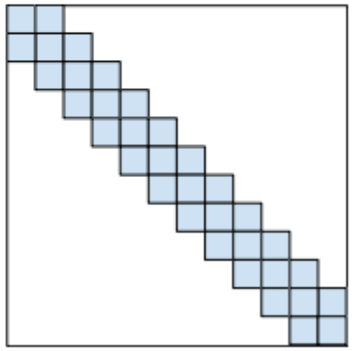
(d) Global+sliding window

- Big Bird

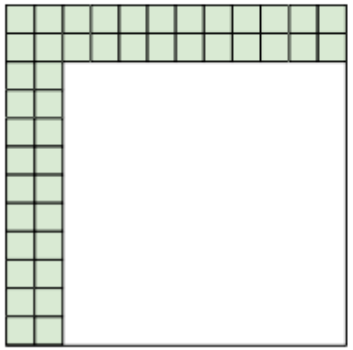
https://arxiv.org/abs/2007.14062



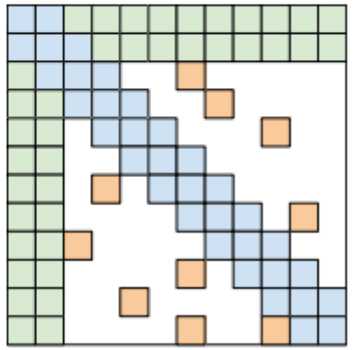
(a) Random attention



(b) Window attention

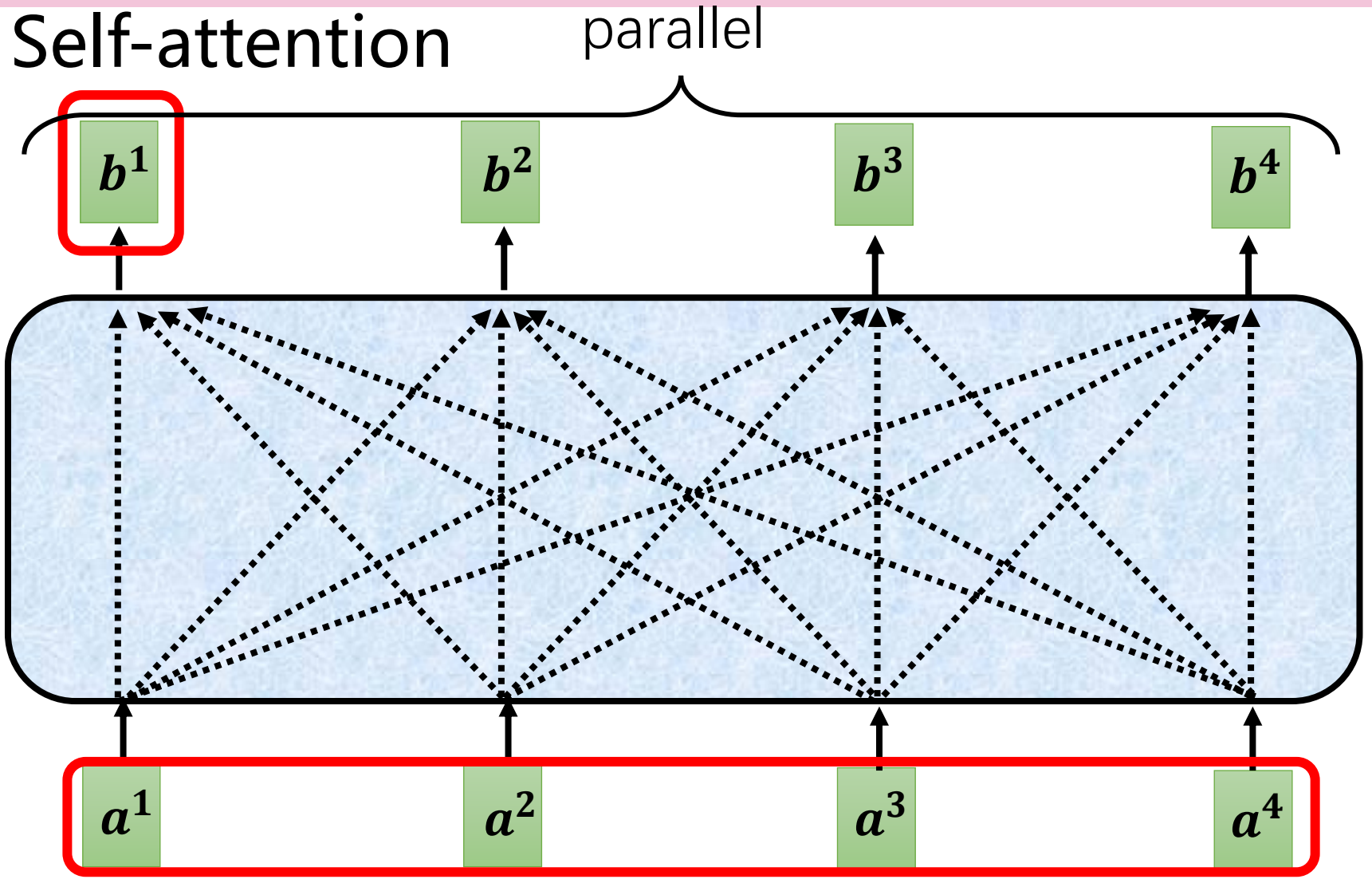


(c) Global Attention



(d) BIGBIRD

Self-attention



$$Q = W^q I$$

$$K = W^k I$$

$$V = W^v I$$

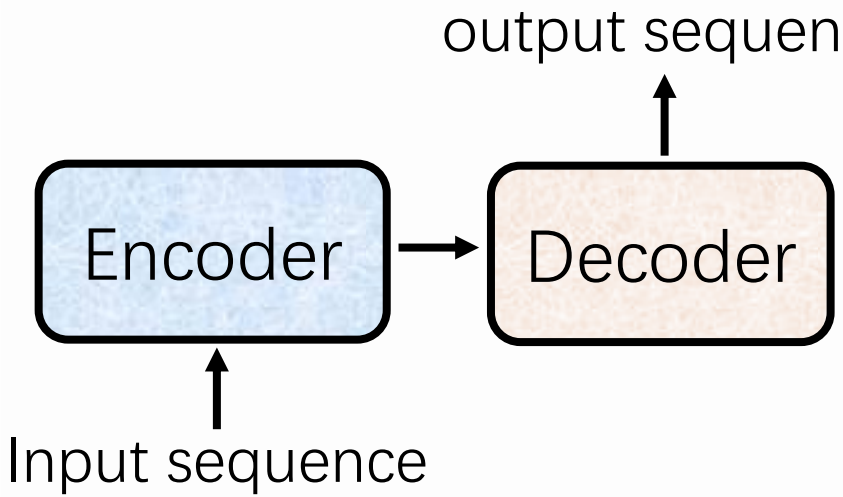
$$A' \leftarrow A = K^T Q$$

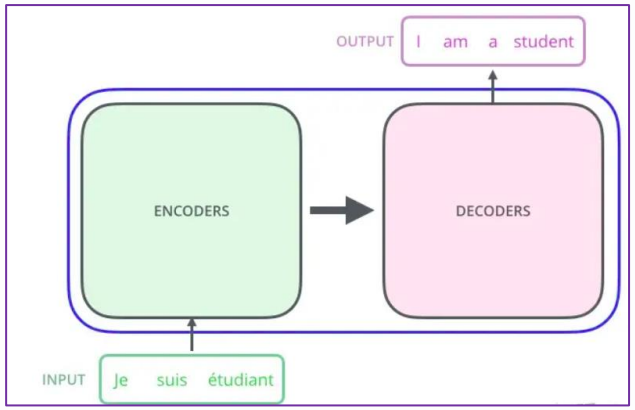
$$O = V A'$$

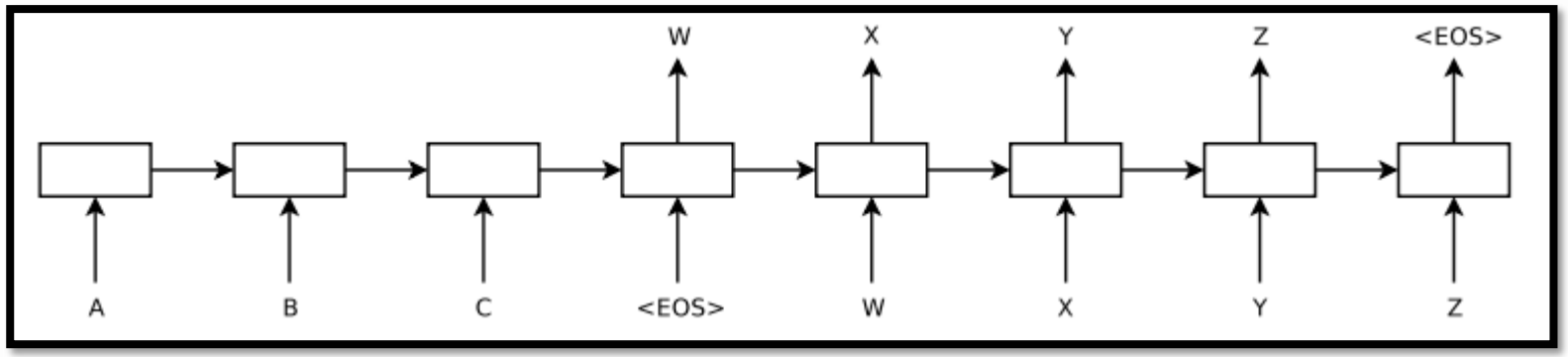
02

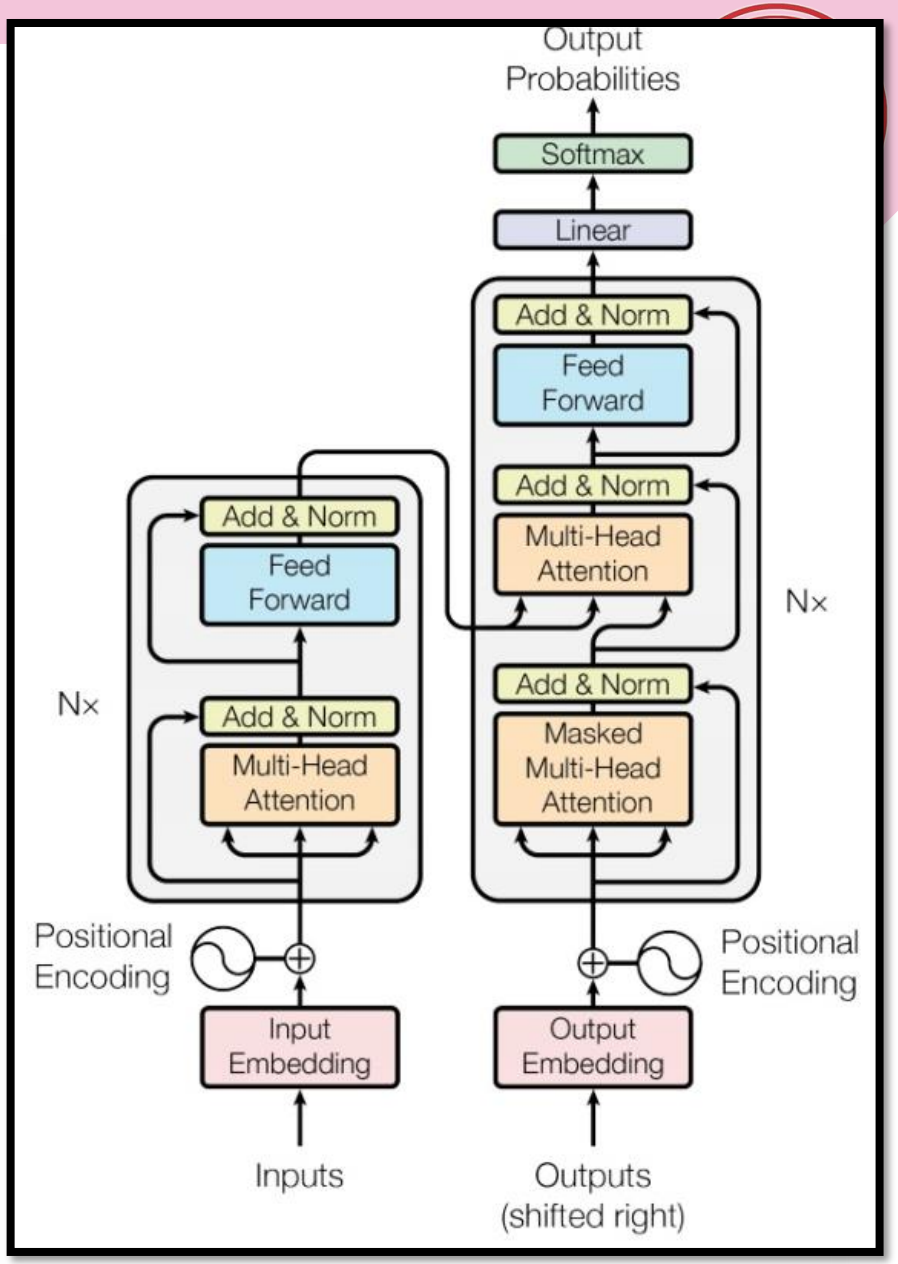
Transformer

Seq2seq



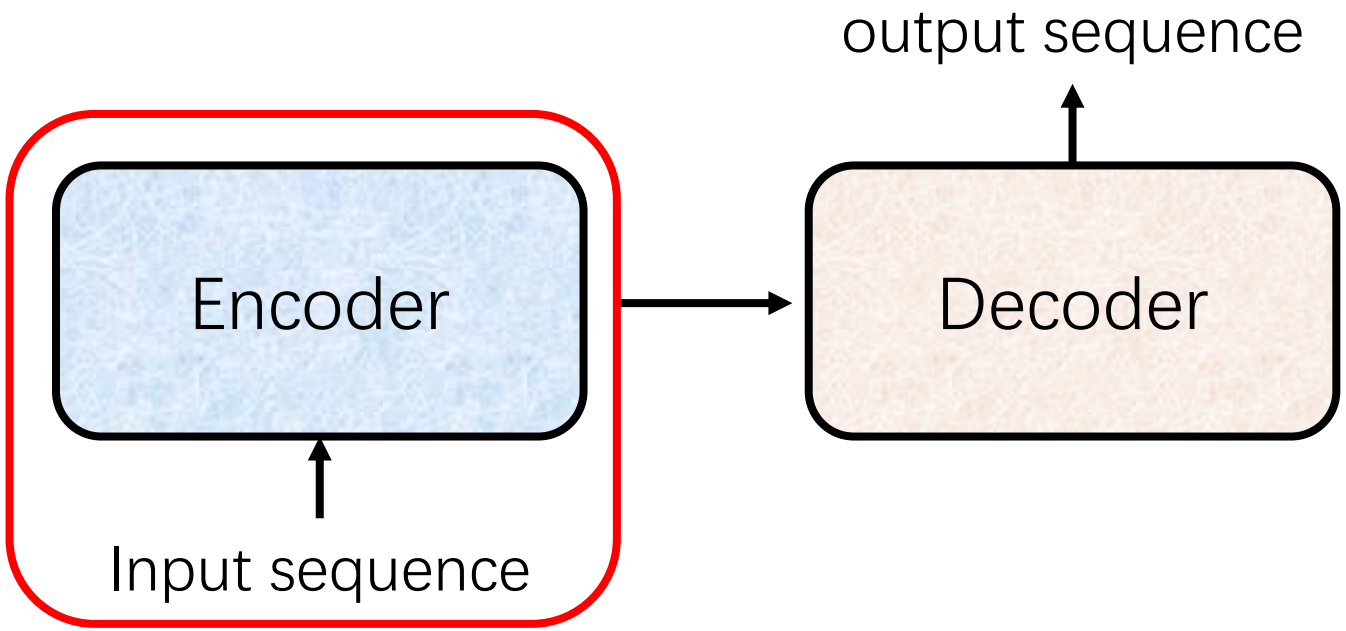






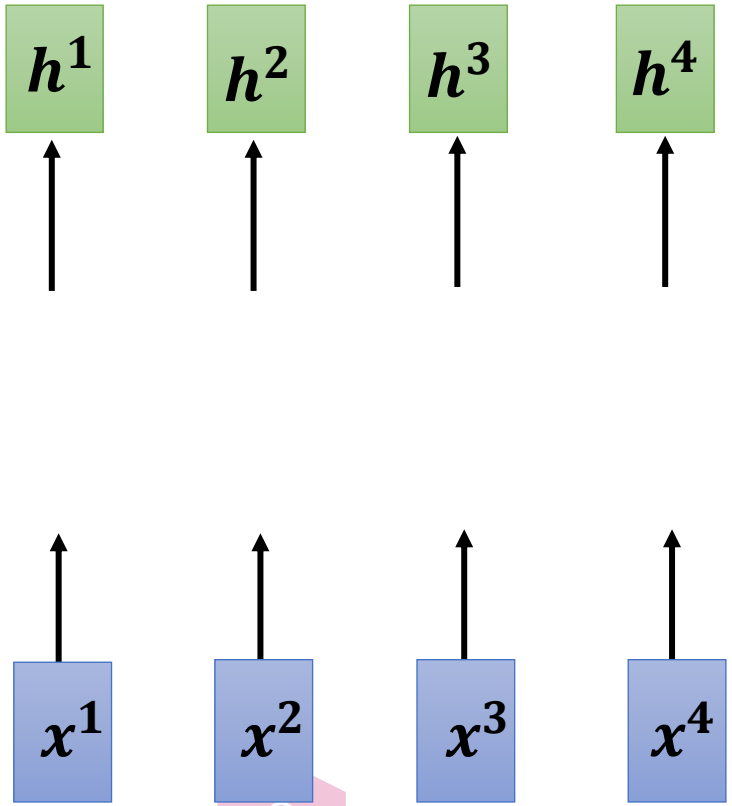
Transformer

https://arxiv.org/abs/1706.03762

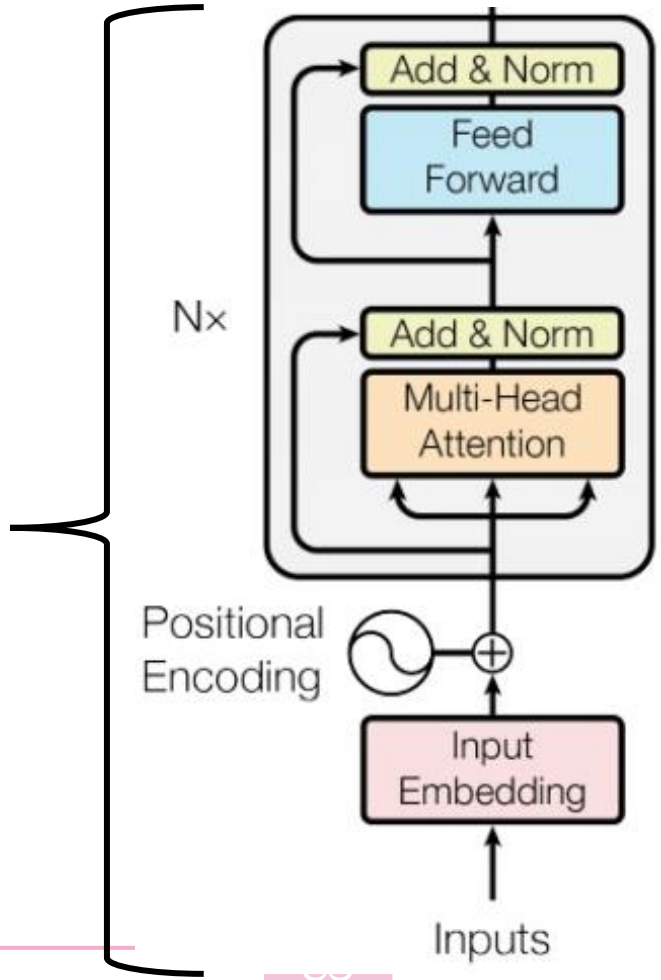

output sequence
Encoder
Decoder
Input sequence

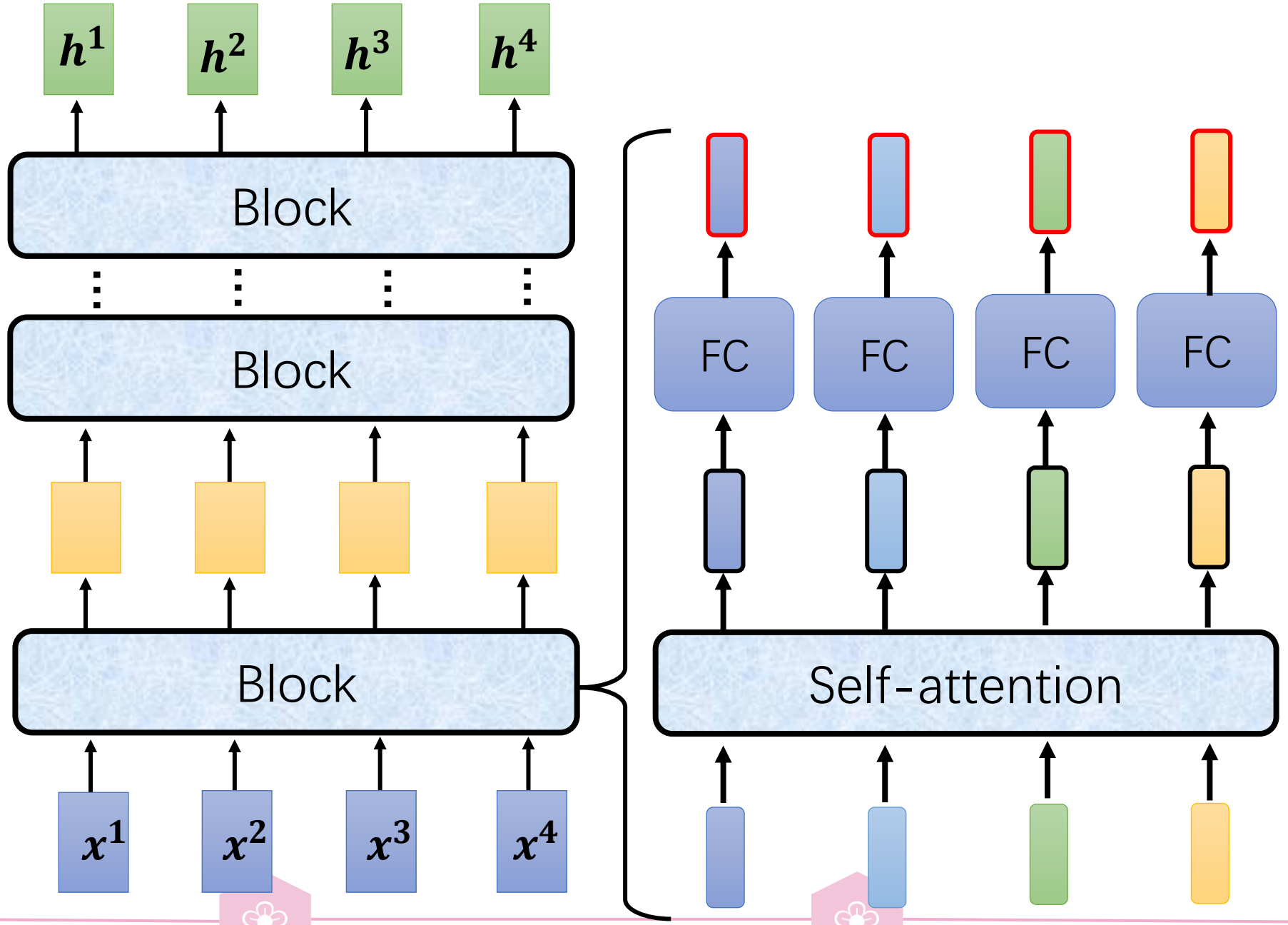
Encoder

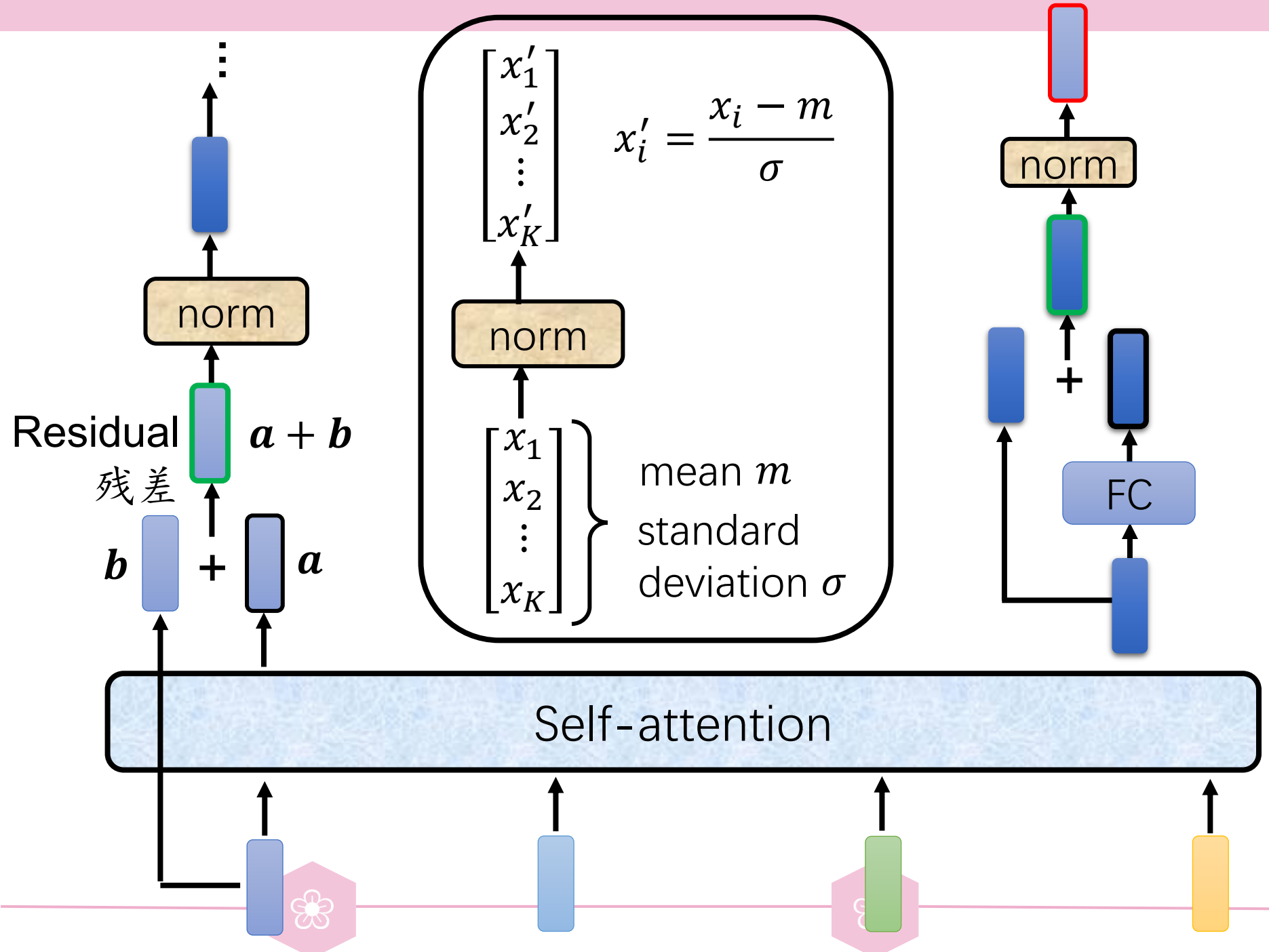
You can use **RNN** or **CNN**.



Transformer's Encoder




h^1
h^2
h^3
h^4
Block
Block
Block
x^1
x^2
x^3
x^4
FC
FC
FC
FC
Self-attention



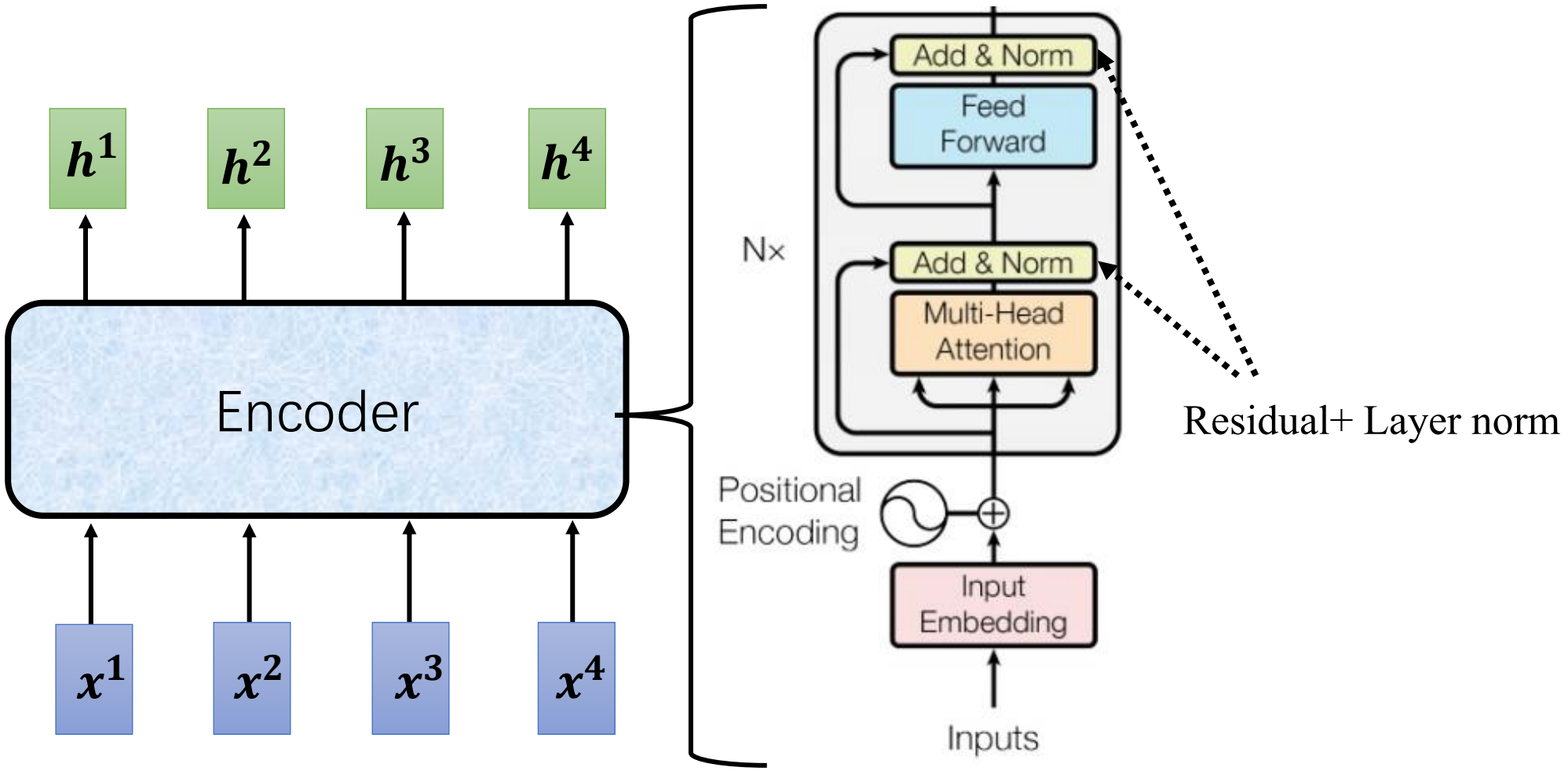
残差：
稳定训练、帮助梯度传播、避免退化问题

Norm:
稳定训练、加速收敛

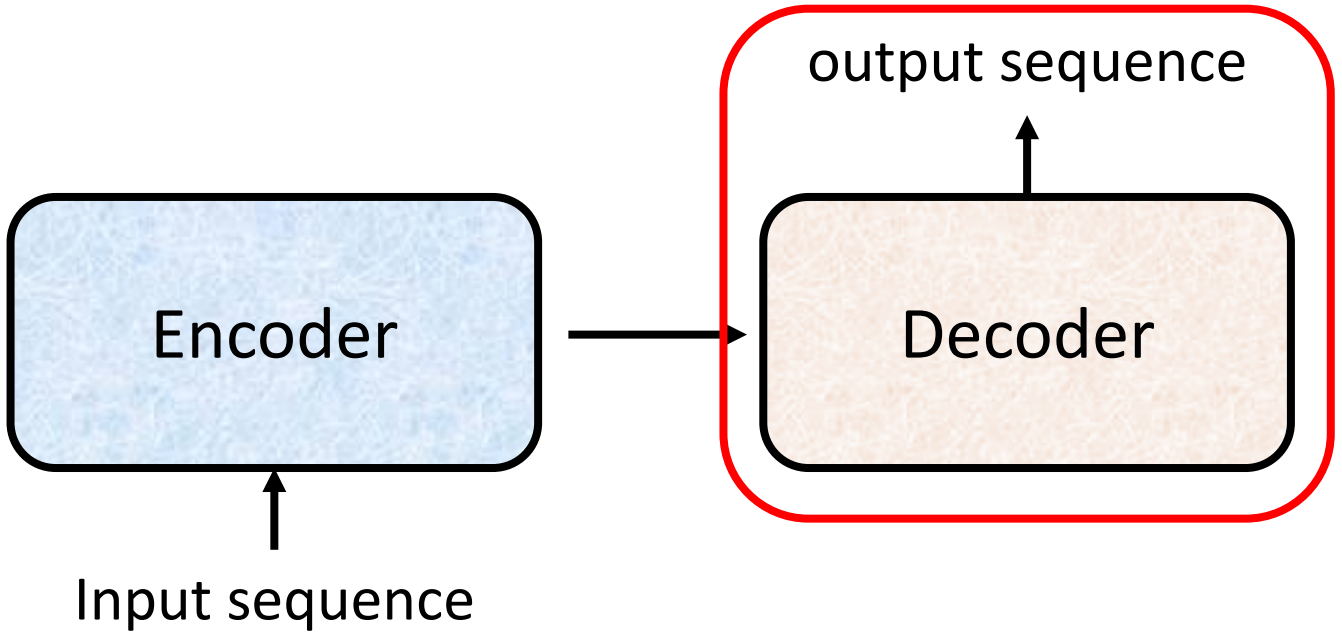
LayerNorm: 单个样本所有特征维度归一化

BatchNorm: 一个batch中同一通道的所有样本归一化（CNN）

Encoder

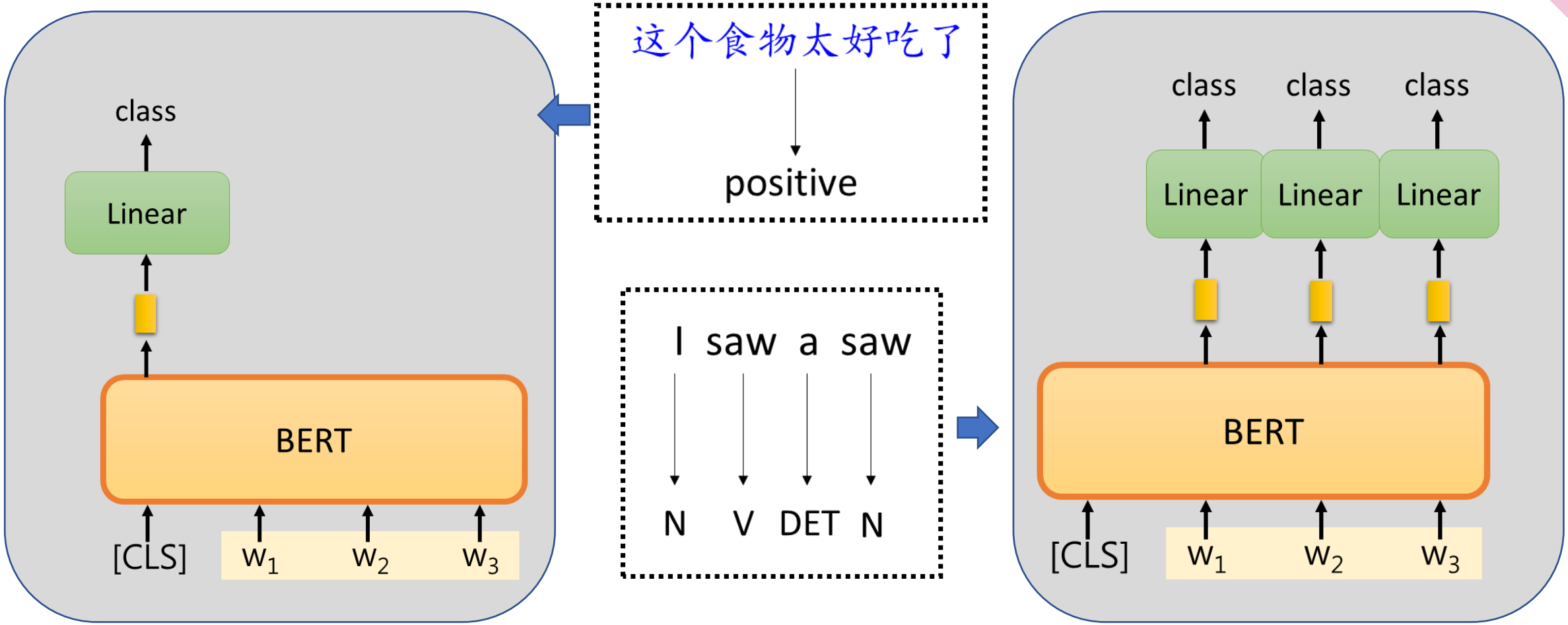

h^1
h^2
h^3
h^4
Encoder
x^1
x^2
x^3
x^4
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Positional Encoding
Input Embedding
Inputs
Residual+ Layer norm

Decoder



Autoregressive (AT)

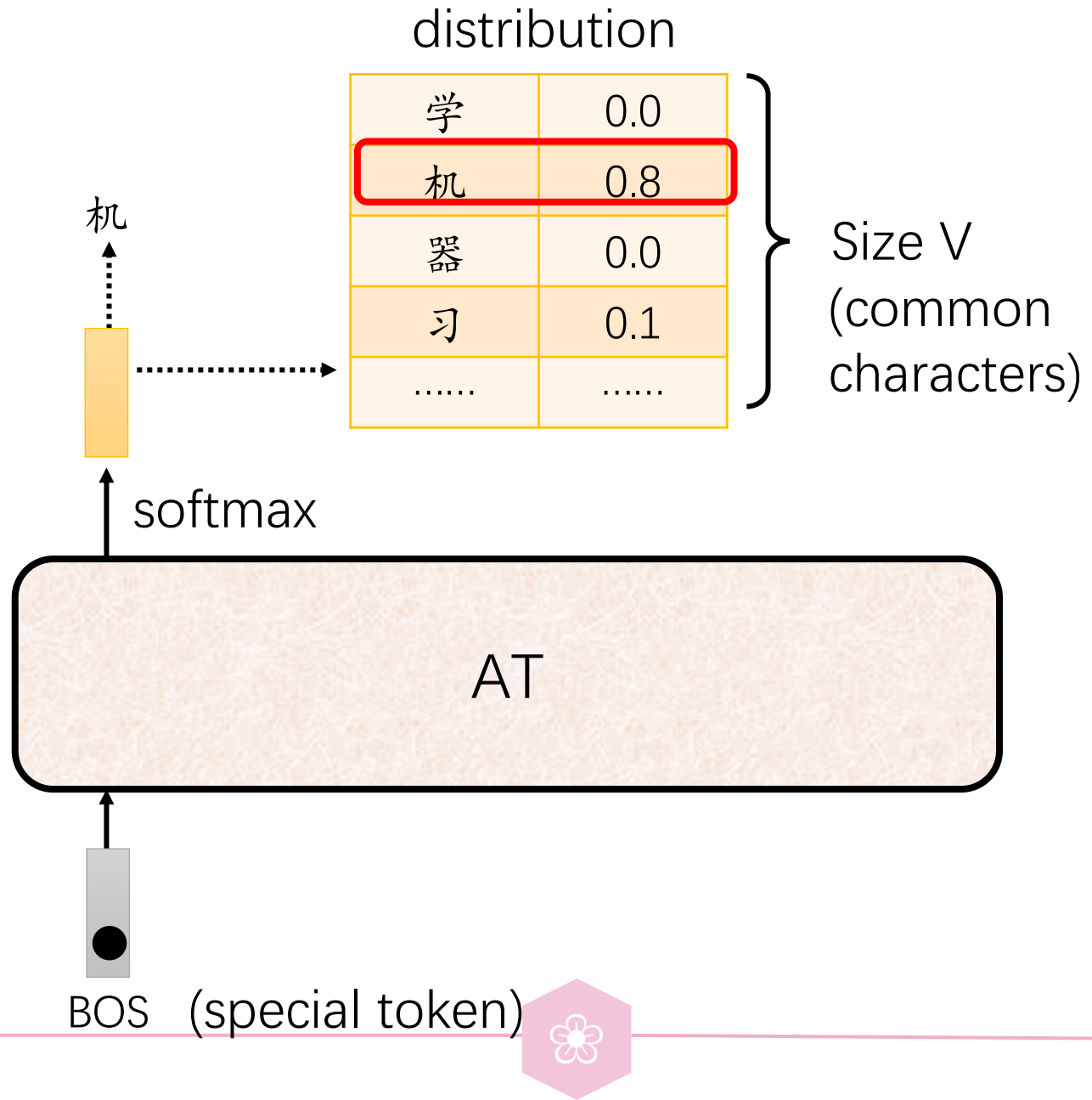
Bert


class
Linear
BERT
[CLS]
w_1
w_2
w_3
这个食物太好吃了
positive
I saw a saw
N V DET N
class
class
class
Linear
Linear
Linear
BERT
[CLS]
w_1
w_2
w_3

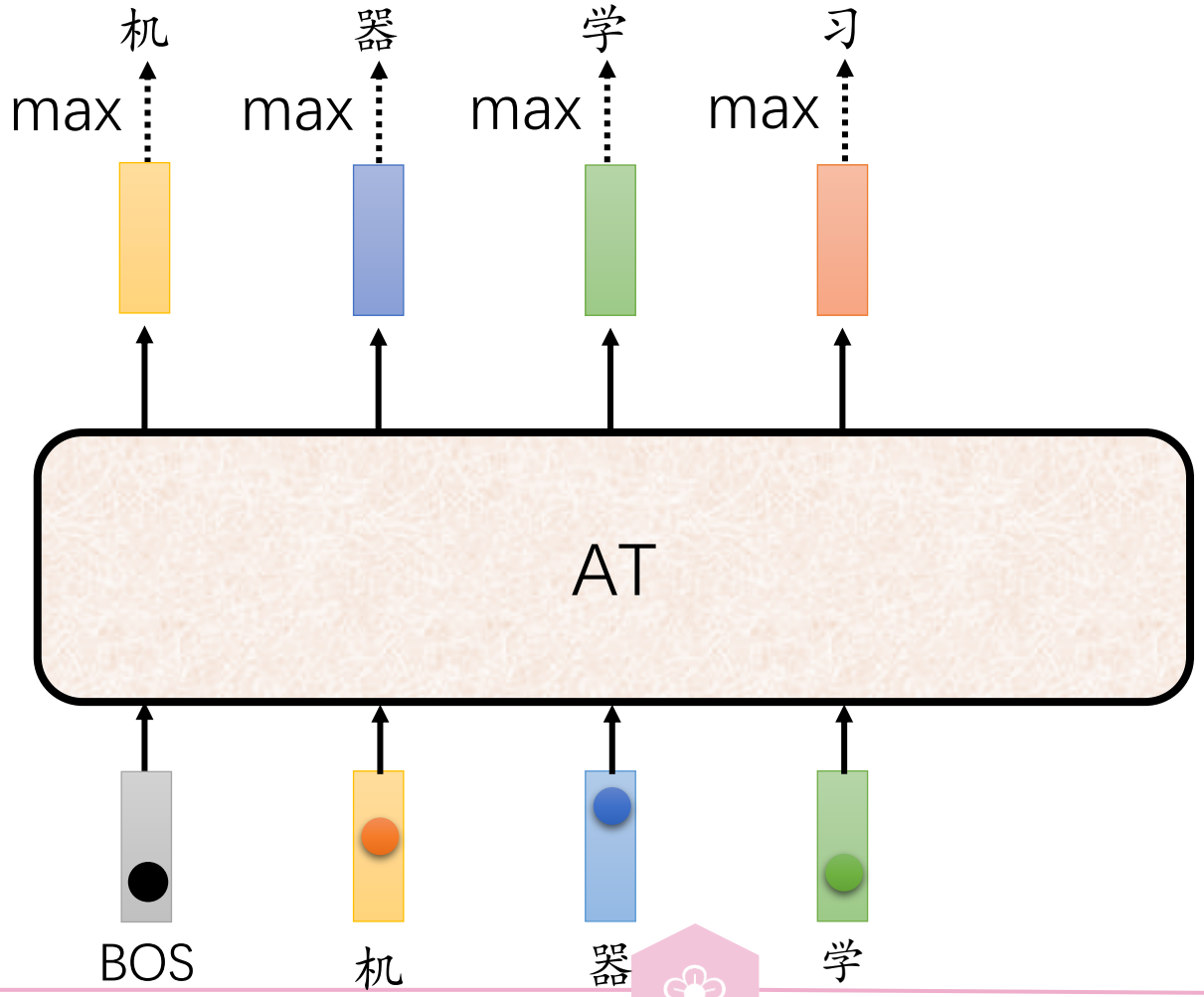
自回归Autoregressive

- 将预测对象按照时间顺序排列起来，构成一个所谓的时间序列
  - 从过去预测未来，依次生成每一个输出值
- 目标函数：$-log\sum_{t=1}^{T}p(y_t|y_{<t},x)$

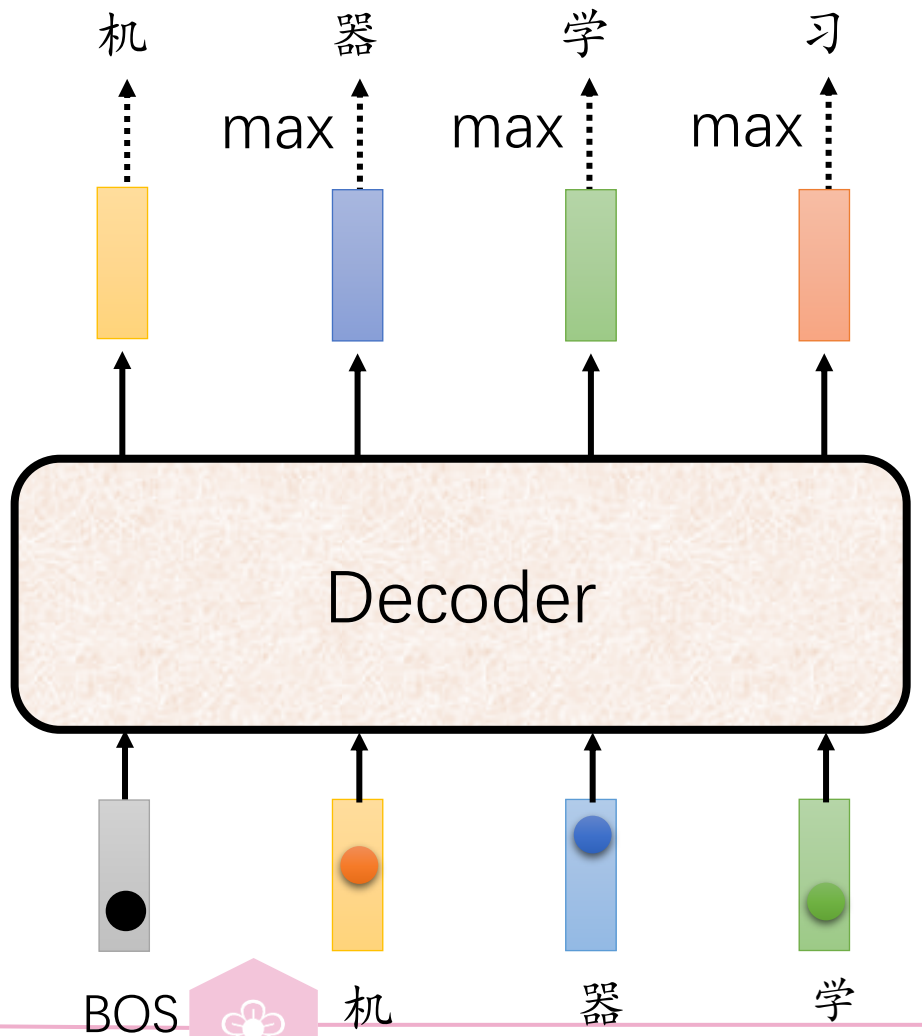
AT

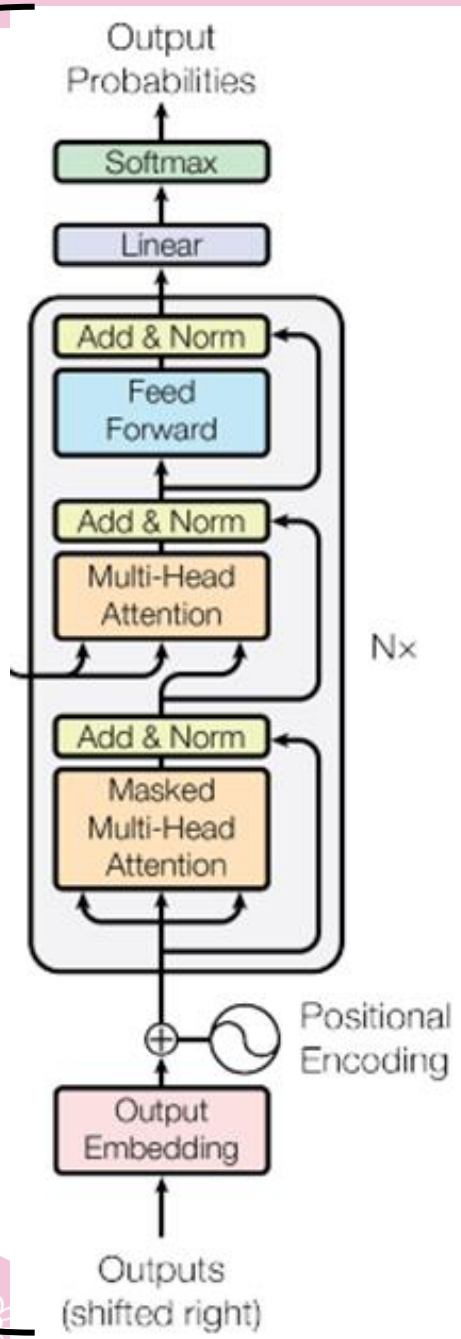

distribution
学 0.0
机 0.8
器 0.0
习 0.1
....
Size V
(common
characters)
机
softmax
AT
BOS
(special token)

AT



Decoder

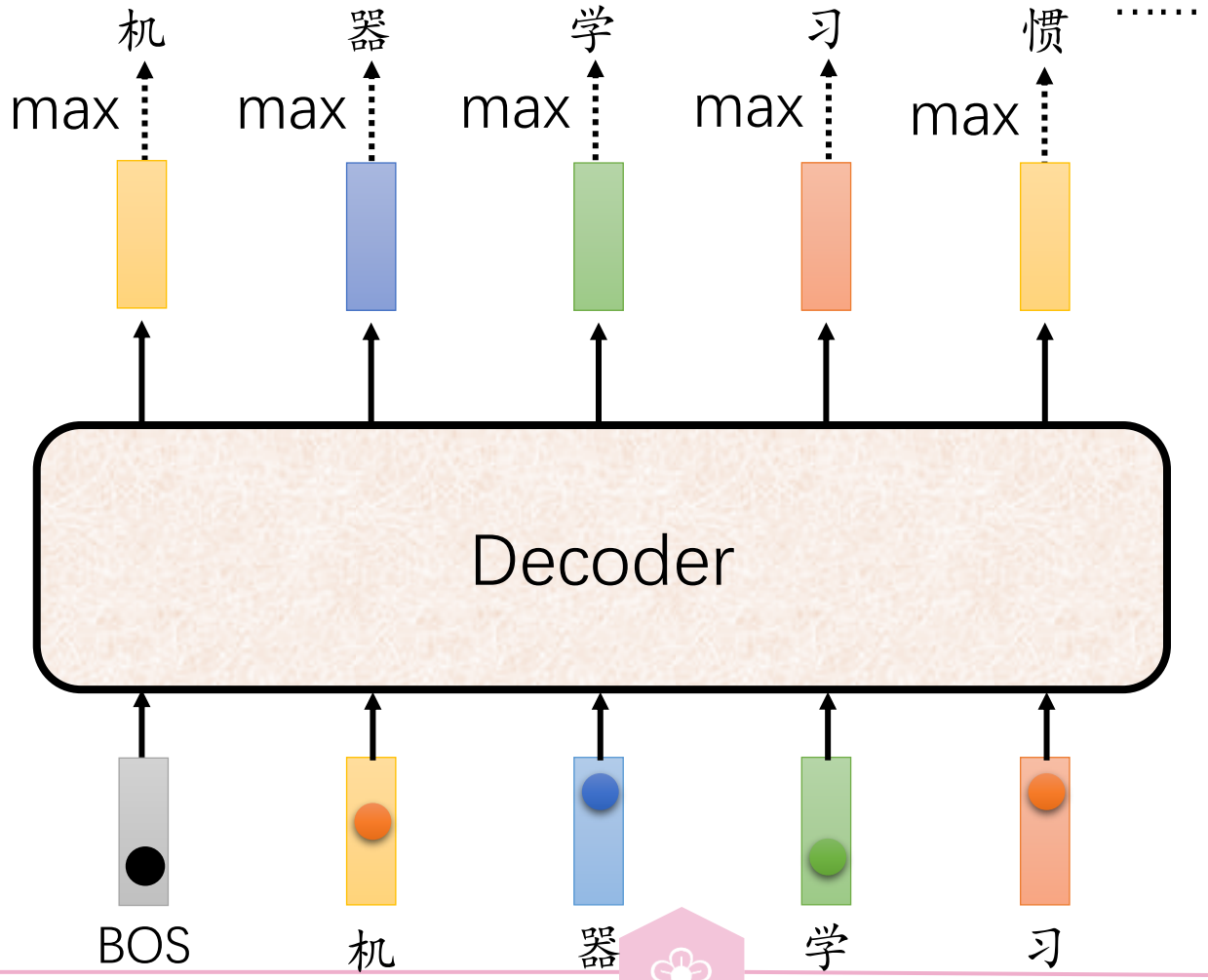

机
器
学
习
max
max
max
Decoder
BOS
机
器
学


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Positional Encoding
Output Embedding
Outputs (shifted right)

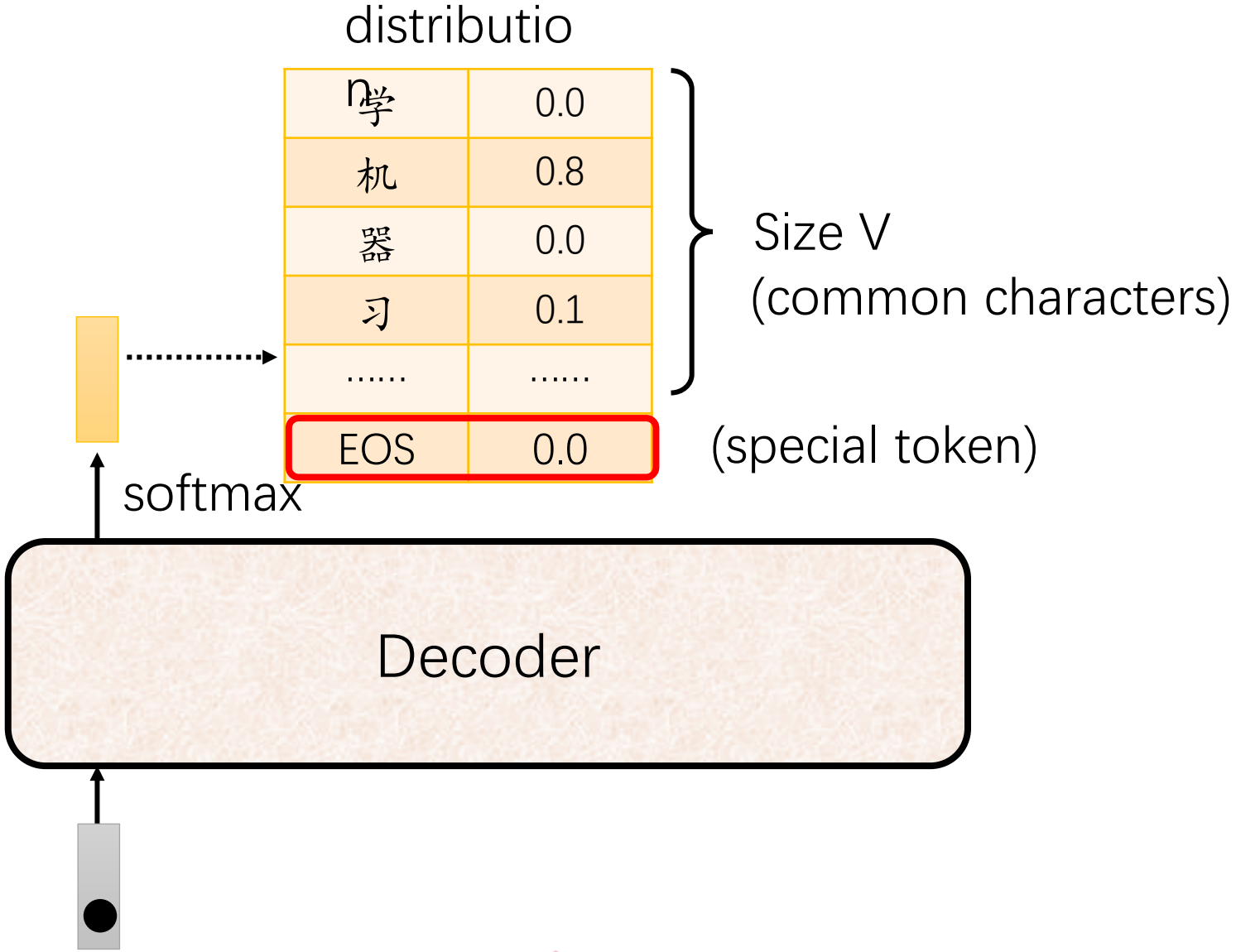
Decoder

由于真实的输出序列长度未知.

Never stop!

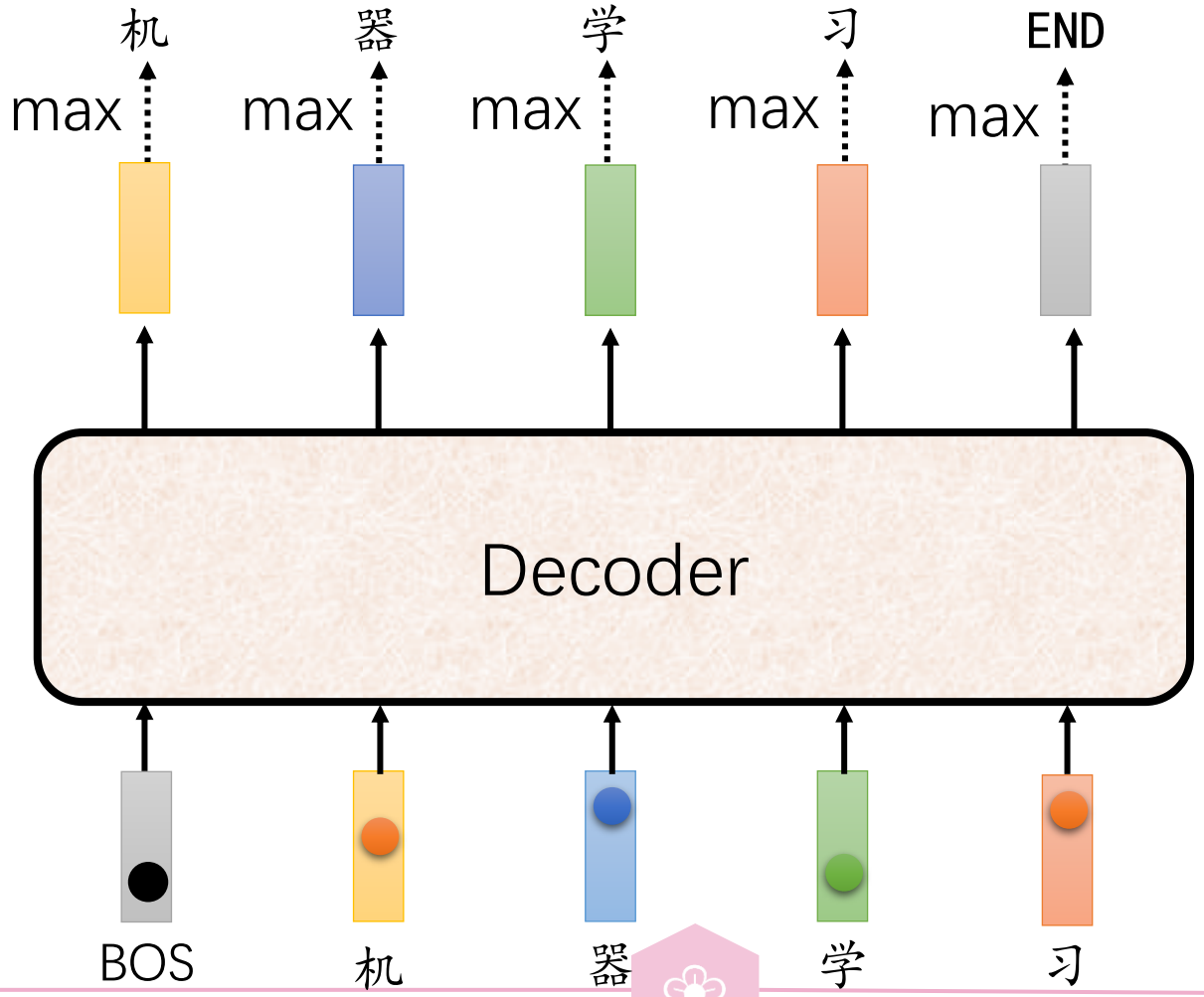


Decoder

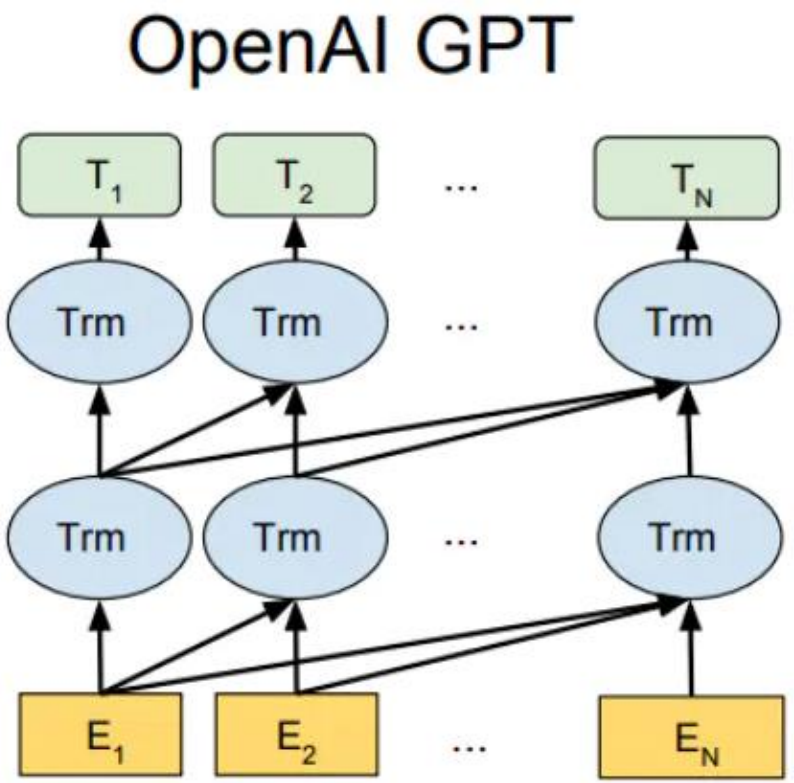

distributio
学 0.0
机 0.8
器 0.0
习 0.1
....
EOS 0.0
Size V
(common characters)
(special token)
softmax
Decoder

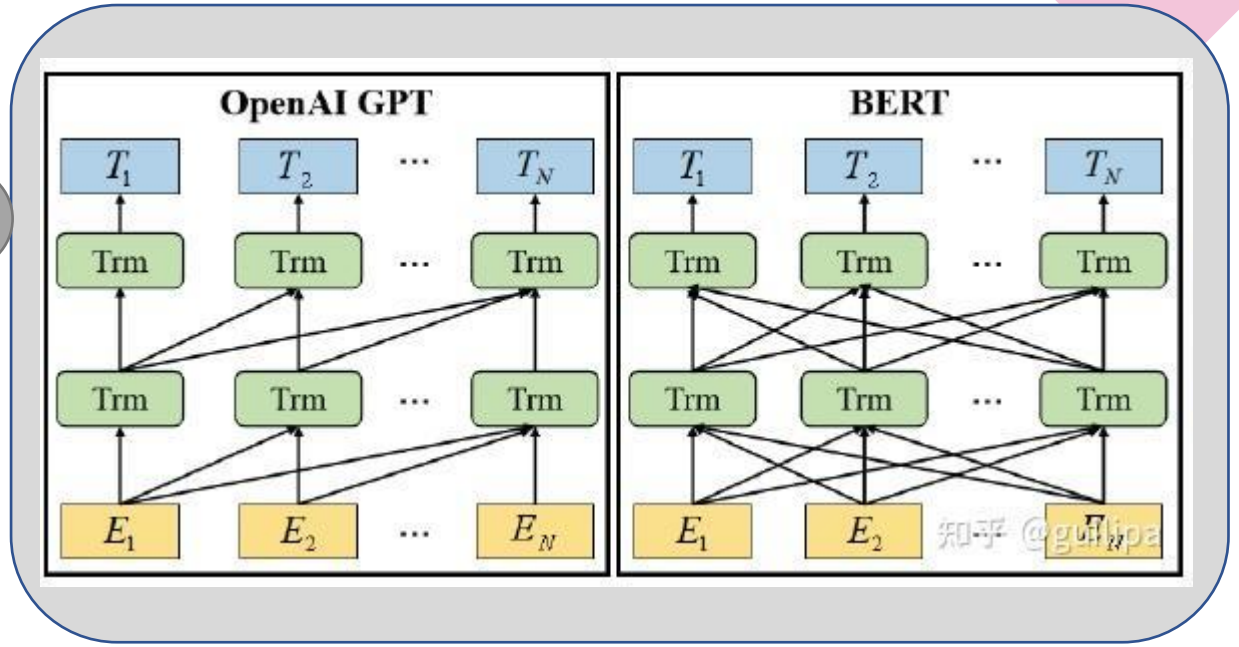
Decoder

Stop at here!

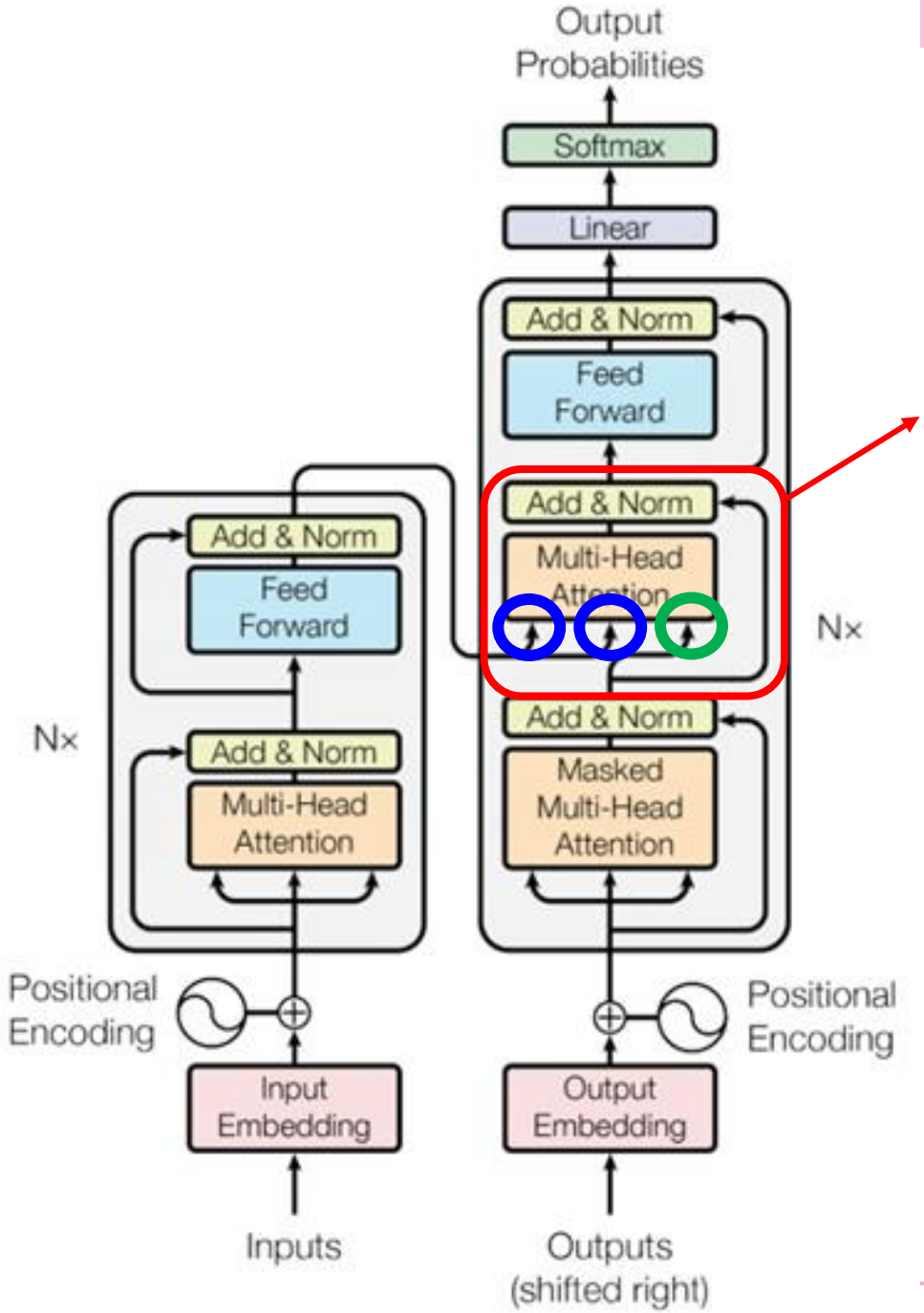


GPT


OpenAI GPT
T1
T2
...
TN
Trm
Trm
...
Trm
Trm
Trm
...
Trm
E1
E2
...
EN

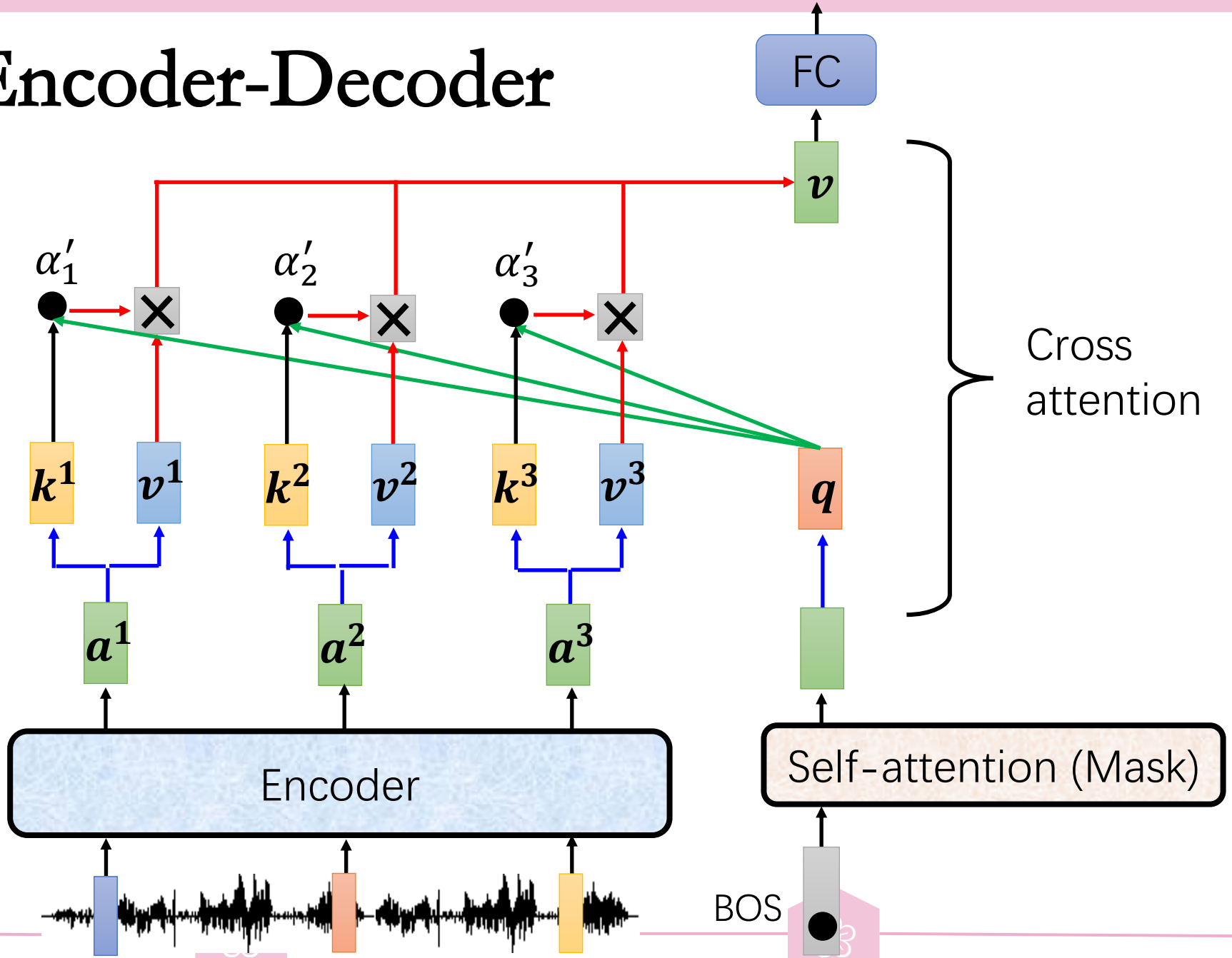

OpenAI GPT
BERT
T1
T2
...
TN
Trm
E1
E2
EN

Encoder-Decc



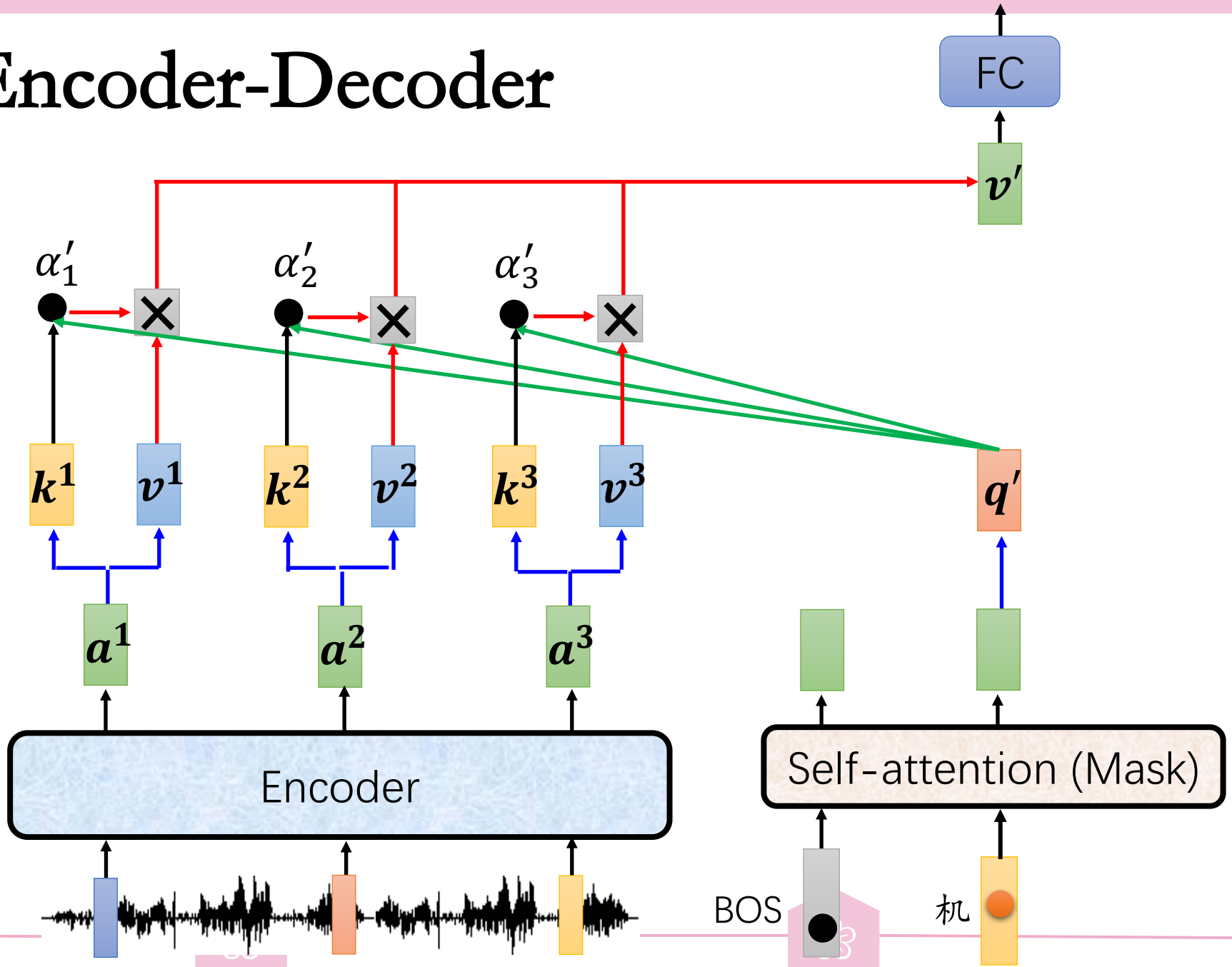
Cross attention

Encoder-Decoder

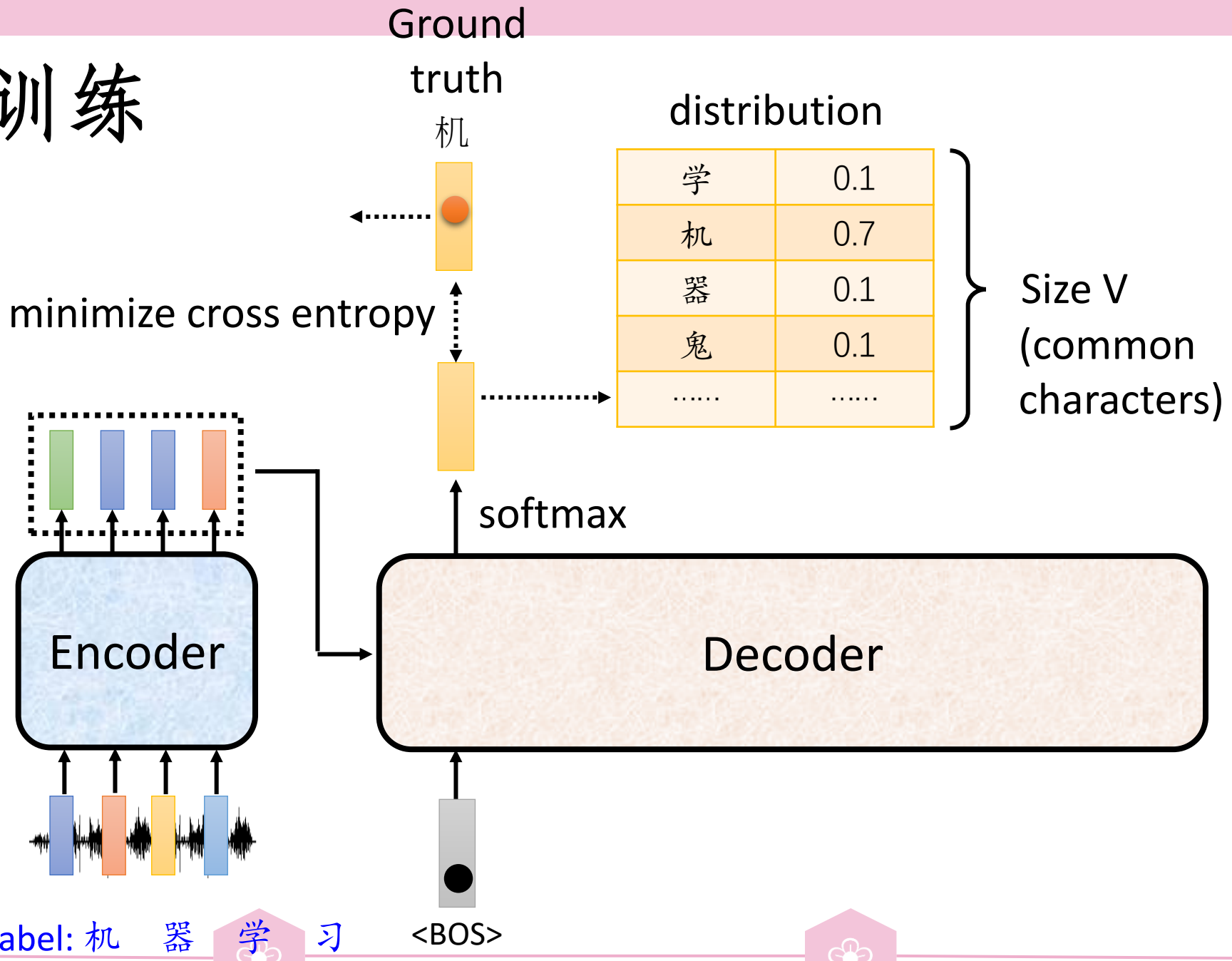


Cross attention

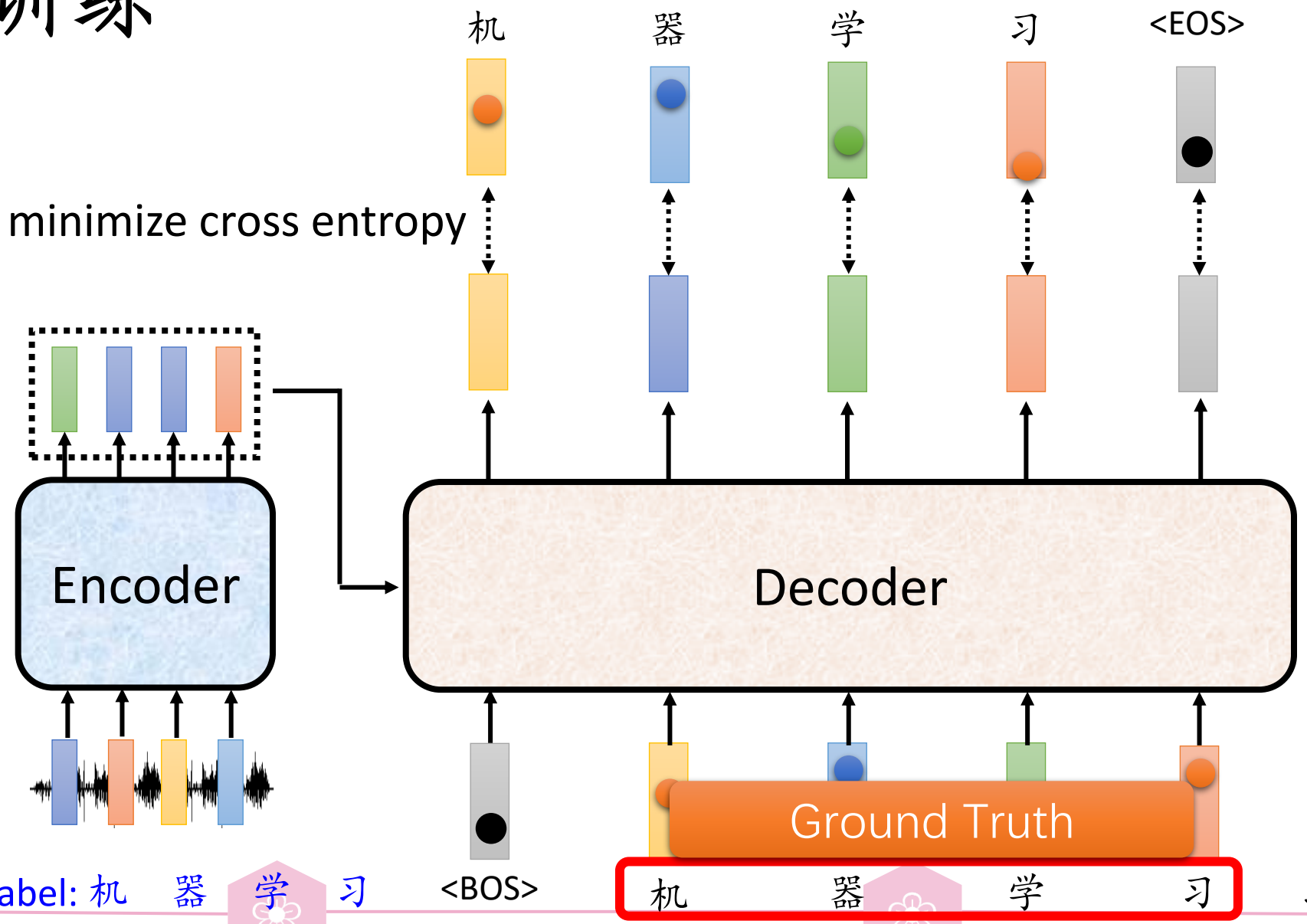
Encoder-Decoder



训练


Ground truth
机
distribution
学 0.1
机 0.7
器 0.1
鬼 0.1
…… ……
Size V (common characters)
minimize cross entropy
softmax
Encoder
Decoder
<BOS>
Label: 机 器 学 习

训练

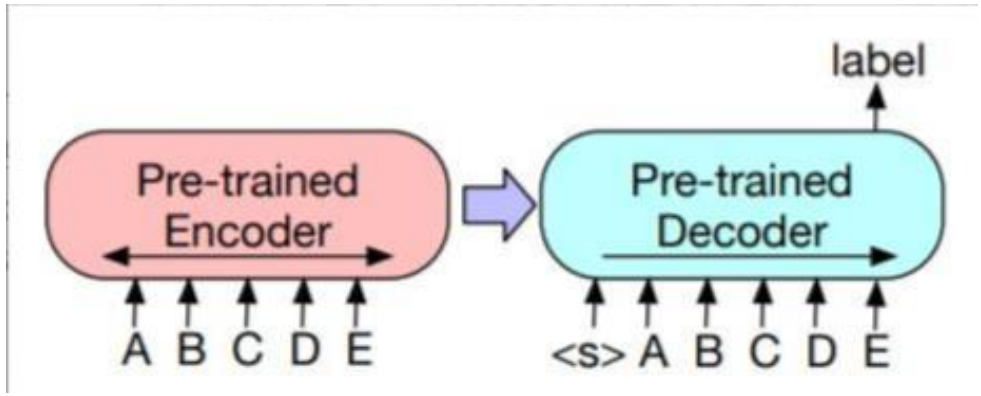


Label: 机 器 学 习

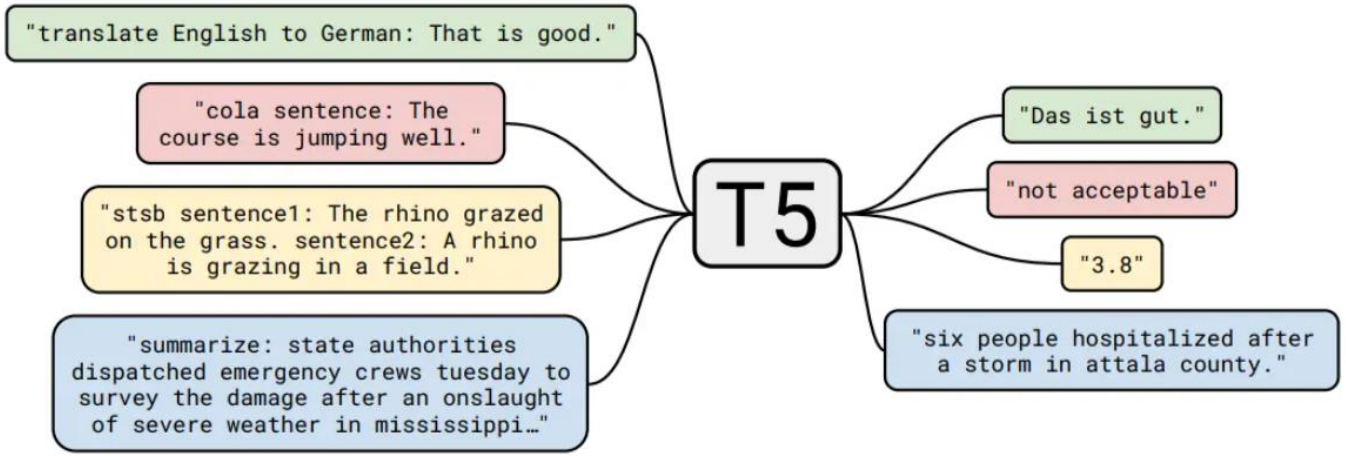
真实输入.

Encoder-decoder架构

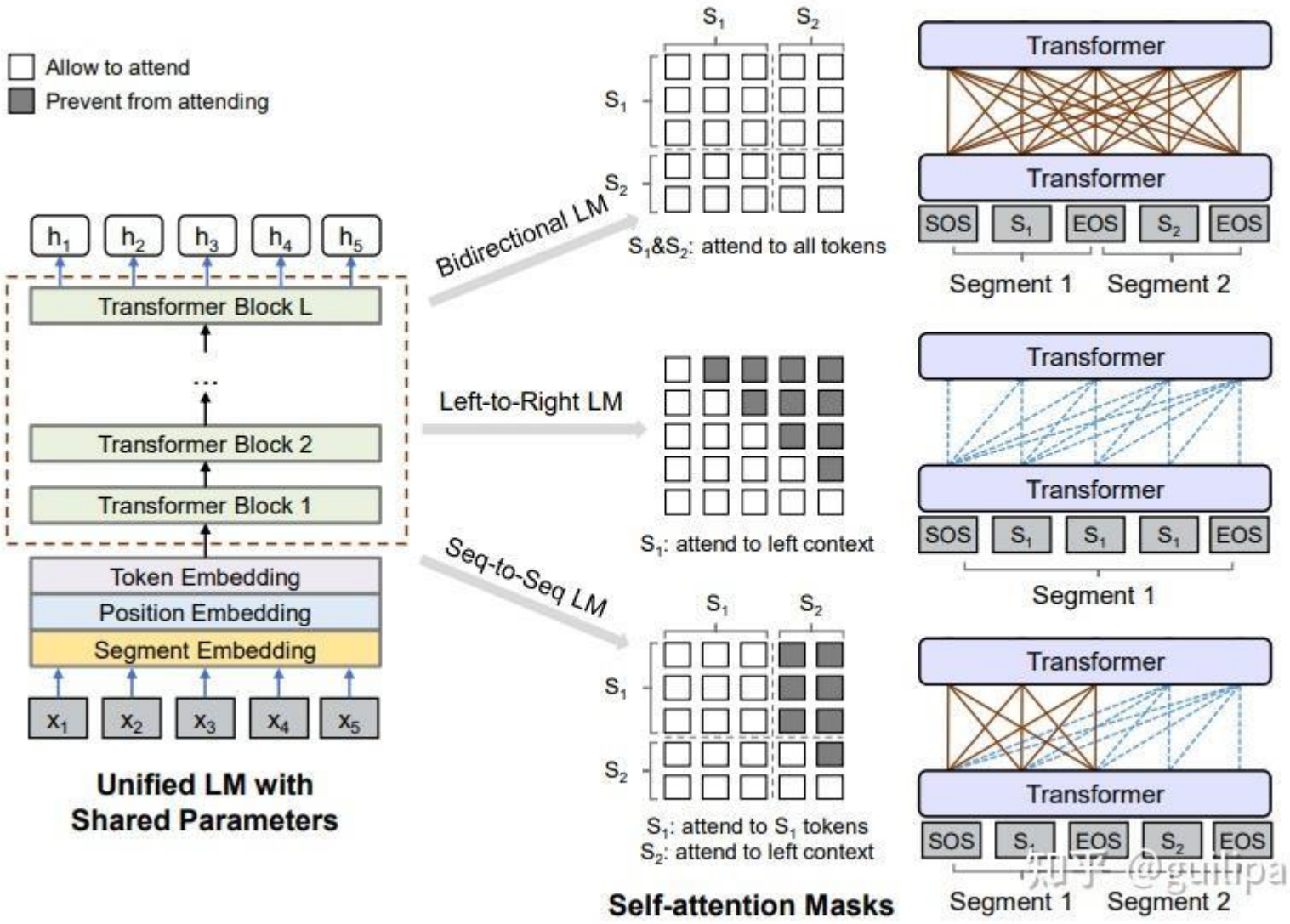
BART



T5



UniLM


Allow to attend
Prevent from attending
h1 h2 h3 h4 h5
Transformer Block L
...
Transformer Block 2
Transformer Block 1
Token Embedding
Position Embedding
Segment Embedding
x1 x2 x3 x4 x5
Unified LM with Shared Parameters
Bidirectional LM
Left-to-Right LM
Seq-to-Seq LM
S1 S2
S1&S2: attend to all tokens
S1: attend to left context
S1: attend to S1 tokens
S2: attend to left context
Self-attention Masks
Transformer
Transformer
SOS S1 EOS S2 EOS
Segment 1 Segment 2
Transformer
Transformer
SOS S1 S1 S1 EOS
Segment 1
Transformer
Transformer
SOS S1 EOS S2 EOS
Segment 1 Segment 2