北京大学
PEKING UNIVERSITY

Transformer

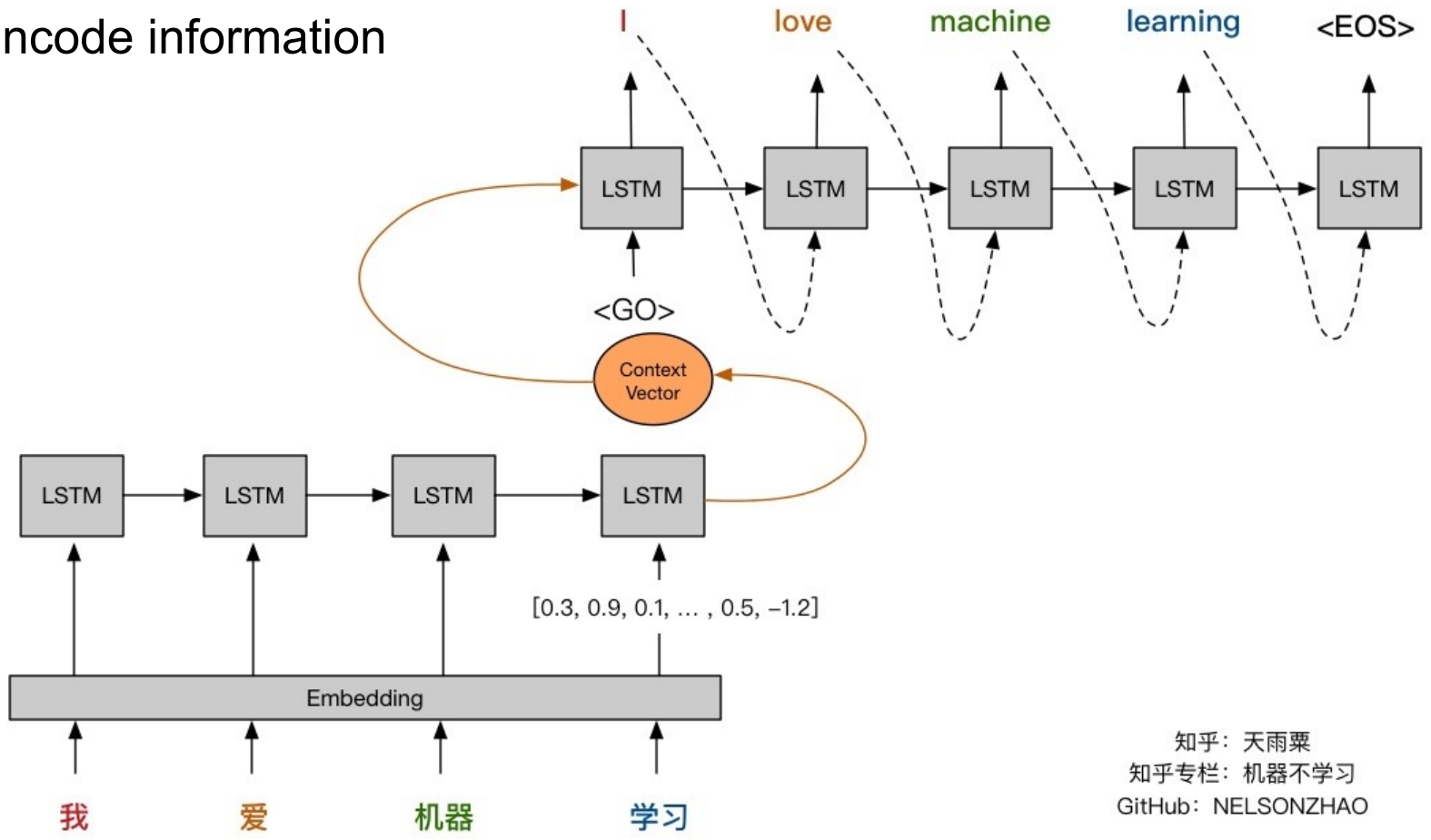
Kun Yuan

Center for Machine Learning Research @ Peking University

Oct. 24, 2023

Traditional seq2seq model

RNN is hard to encode information that is far away



知乎：天雨粟
知乎专栏：机器不学习
GitHub：NELSONZHAO

Attention

How to capture the most valuable information from a pool of candidate?

Consider a pool of candidate information D = (k1,v1),(k2,v2),…,(km,vm)

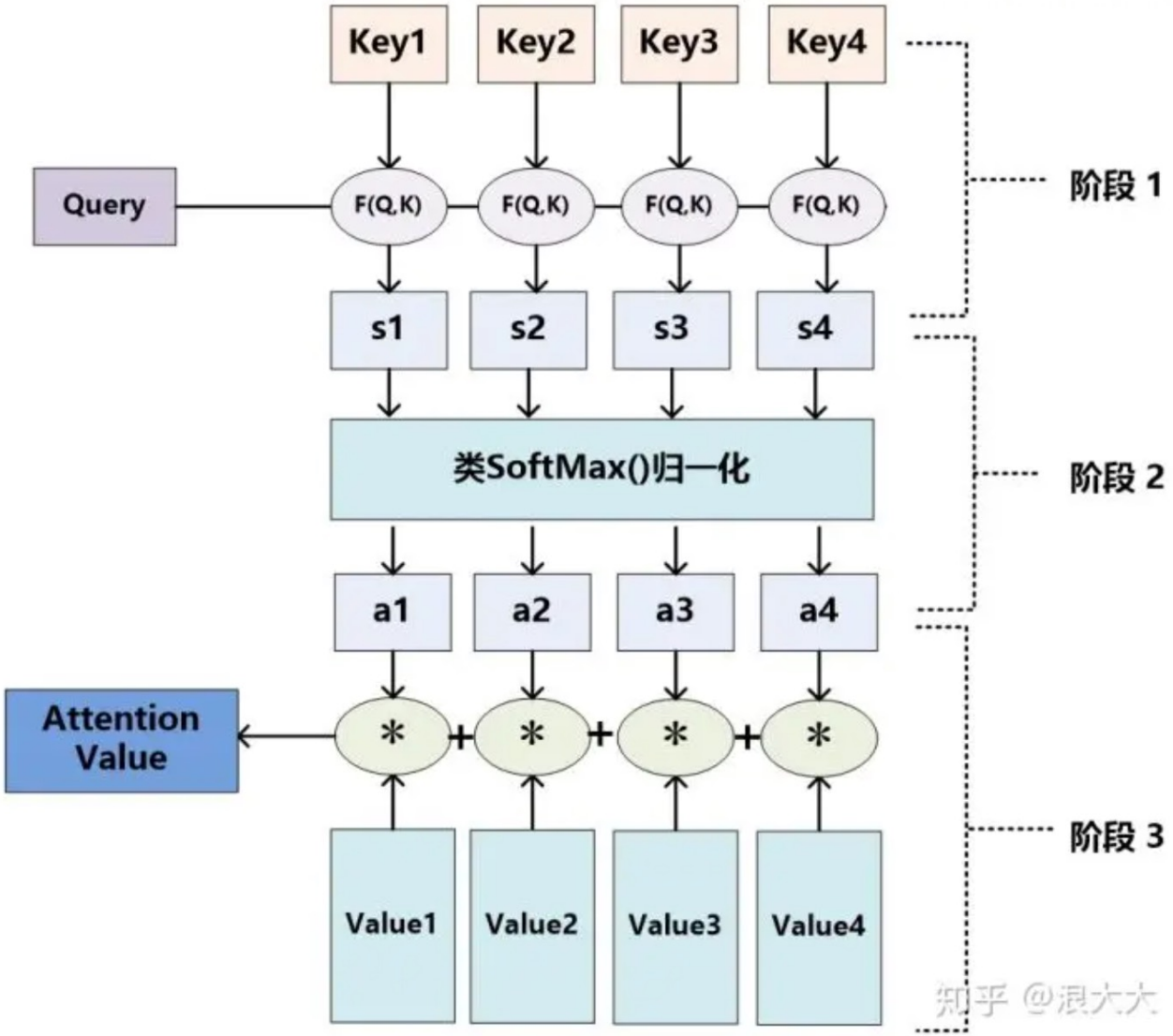
Given a query q, we can capture the most valuable information as follows

$$\text{Attention}(\mathbf{q}, \mathcal{D}) \stackrel{\text{def}}{=} \sum_{i=1}^{m} \alpha(\mathbf{q}, \mathbf{k}_i)\mathbf{v}_i,$$

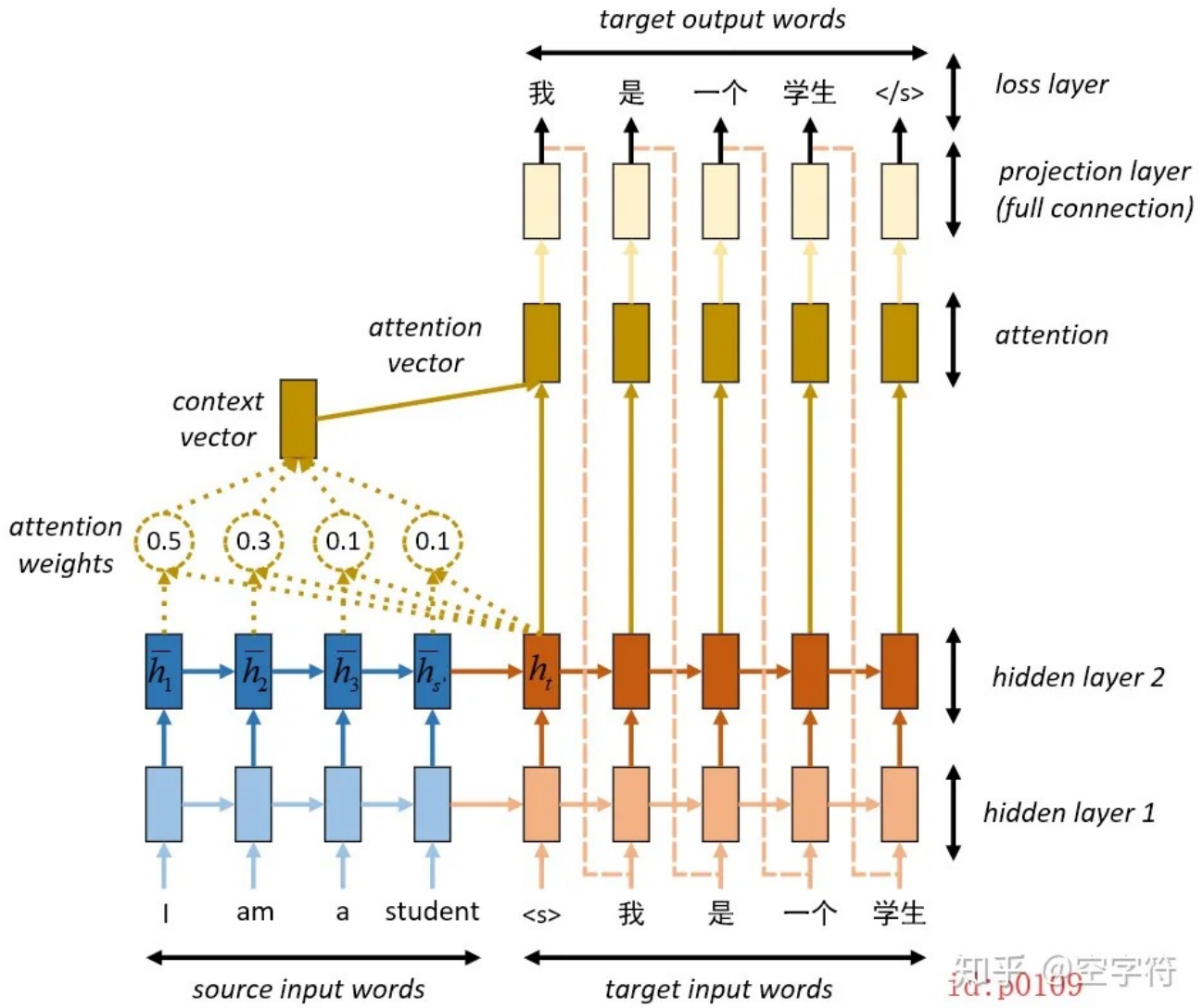
where weight α is to evaluate how close the query q is to key ki

$$\alpha(\mathbf{q}, \mathbf{k}_i) = \text{softmax}(a(\mathbf{q}, \mathbf{k}_i)) = \frac{\exp(\mathbf{q}^\top \mathbf{k}_i/\sqrt{d})}{\sum_{j=1} \exp(\mathbf{q}^\top \mathbf{k}_j/\sqrt{d})}.$$

Attention


Key1
Key2
Key3
Key4
Query
F(Q,K)
F(Q,K)
F(Q,K)
F(Q,K)
s1
s2
s3
s4
类SoftMax()归一化
a1
a2
a3
a4
Attention Value
*
*
*
*
Value1
Value2
Value3
Value4
阶段 1
阶段 2
阶段 3
知乎 @浪大大

Seq2Seq with attention



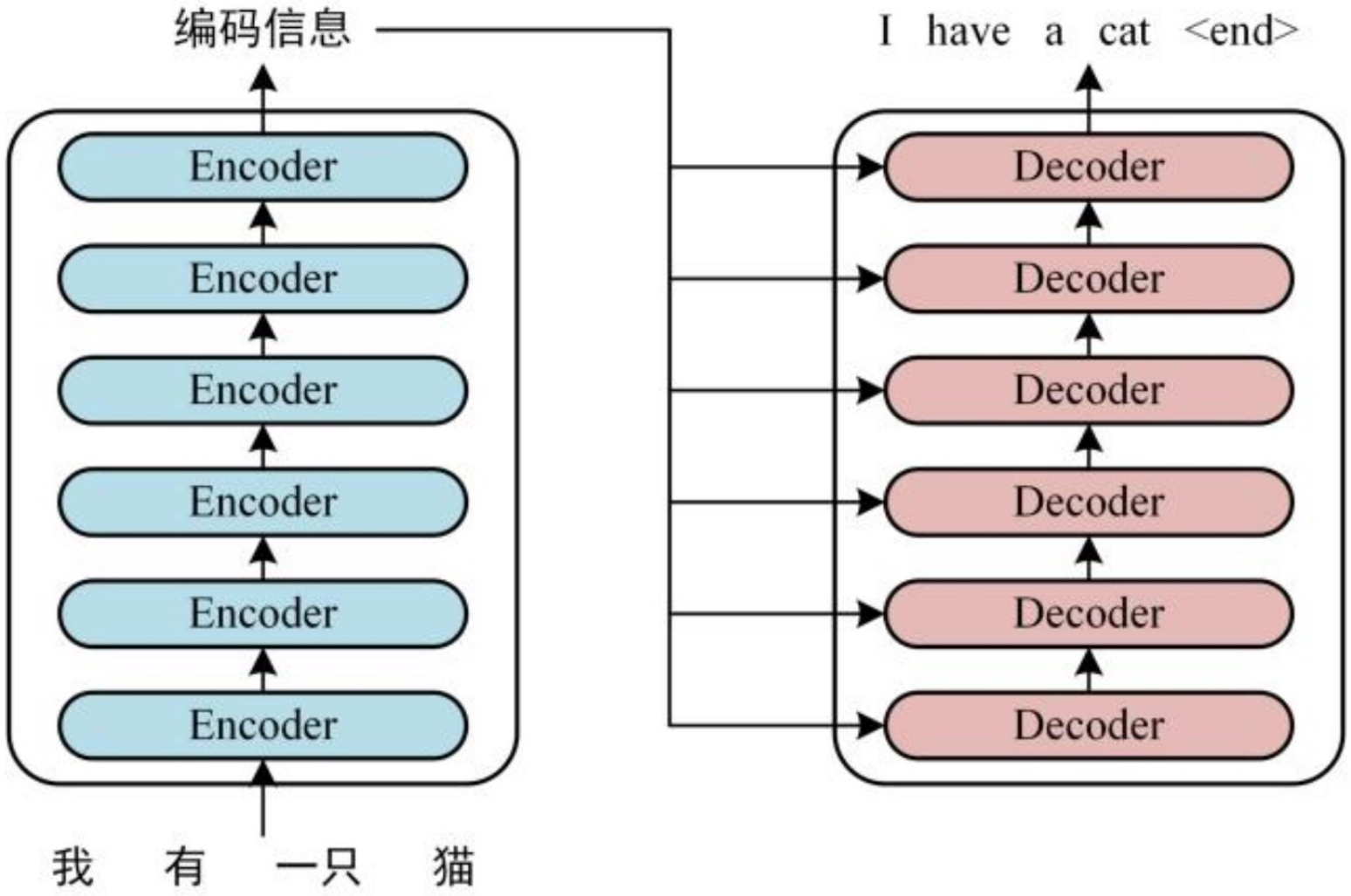
How to get representation h?

One way is RNN

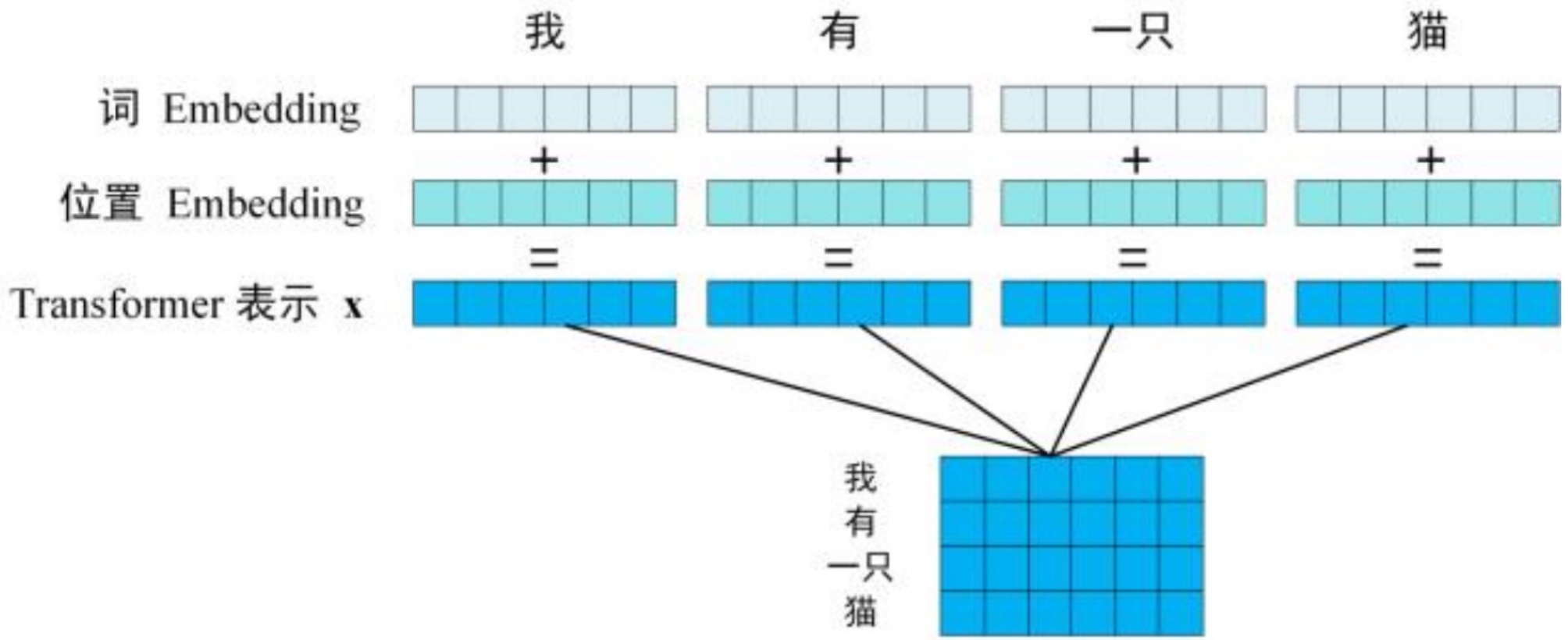
The other is Transformer!

Transformer overview

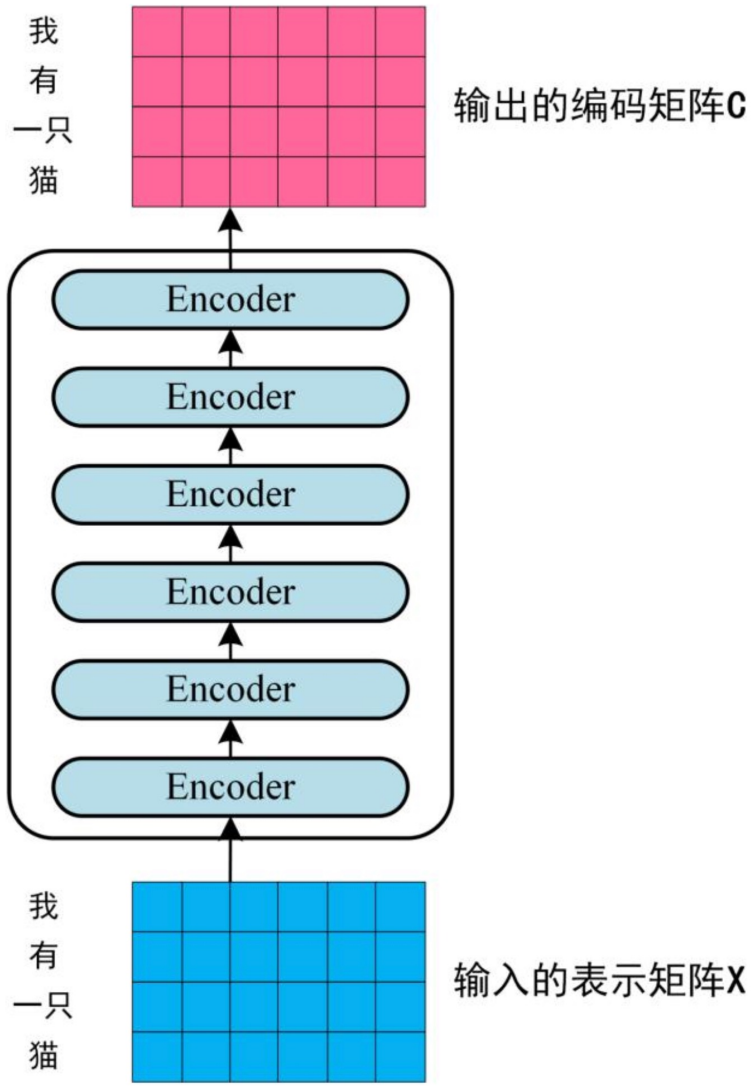
Transformer also utilizes seq2seq



Step 1: represent input information

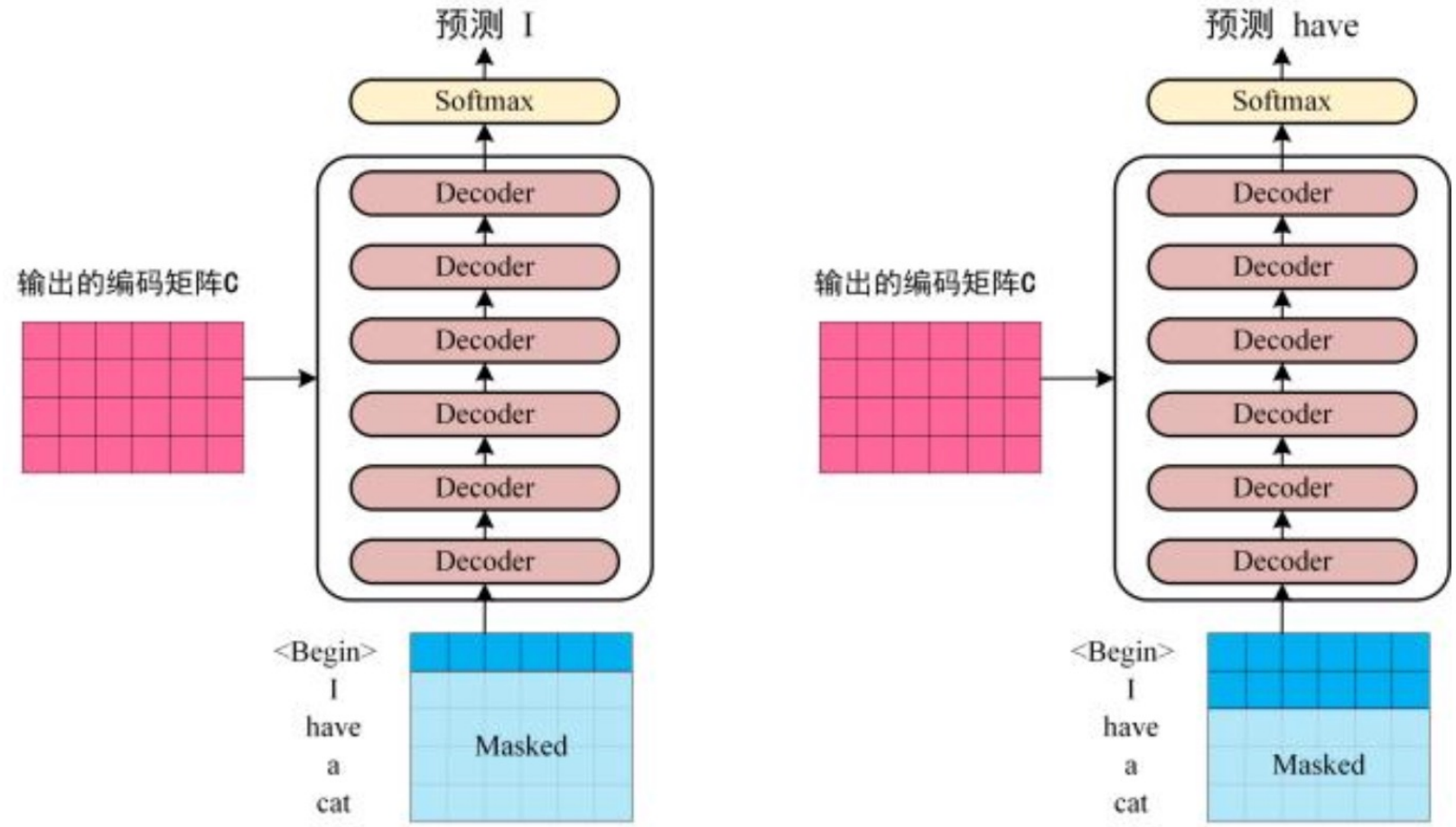


Step 2: Encode input



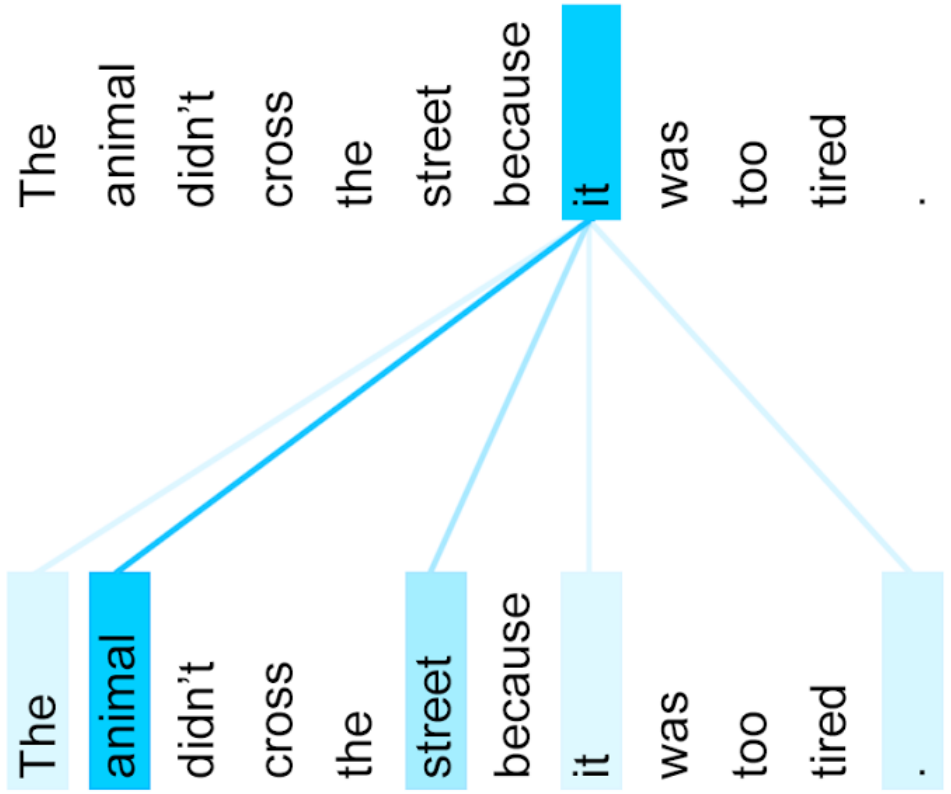
Transformer Encoder 编码句子信息

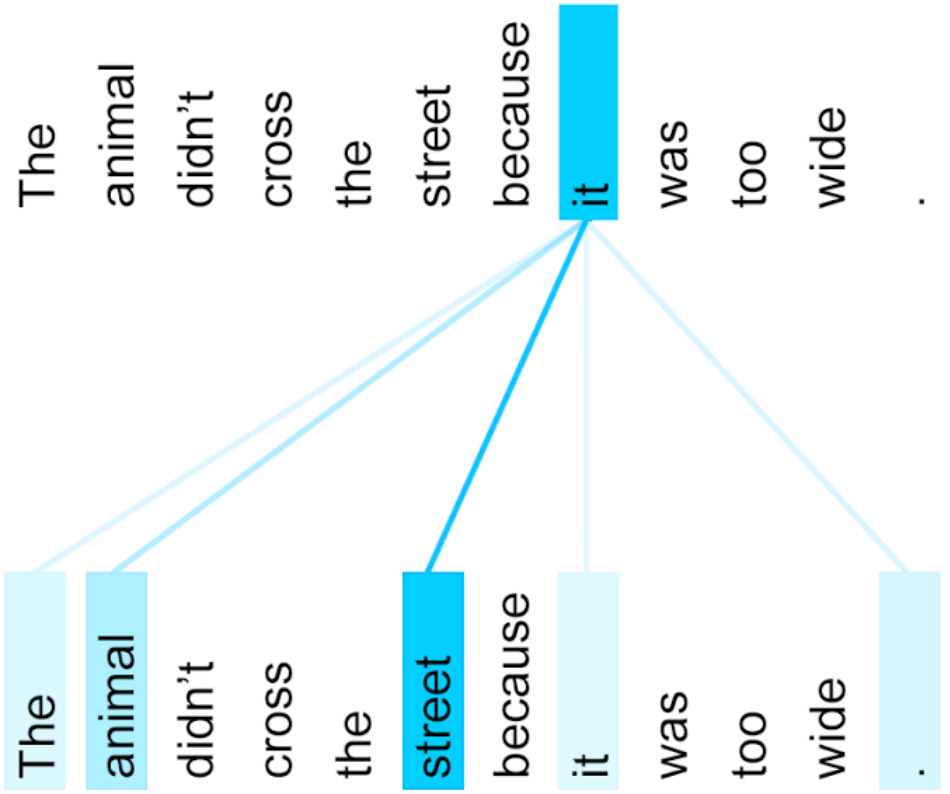
Step 3: Decode and predict



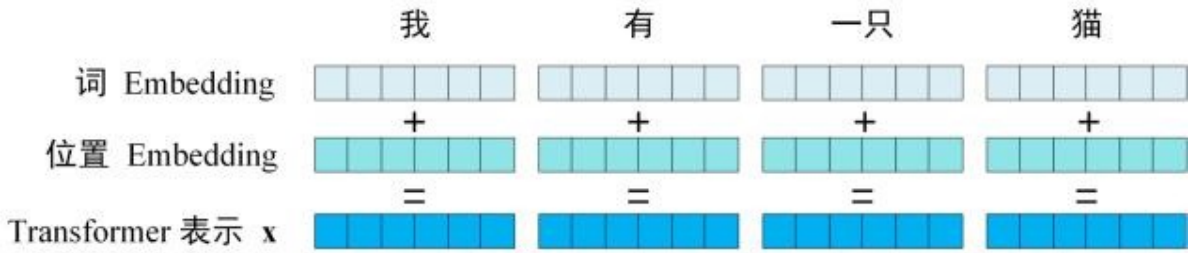
Transofrmer Decoder 预测

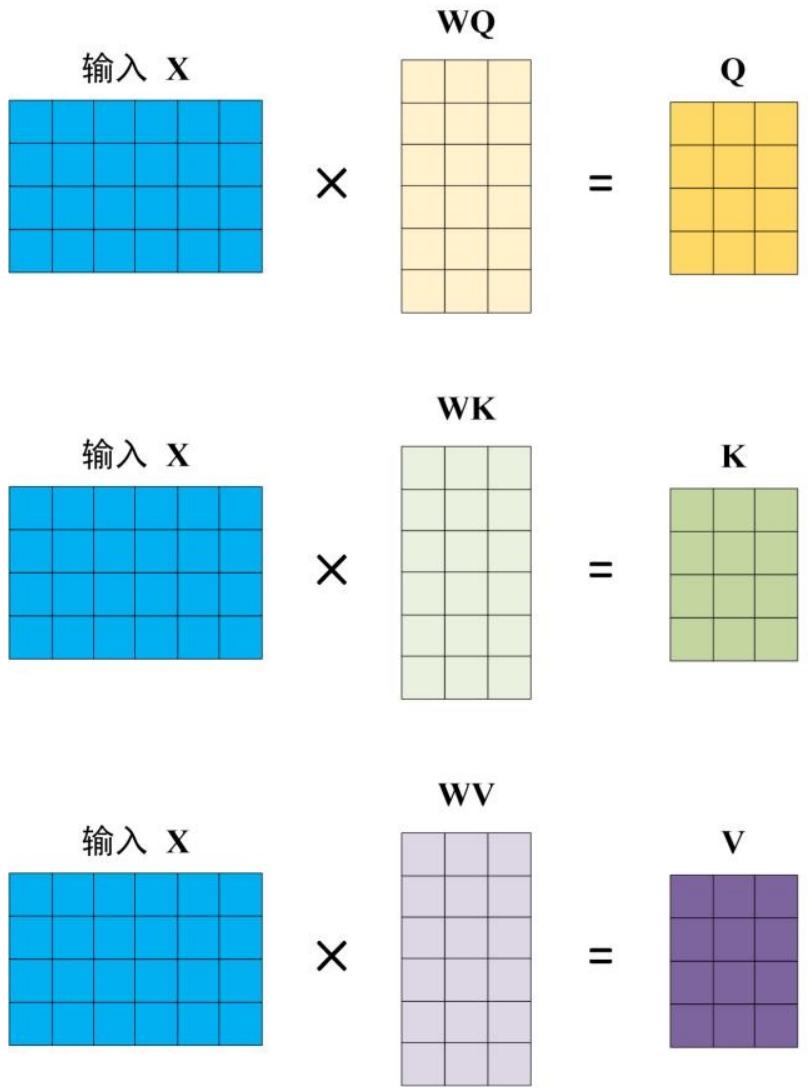
Encode: self-attention





Encode: self-attention

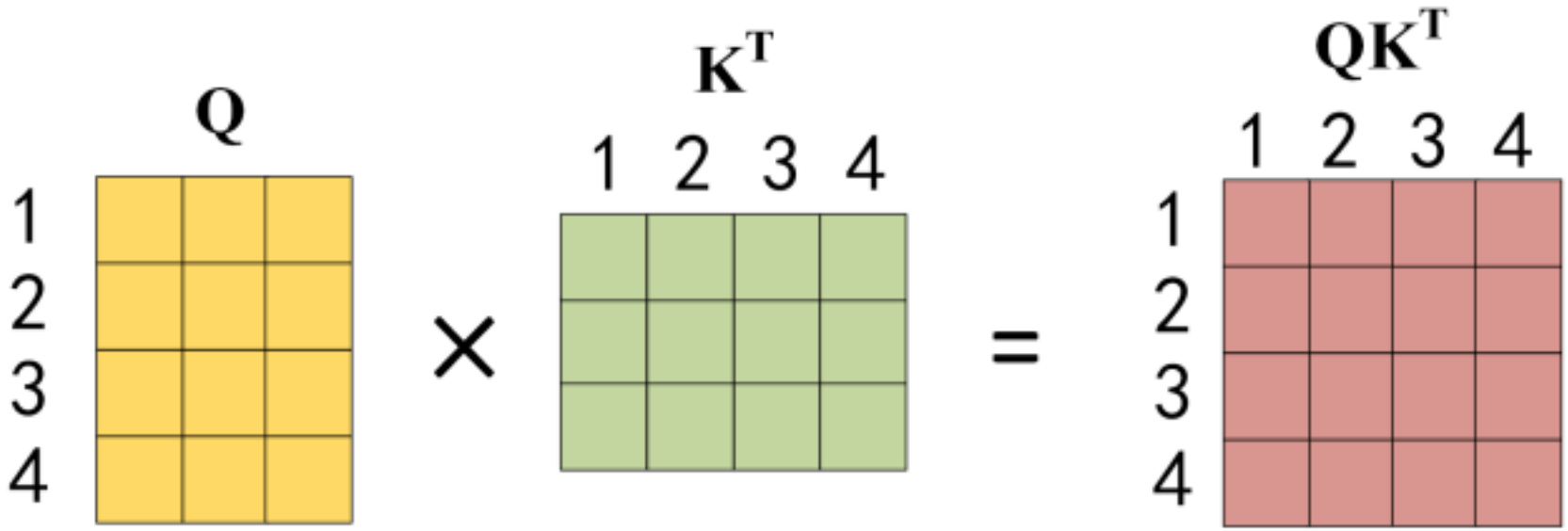




Encode: self-attention

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

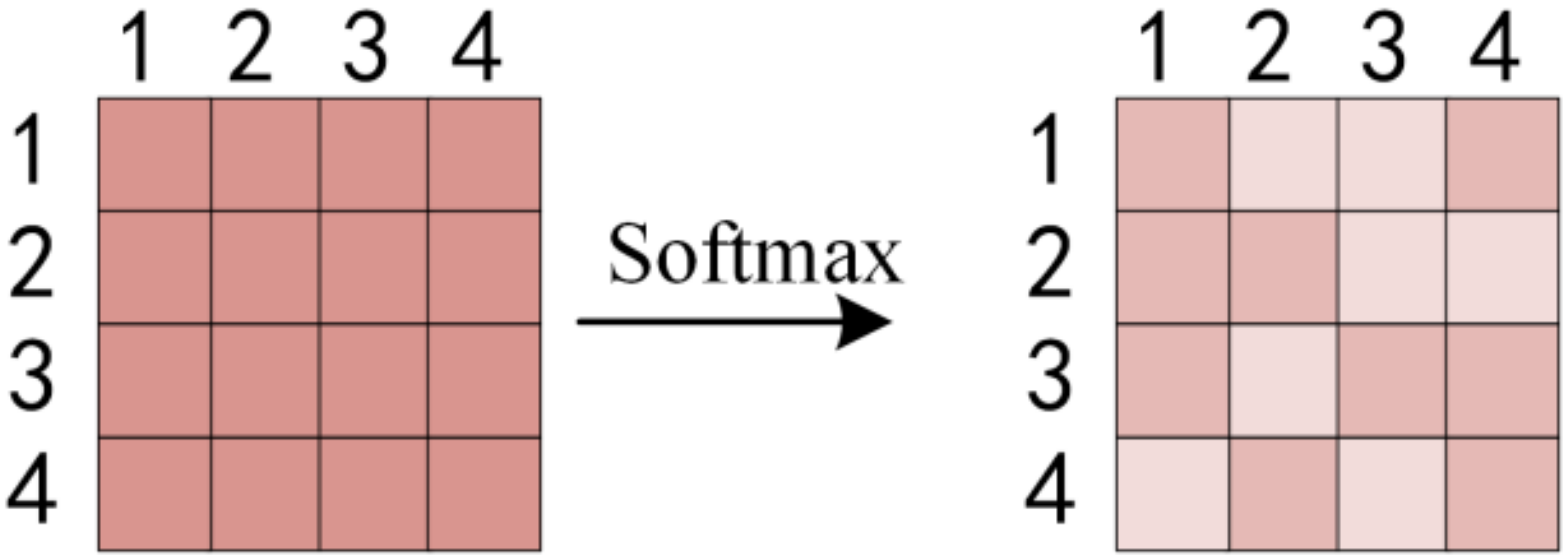
d_k是Q, K矩阵的列数，即向量维度



Encode: self-attention

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

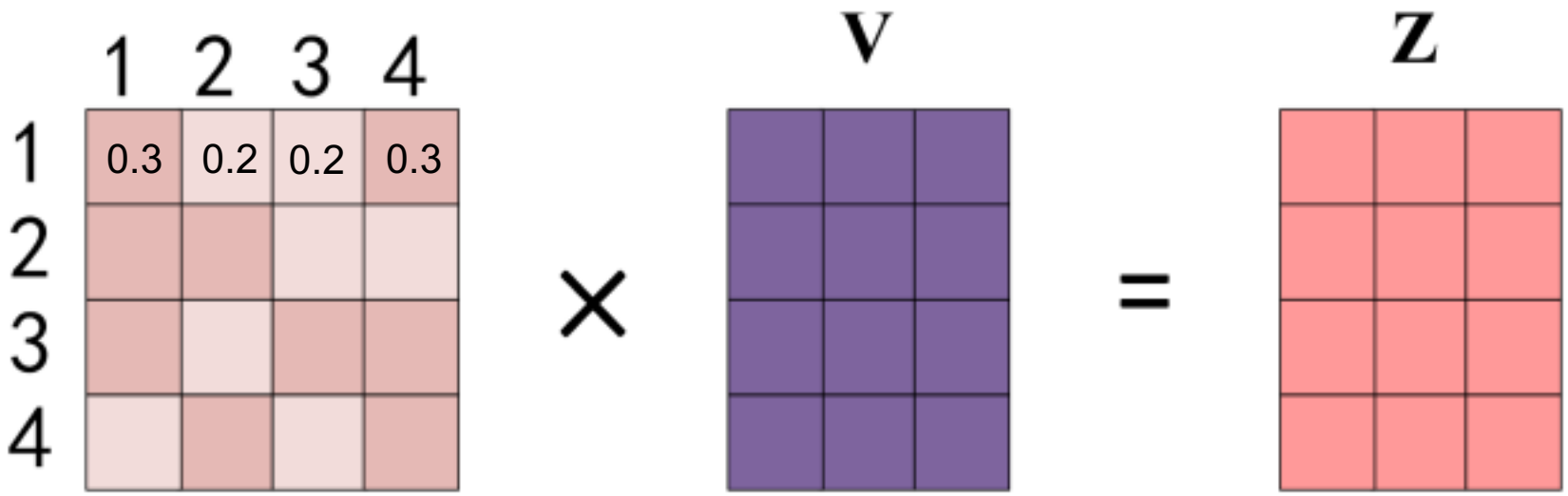
d_k是Q, K矩阵的列数，即向量维度



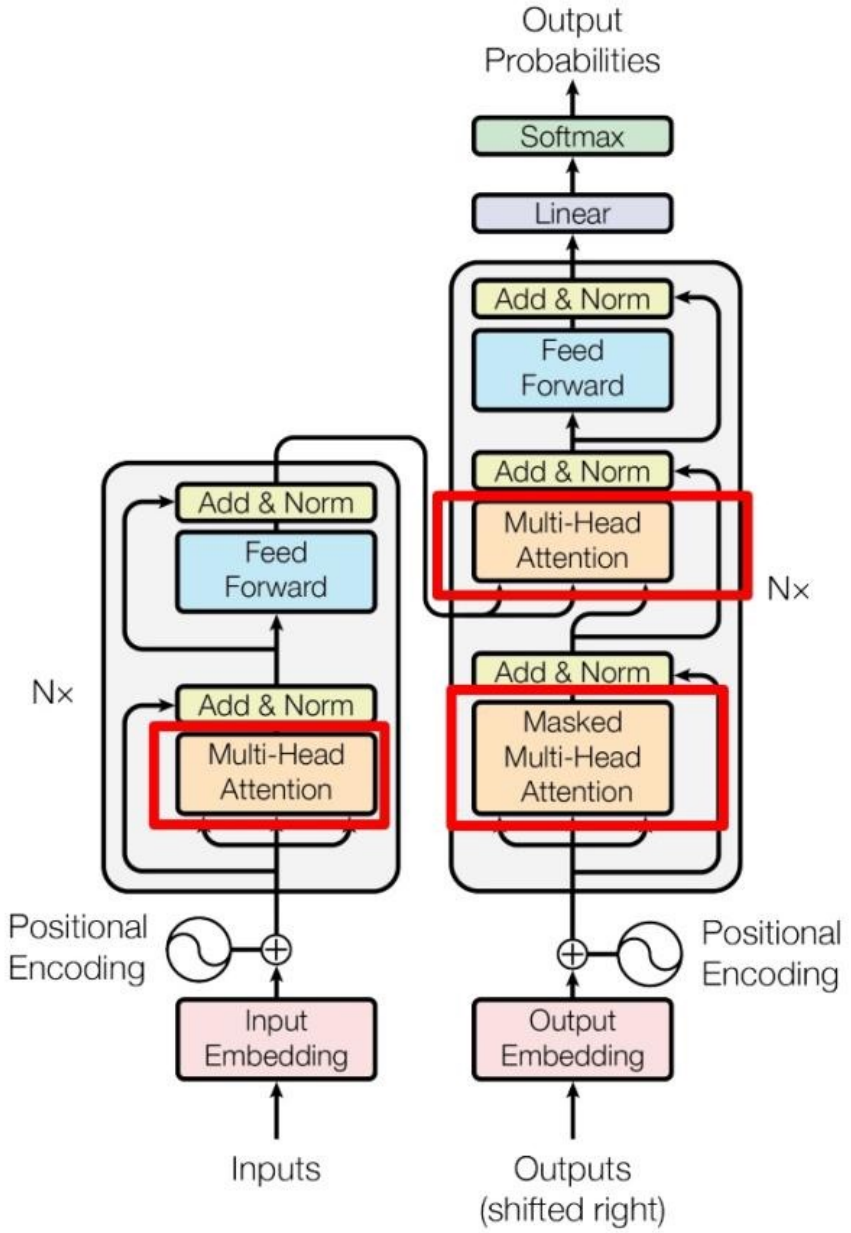
Encode: self-attention

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

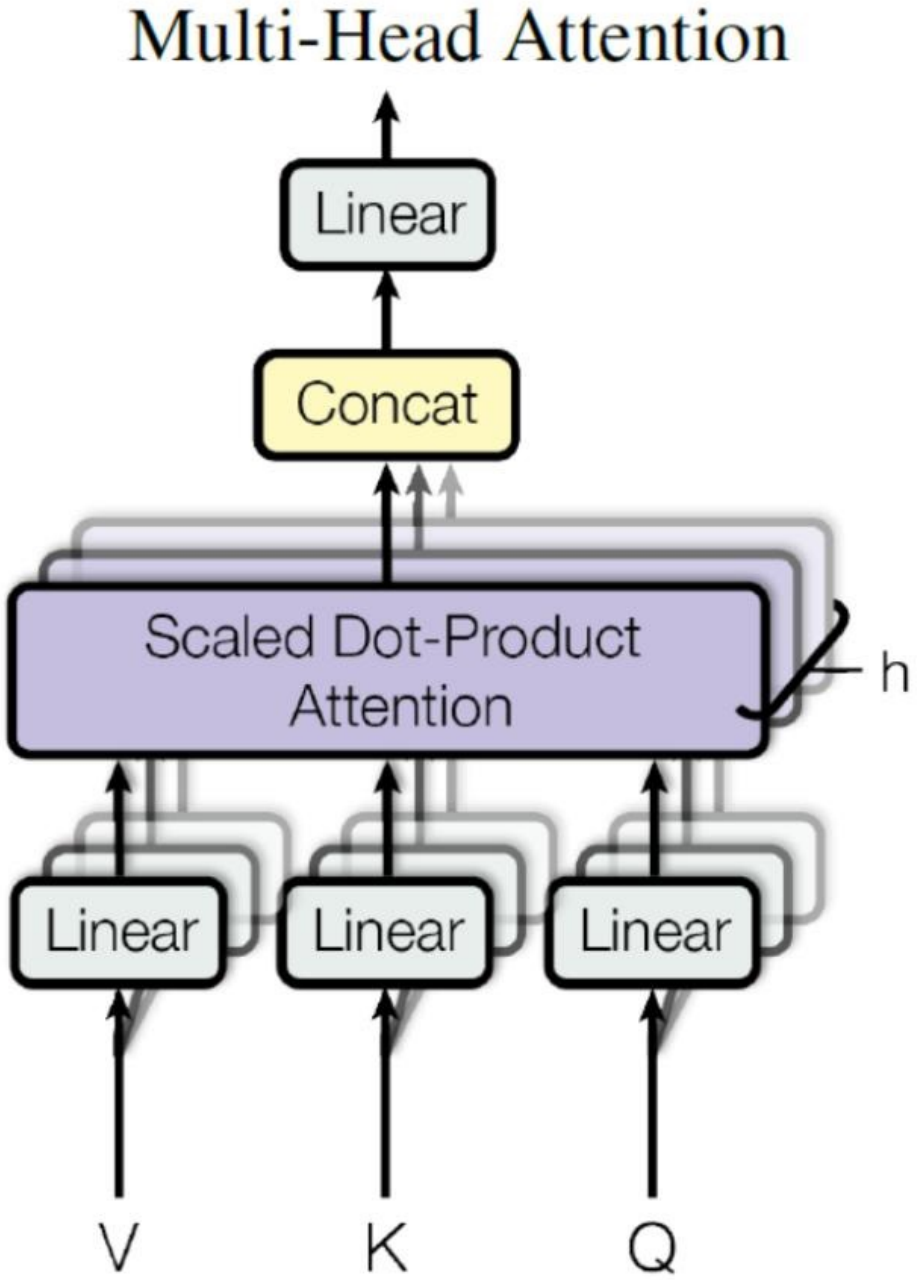
d_k是Q, K矩阵的列数，即向量维度



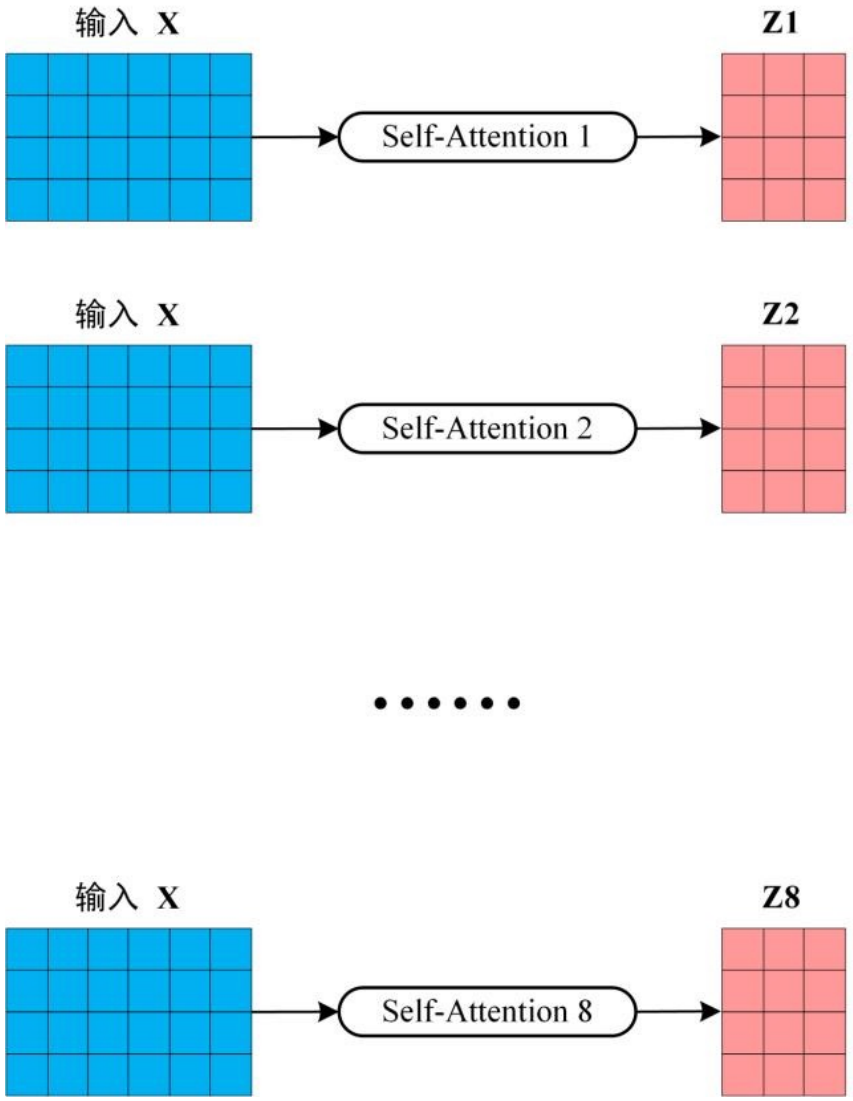
Encode: multi-head attention



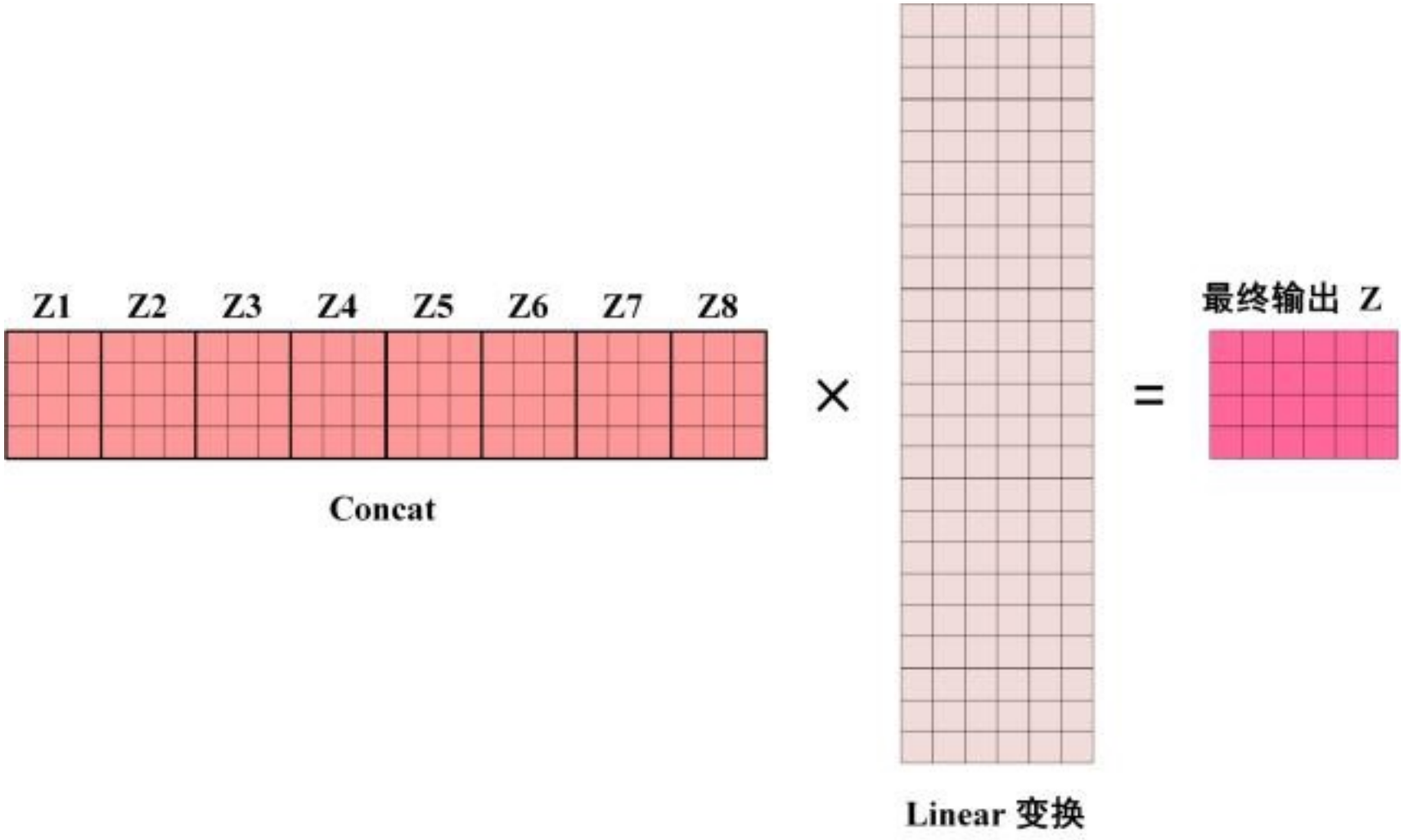
Encode: multi-head attention



Encode: multi-head attention



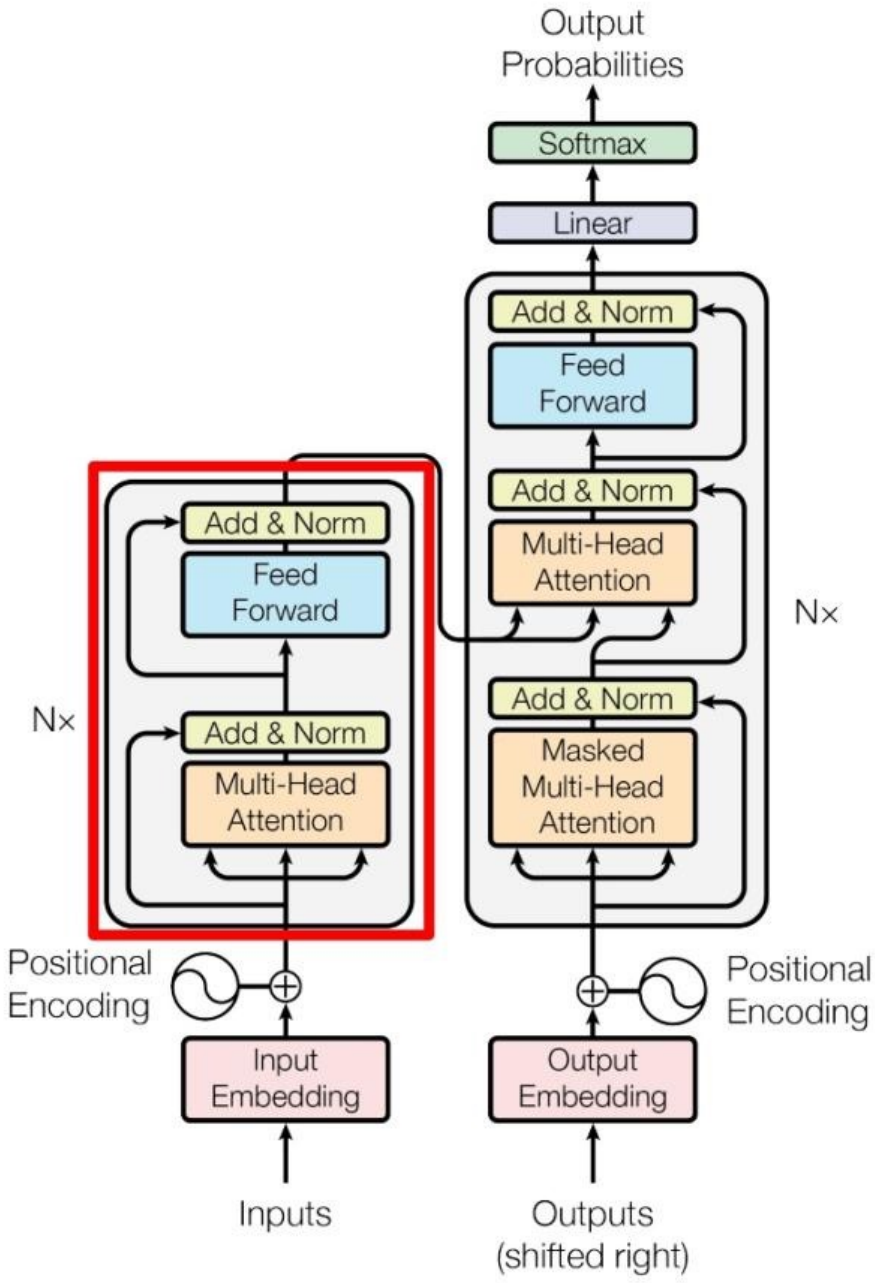
Encode: multi-head attention



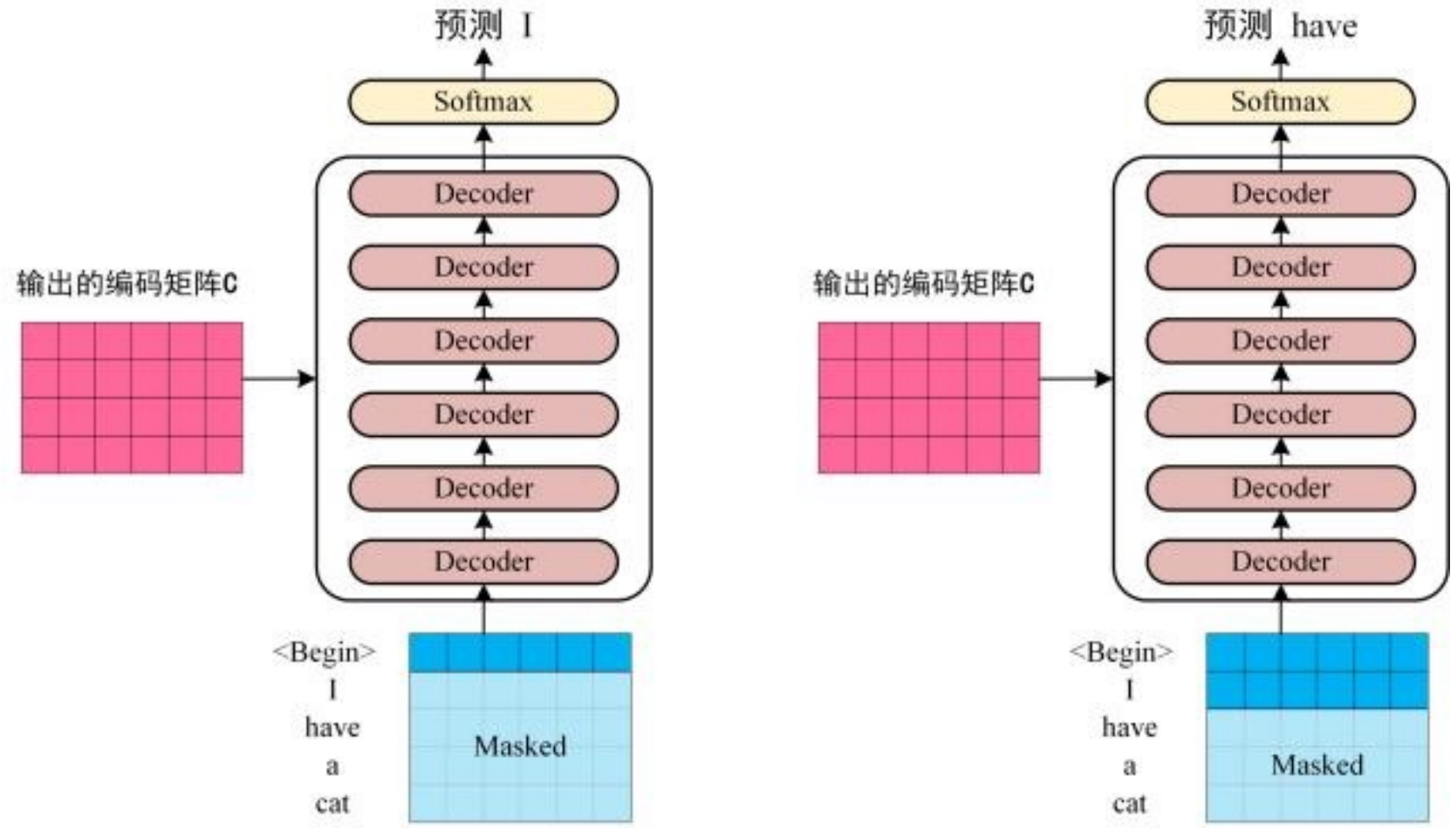
Encode: add and norm

$$\text{LayerNorm}\big(X + \text{MultiHeadAttention}(X)\big)$$
$$\text{LayerNorm}\big(X + \text{FeedForward}(X)\big)$$

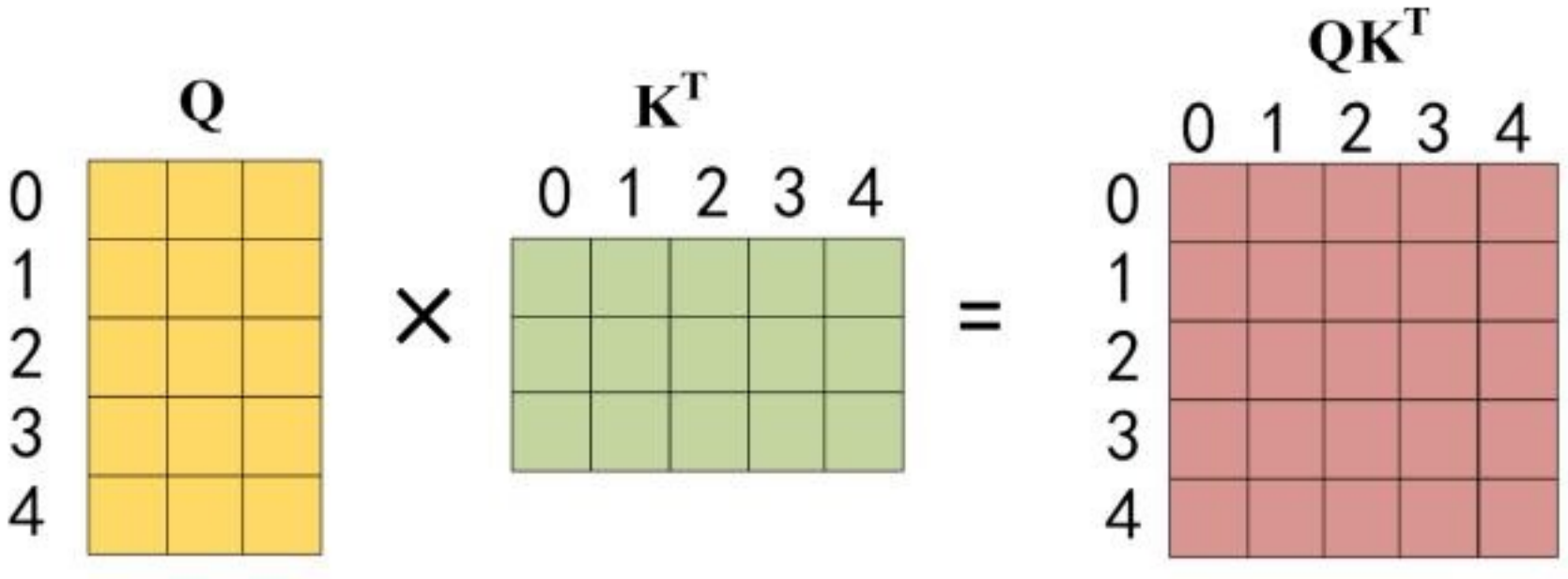
残差链接，解决深层训练问题



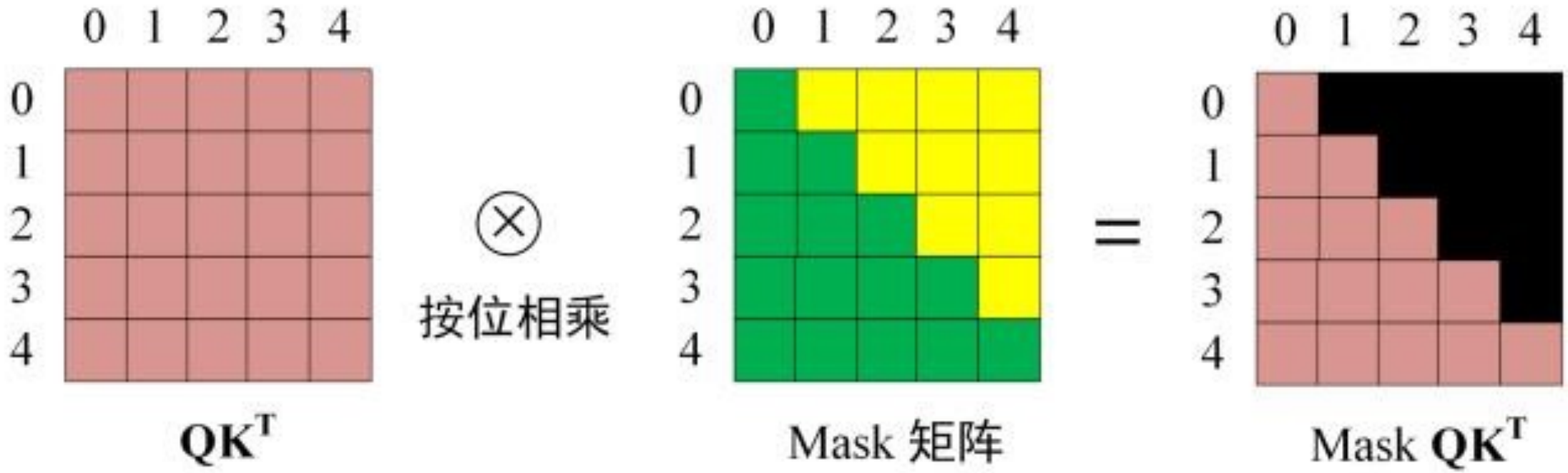
Decoder: masked self-attention



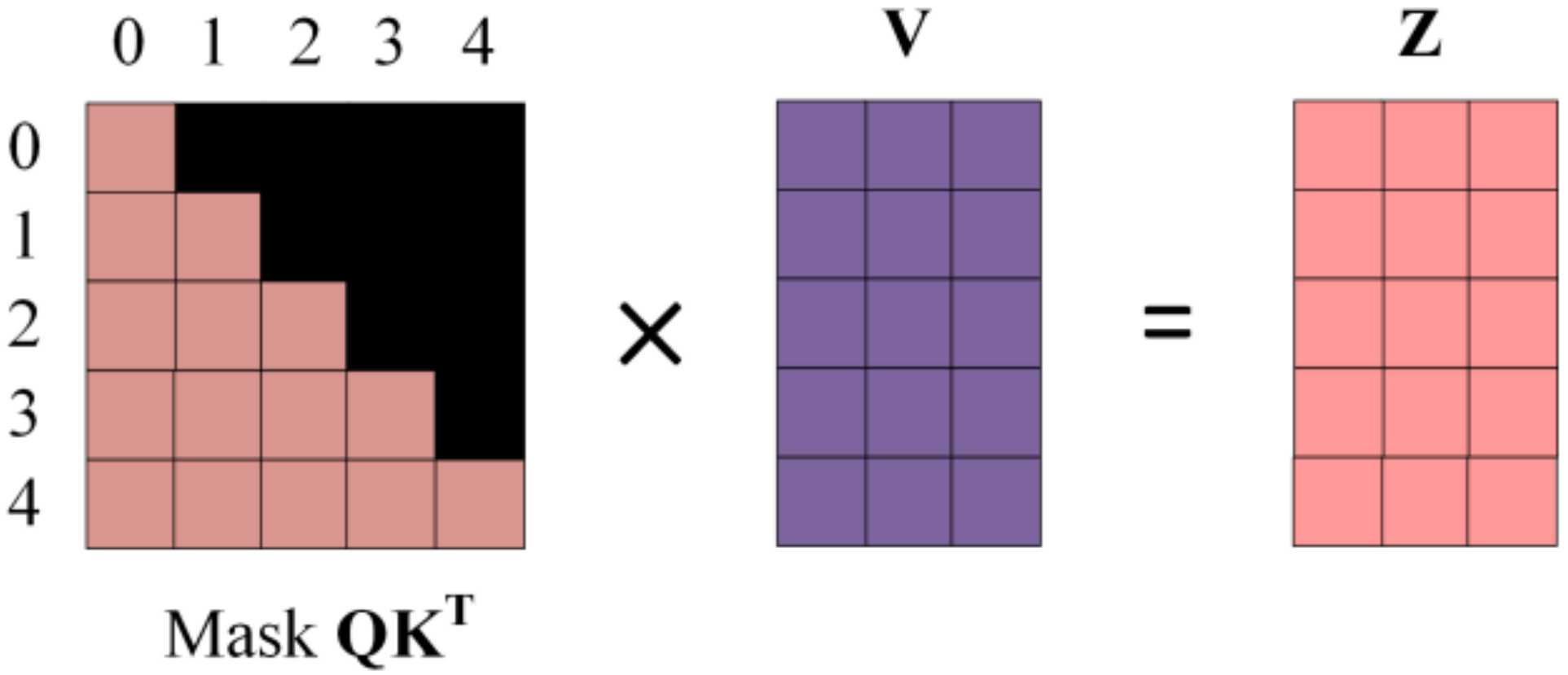
Decoder: masked self-attention



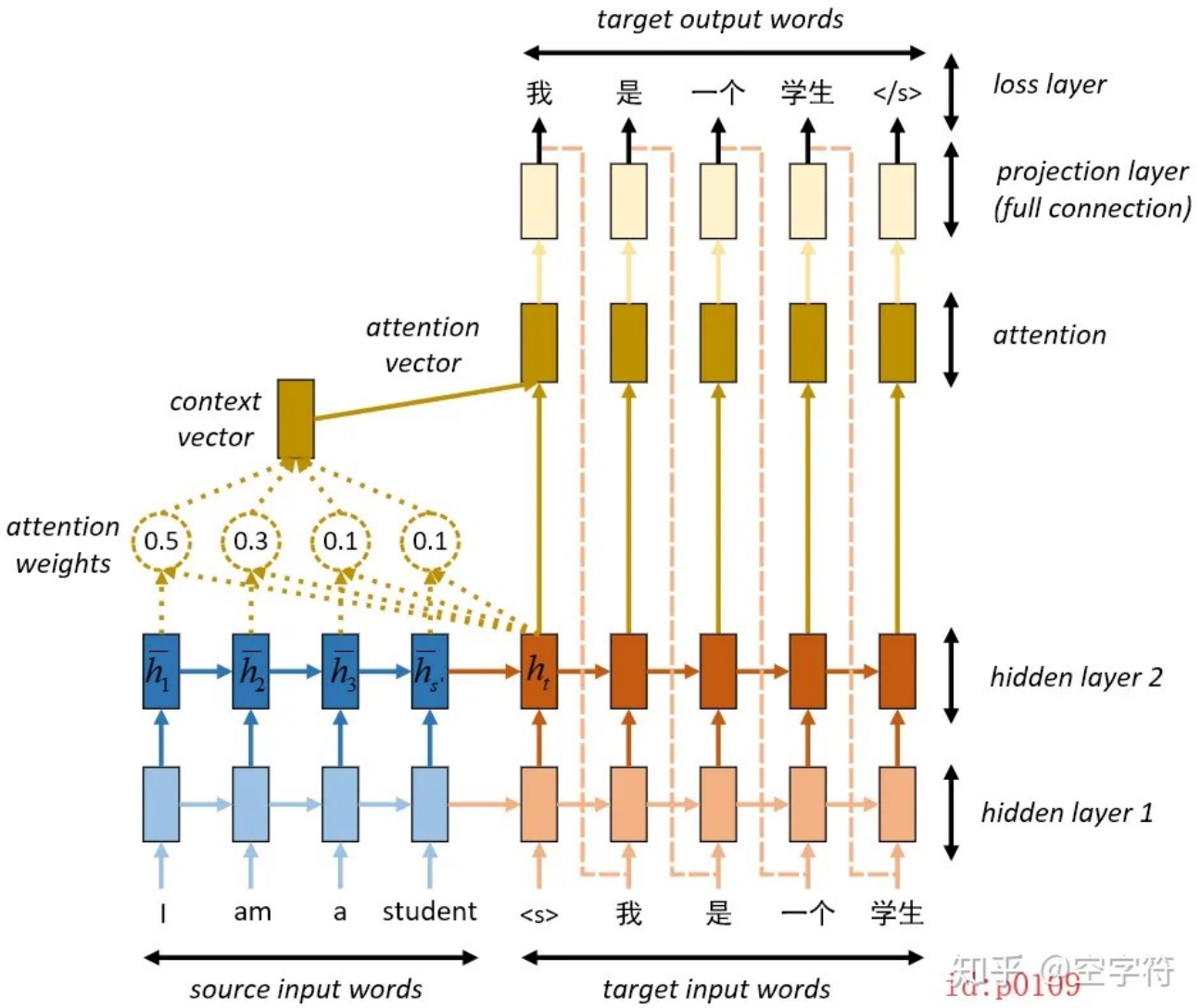
Decoder: masked self-attention

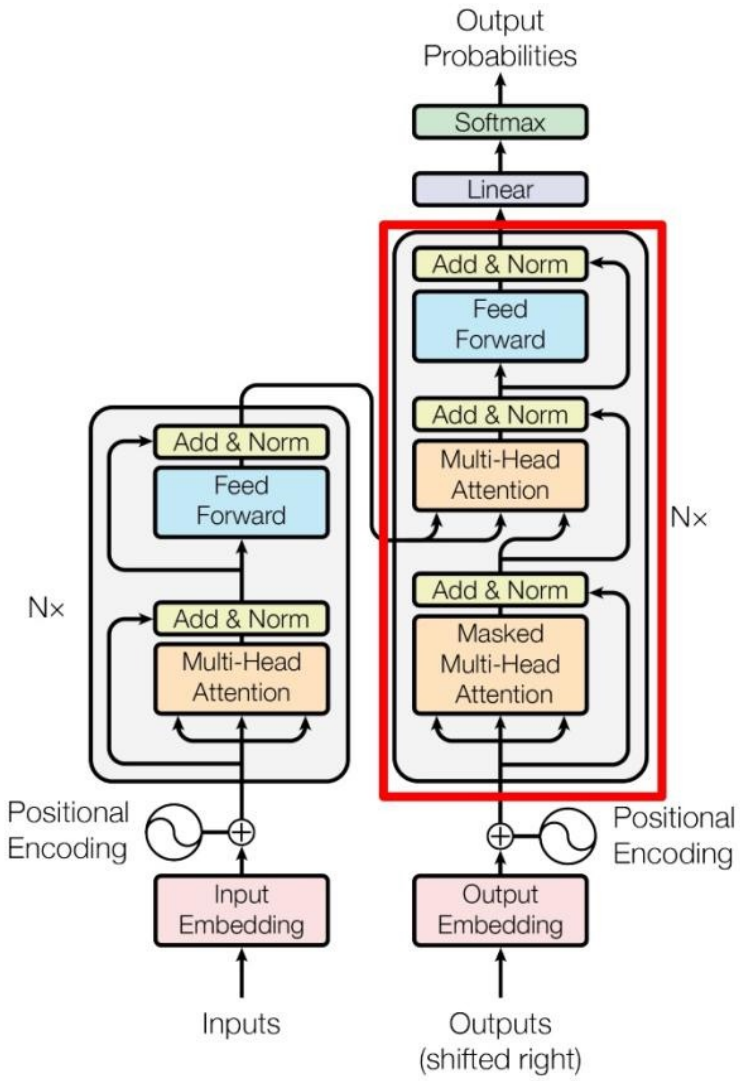


Decoder: masked self-attention

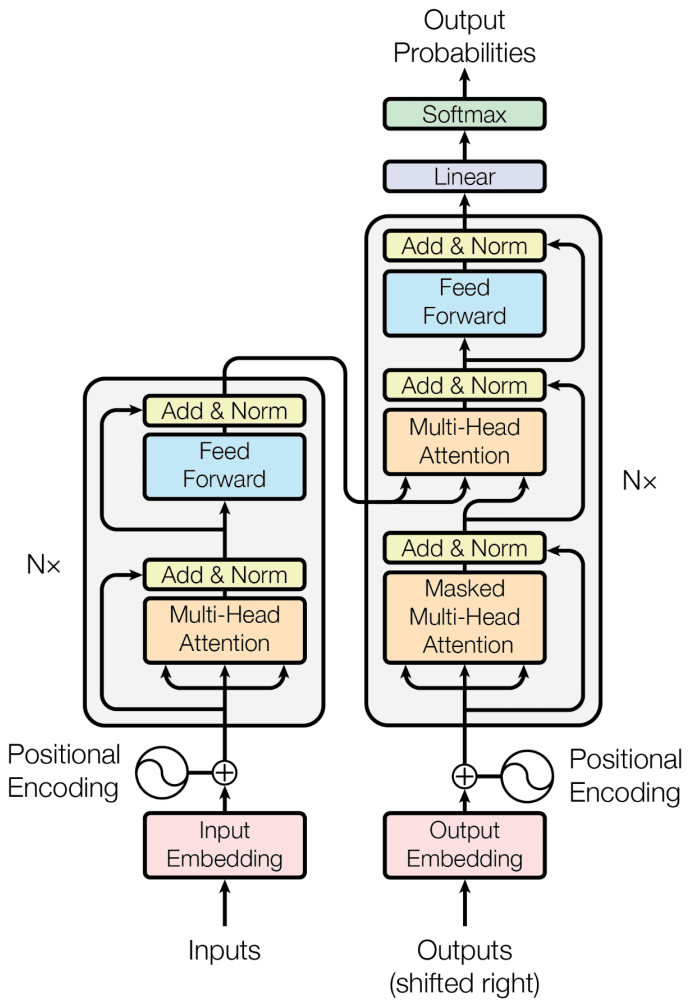


Seq2Seq


target output words
我
是
一个
学生
</s>
loss layer
projection layer
(full connection)
attention
attention vector
context vector
attention weights
0.5
0.3
0.1
0.1
hidden layer 2
hidden layer 1
I
am
a
student
<s>
我
是
一个
学生
source input words
target input words
知乎 @空字符
id: p0109


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

Summary



Transformer没有时序结构，容易并行

需要引入位置的embedding

内存开销巨大

Reference

https://zhuanlan.zhihu.com/p/338817680