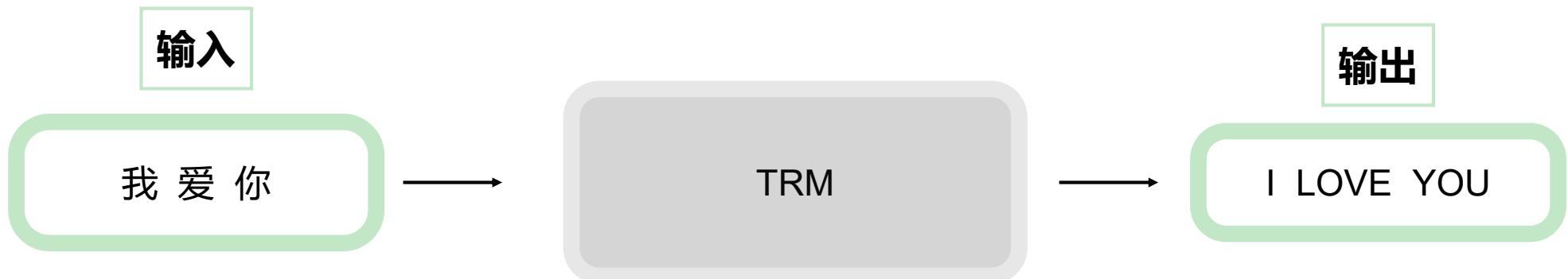




Transformer 从零解读



后台回复【答案解析】



扫码关注微信公众号

文章周更

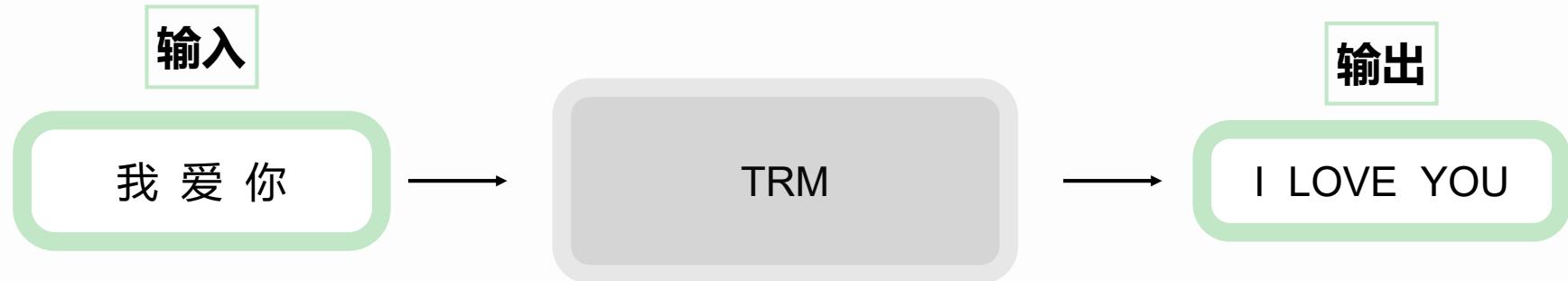
知识分享

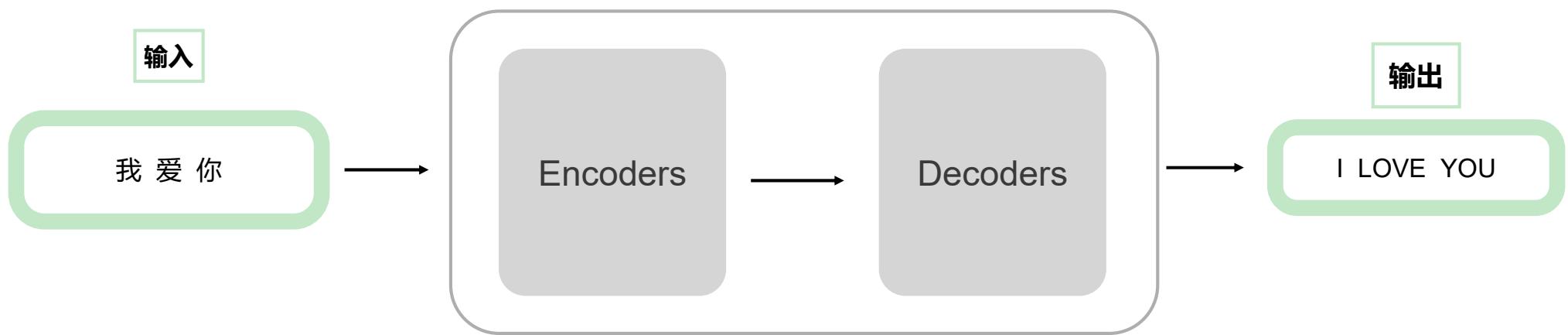
一起进步

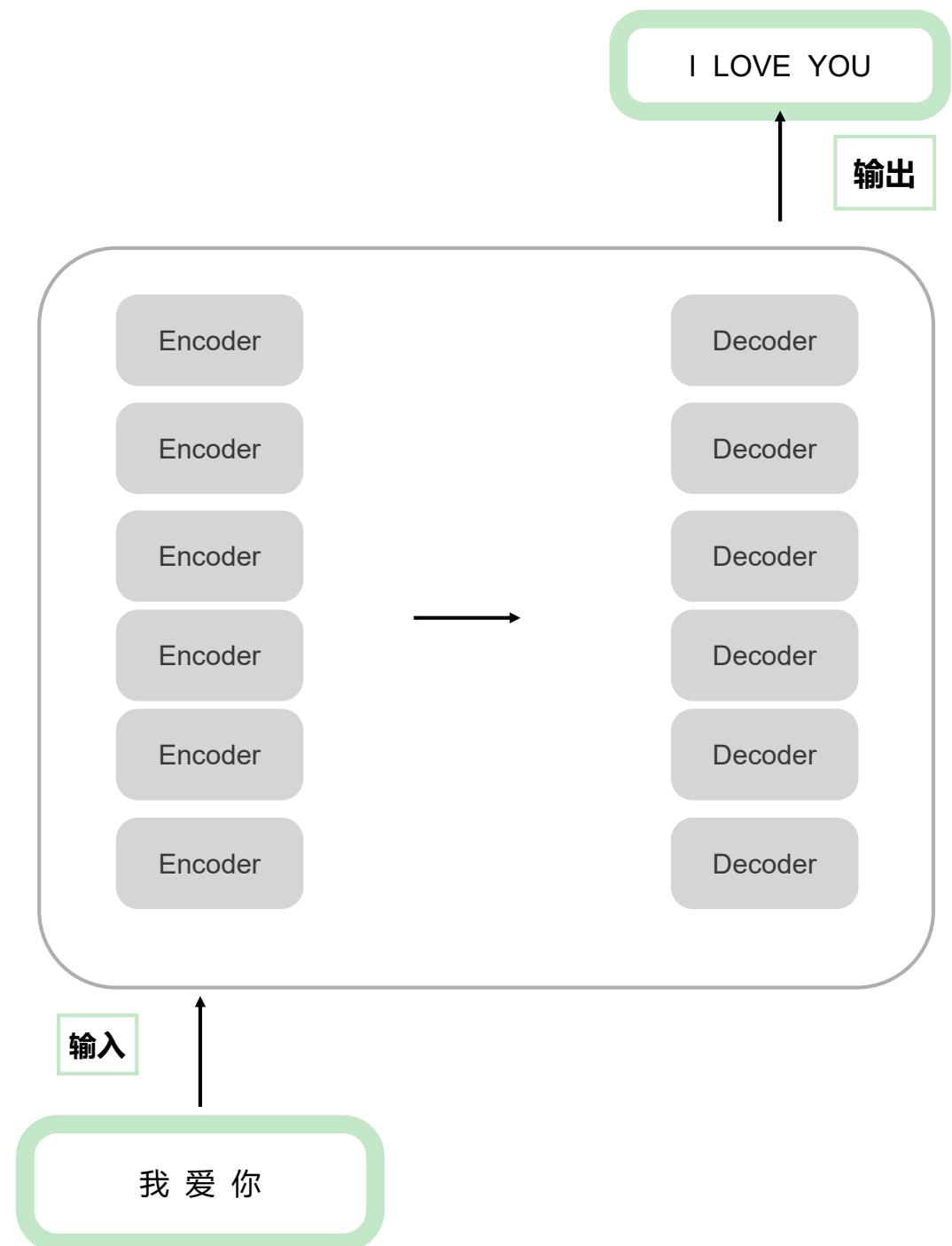
求关注，求点赞，求一切！！

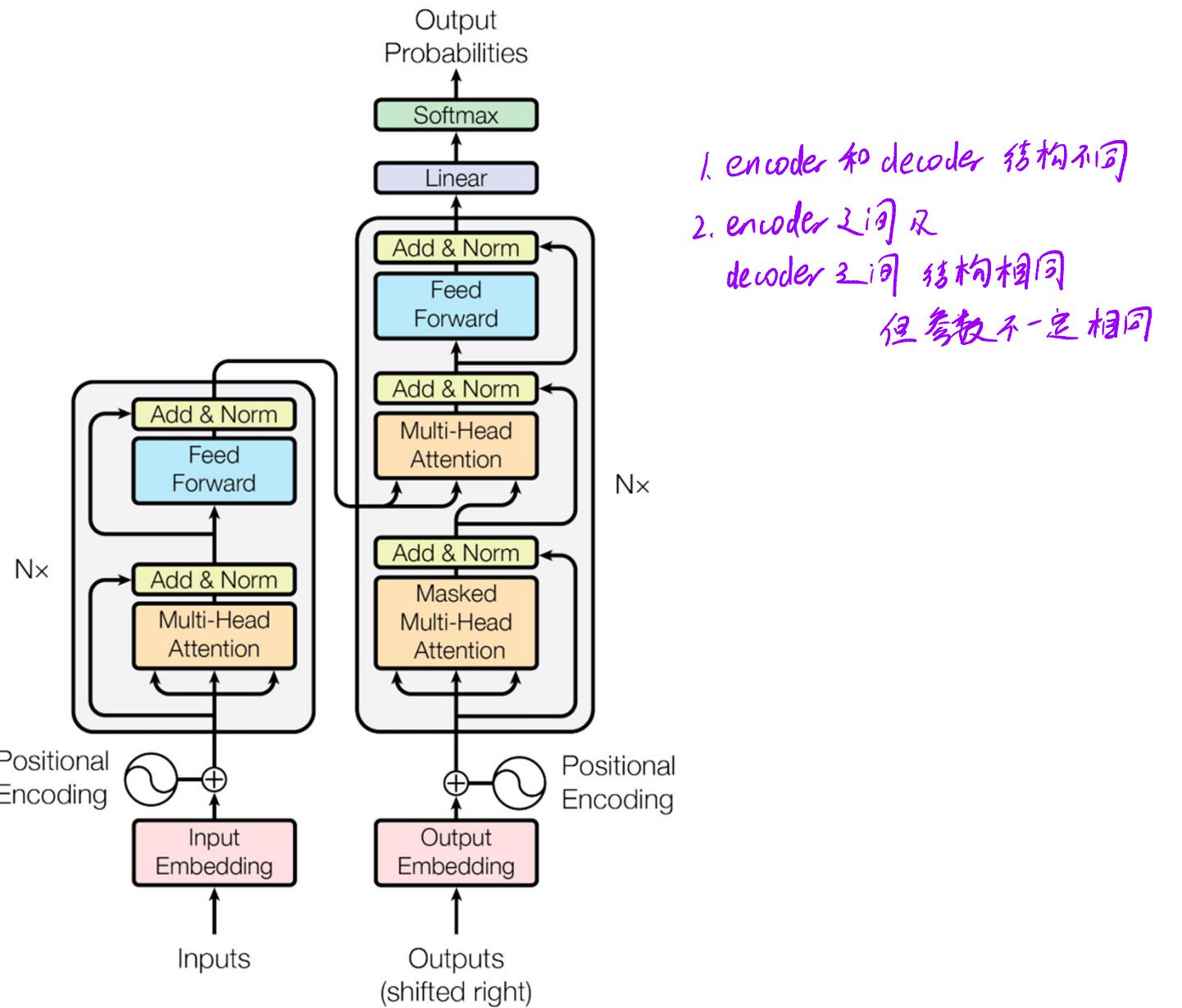
1. 位置编码
2. 多头注意力机制
3. 残差和layerNorm
4. 前馈神经网络
5. TRM面试题讲解

TRM在做一个什么事情？



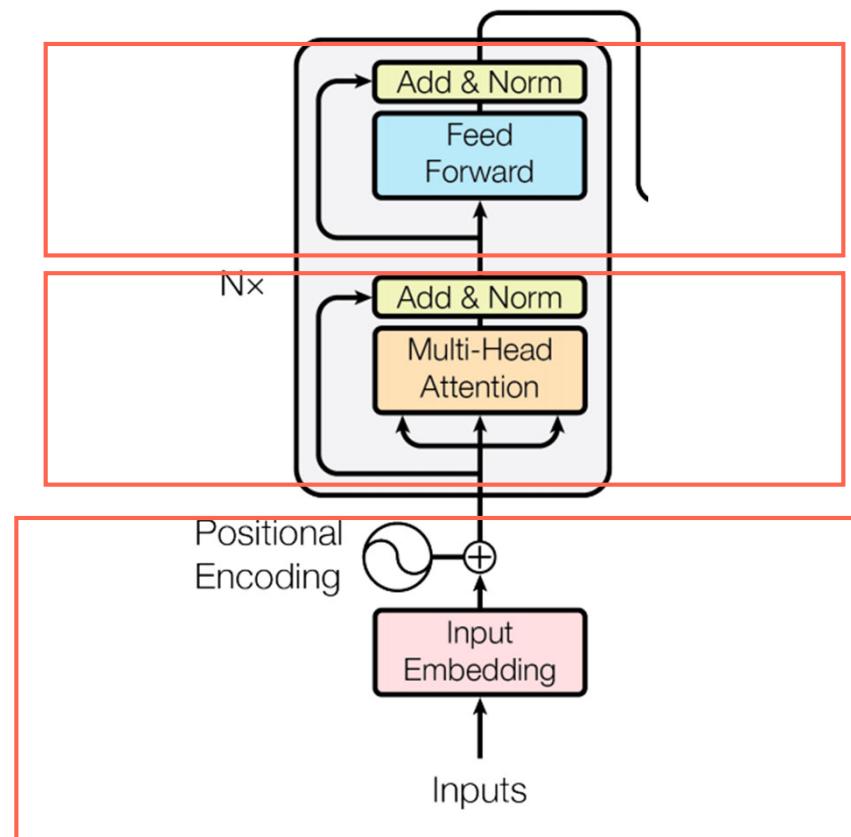






1. encoder 和 decoder 结构不同
 2. encoder 之间 及
 decoder 之间 结构相同
 但参数不一定相同

Figure 1: The Transformer - model architecture.



3 前馈神经网络

2 注意力机制

1 输入部分

输入部分

1. Embedding

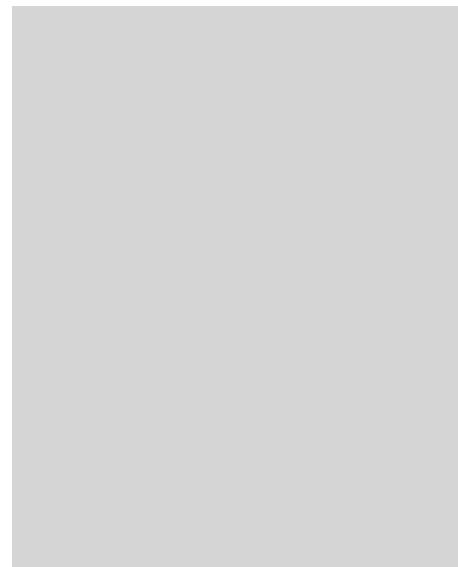
2. 位置嵌入

Embedding

12

我
爱
你
...

512



1. word2vec

2. 随机初始化

词嵌入

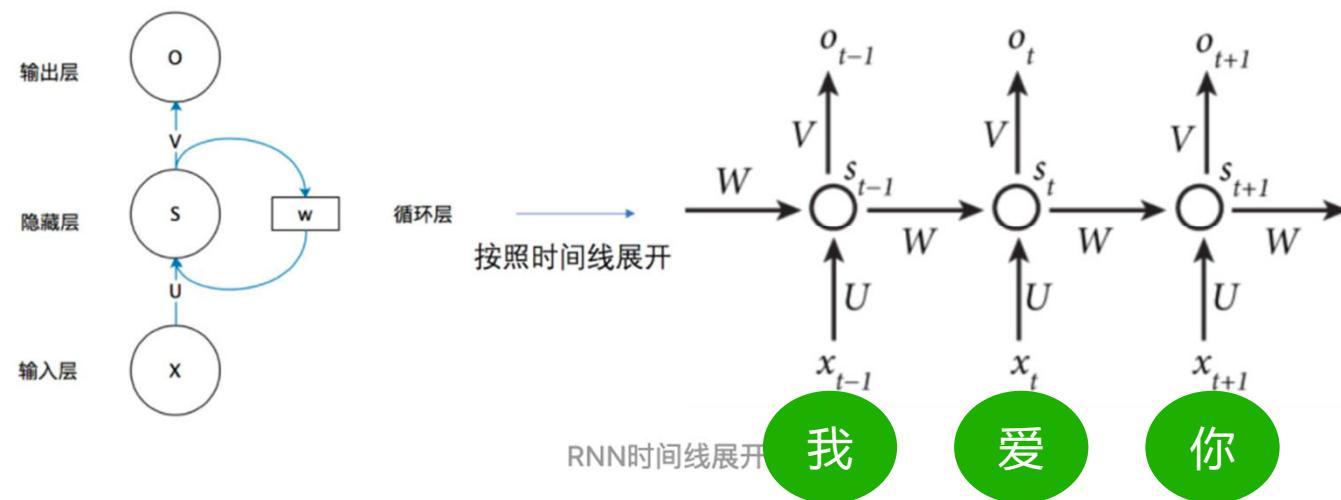
词 → 可计算的向量

通过查表

位置编码

为什么需要：TRNN比RNN 缺少序列关系，所以 加入位置编码

RNN中，参数是同一套

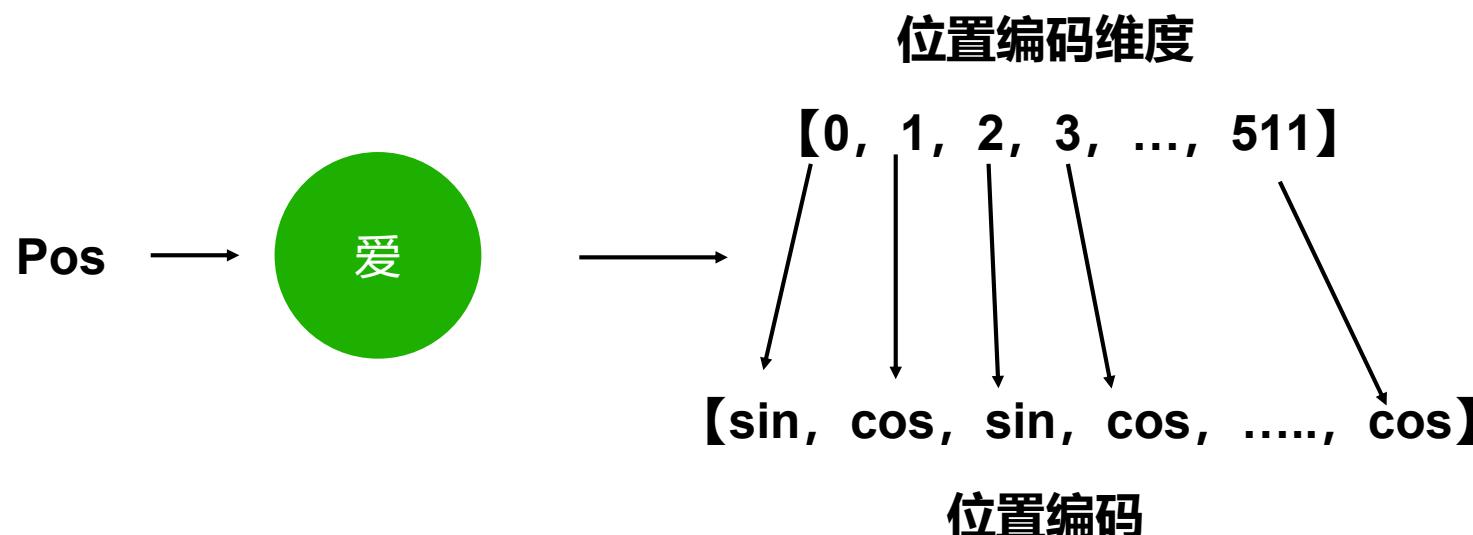


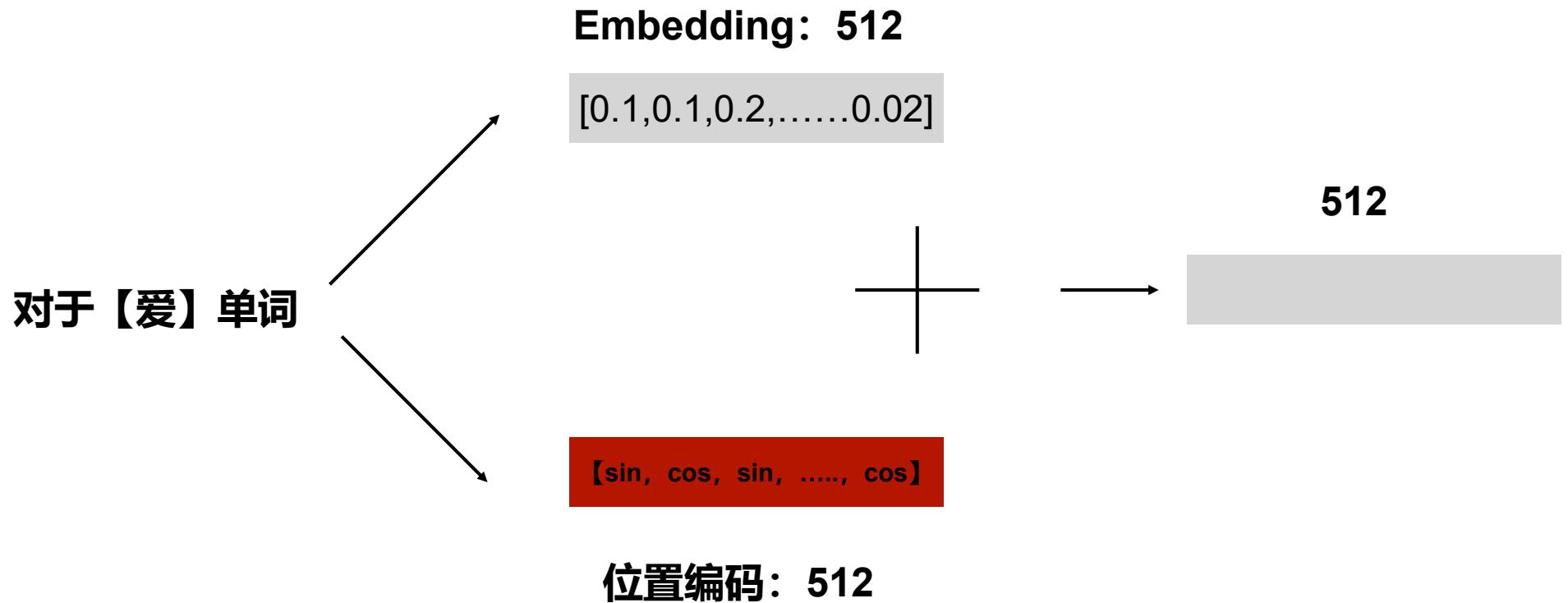
位置编码公式

绝对位置信息

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

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





引申一下为什么位置嵌入会有用

借助上述公式，我们可以得到一个特定位置的 d_{model} 维的位置向量，并且借助三角函数的性质

$$\begin{cases} \sin(\alpha + \beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta \\ \cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta \end{cases} \quad (2)$$

我们可以得到：

$$\begin{cases} PE(pos + k, 2i) = PE(pos, 2i) \times PE(k, 2i + 1) + PE(pos, 2i + 1) \times PE(k, 2i) \\ PE(pos + k, 2i + 1) = PE(pos, 2i + 1) \times PE(k, 2i + 1) - PE(pos, 2i) \times PE(k, 2i) \end{cases} \quad (3)$$

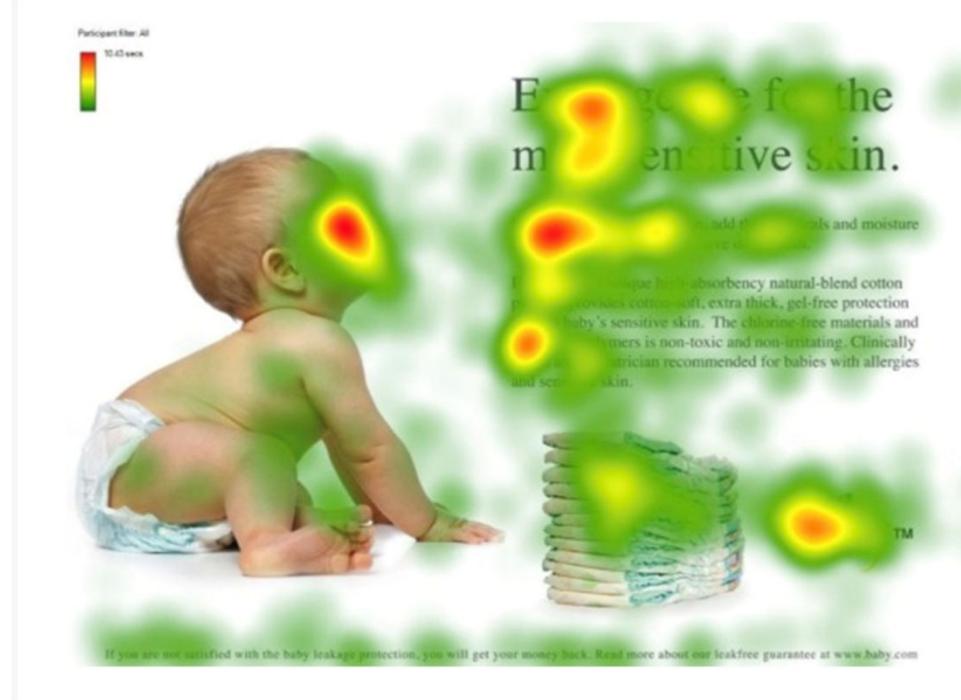
可以看出，对于 $pos+k$ 位置的位置向量某一维 $2i$ 或 $2i+1$ 而言，可以表示为， pos 位置与 k 位置的位置向量的 $2i$ 与 $2i+1$ 维的线性组合，这样的线性组合意味着位置向量中蕴含了相对位置信息。

但是这种相对位置信息会在注意力机制那里消失

注意力机制

1. 基本的注意力机制
2. 在TRM中怎么操作

注意力机制本质



婴儿在干嘛

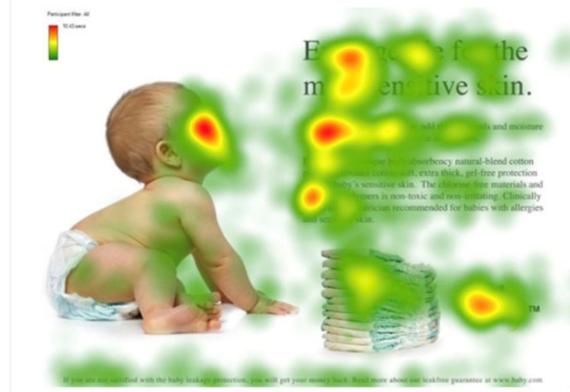
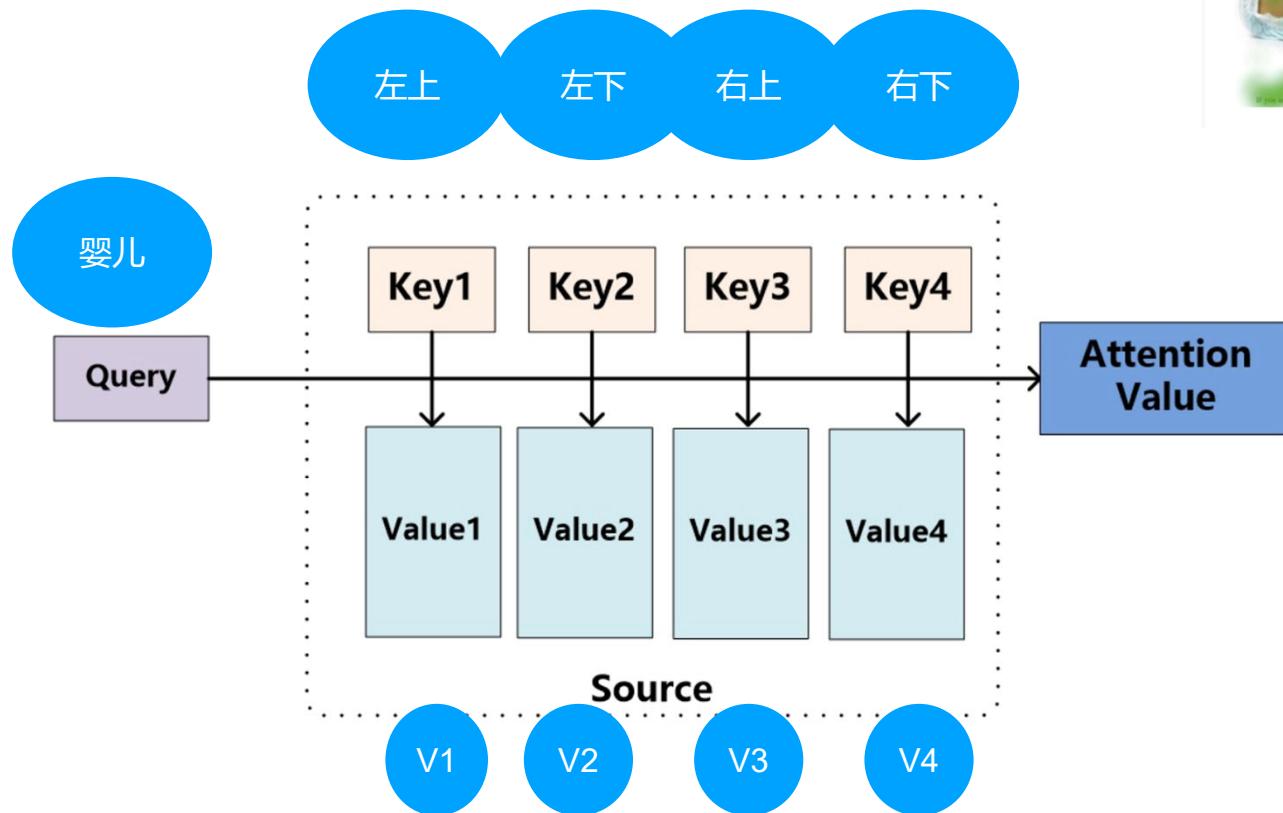
$$\text{Attention}(Q, K, V) = \underbrace{\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V}_{\text{对 } V \text{ 的加权和}}$$

softmax $\left(\frac{QK^T}{\sqrt{d_k}}\right)$ 在做归一化

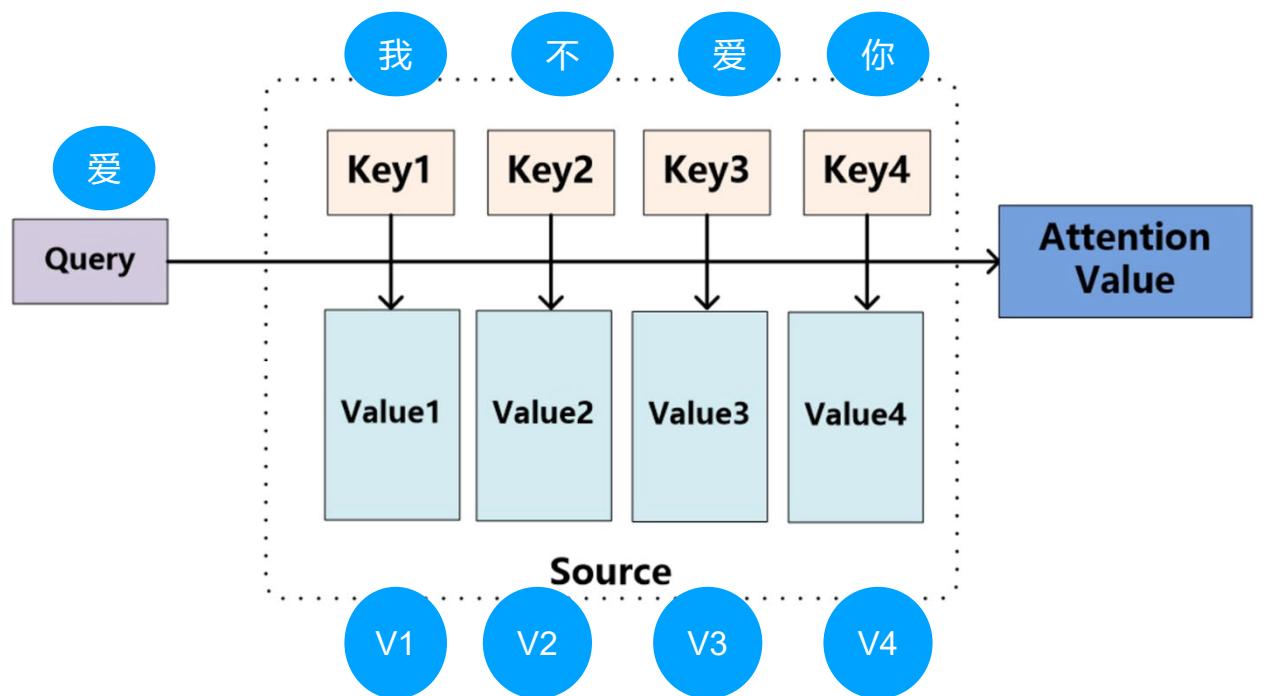
除以 $\sqrt{d_k}$ 的原因:

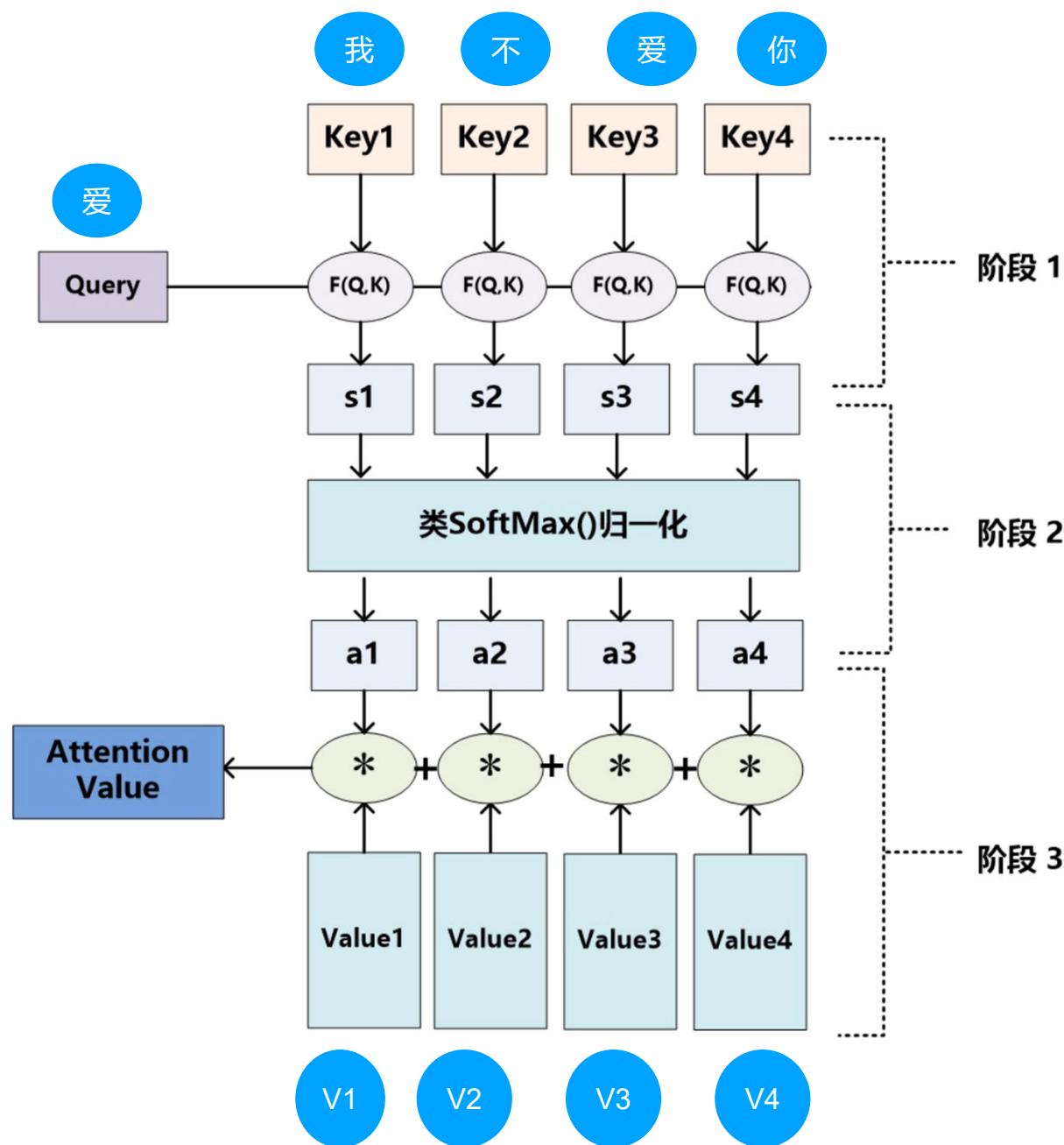
- { ① 防止 QK^T 值太大，导致 softmax 反向传播值太小，梯度消失
- ② 控制方差为 1

从公式角度来看：拿上面图片举例子



从公式角度来看

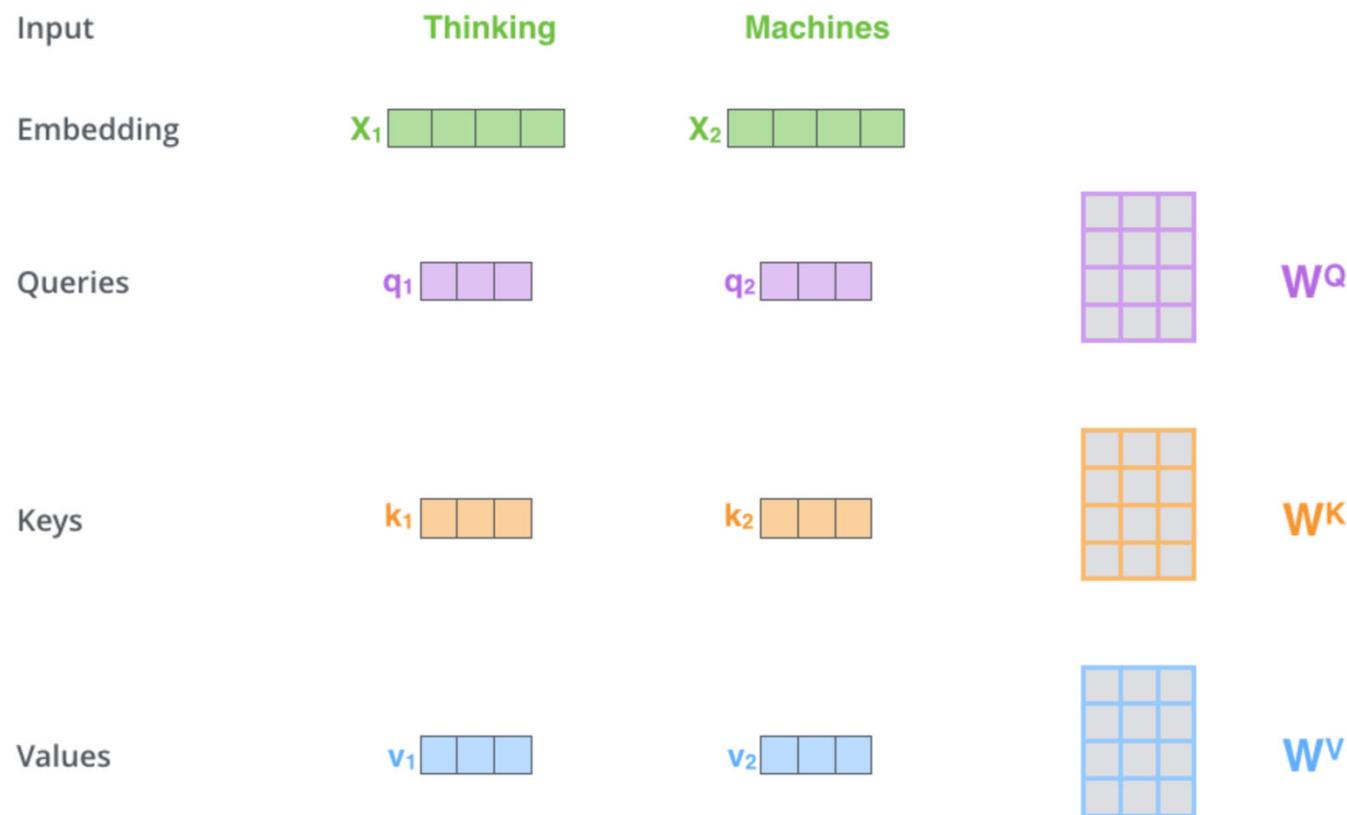




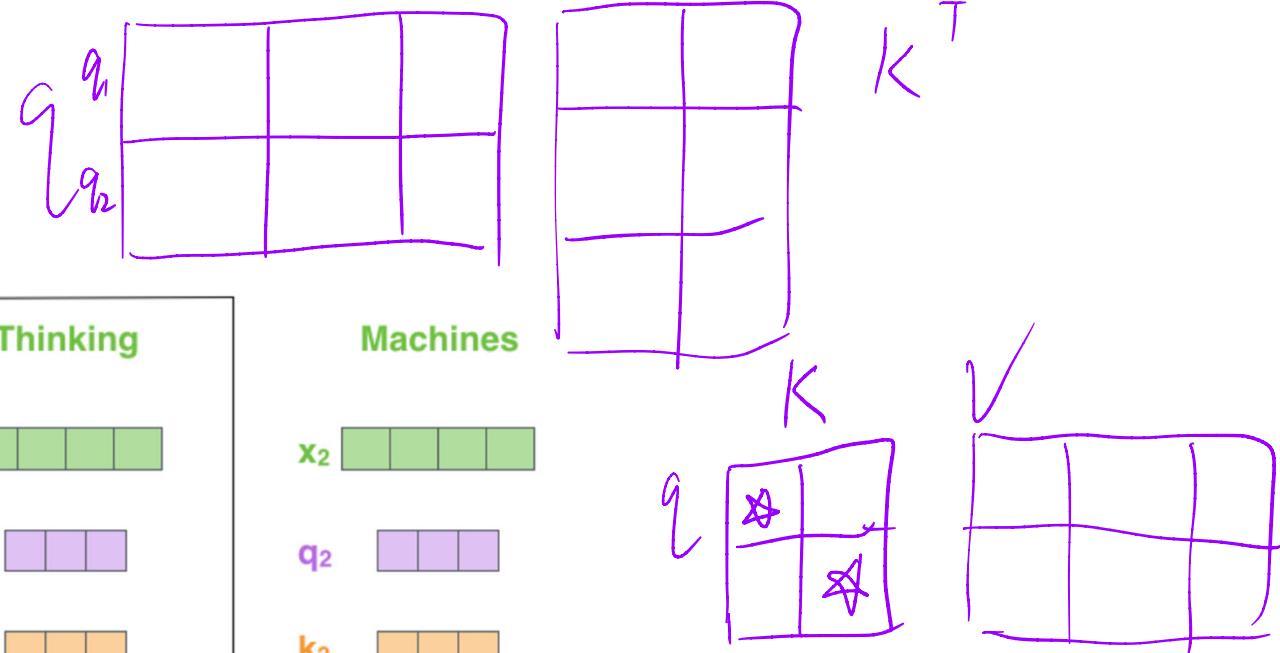
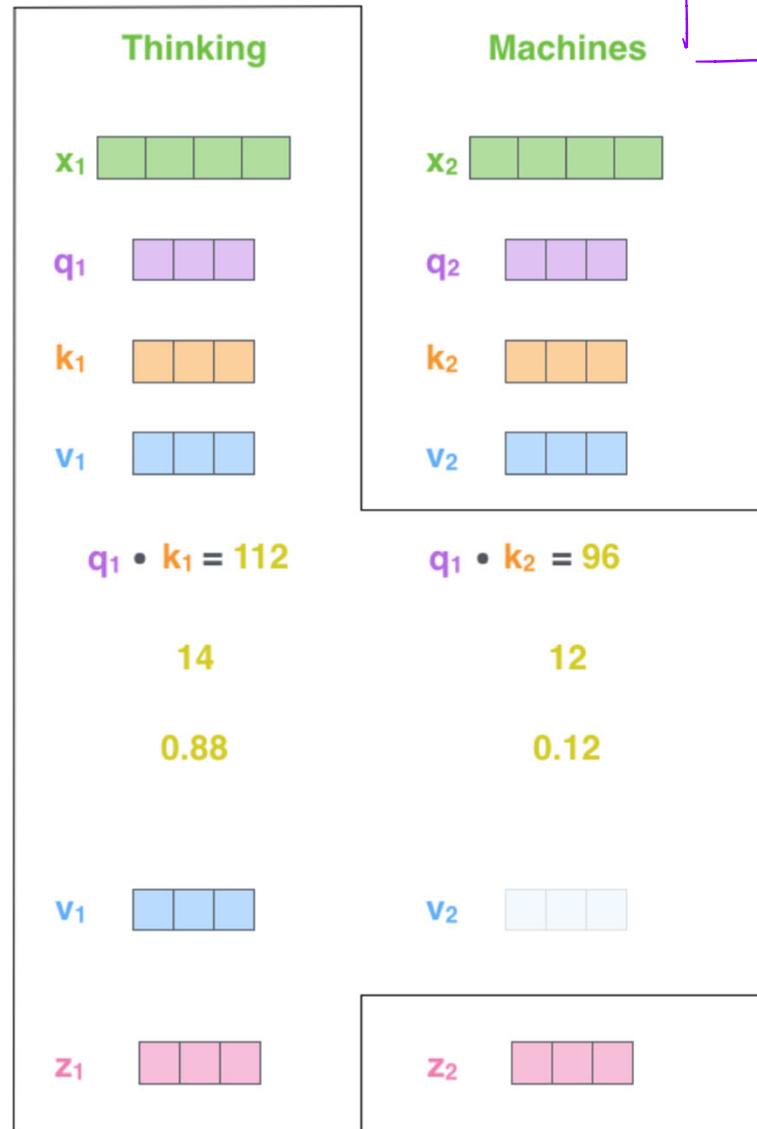
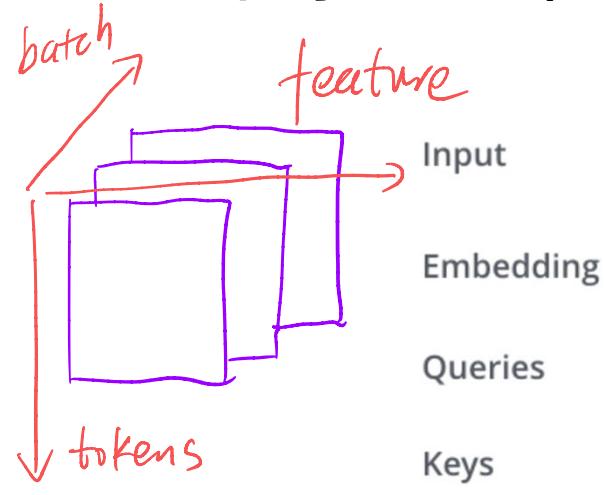
TRM中的注意力

在只有单词向量的情况下，如何获取QKV

对每个单词的一套矩阵参数都相同，但一套矩阵参数中的 $\{W^Q, W^K, W^V\}$ 不相同



计算QK相似度， 得到attention值



实际代码使用矩阵，方便并行

$$\mathbf{X} \times \mathbf{W}^Q = \mathbf{Q}$$

$$\mathbf{X} \times \mathbf{W}^K = \mathbf{K}$$

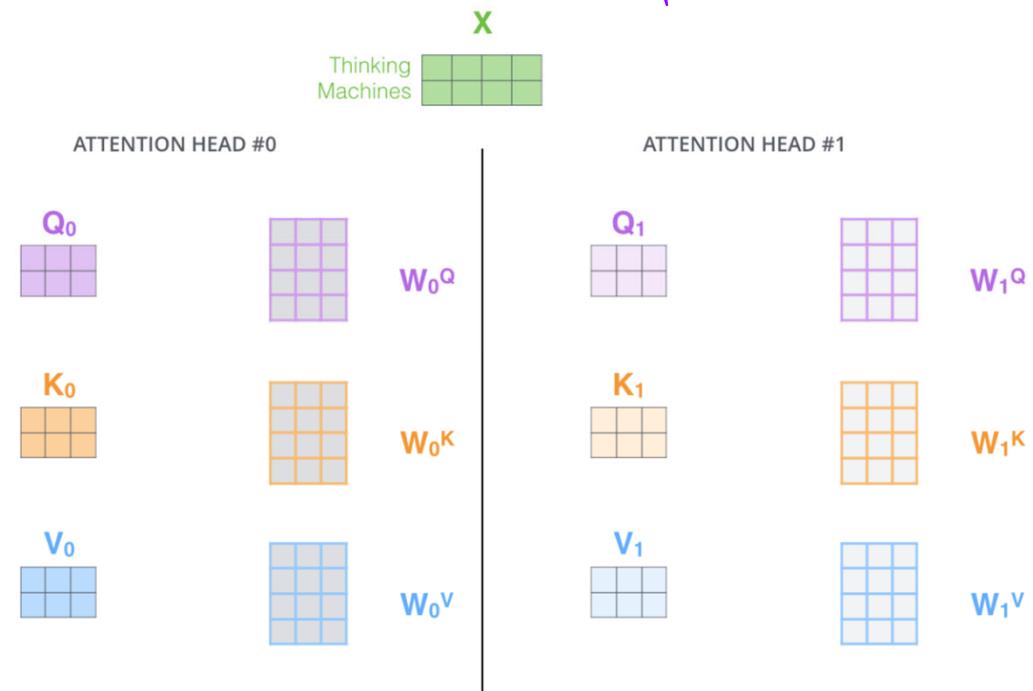
$$\mathbf{X} \times \mathbf{W}^V = \mathbf{V}$$

$$\text{softmax}\left(\frac{\mathbf{Q} \times \mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V} = \mathbf{Z}$$

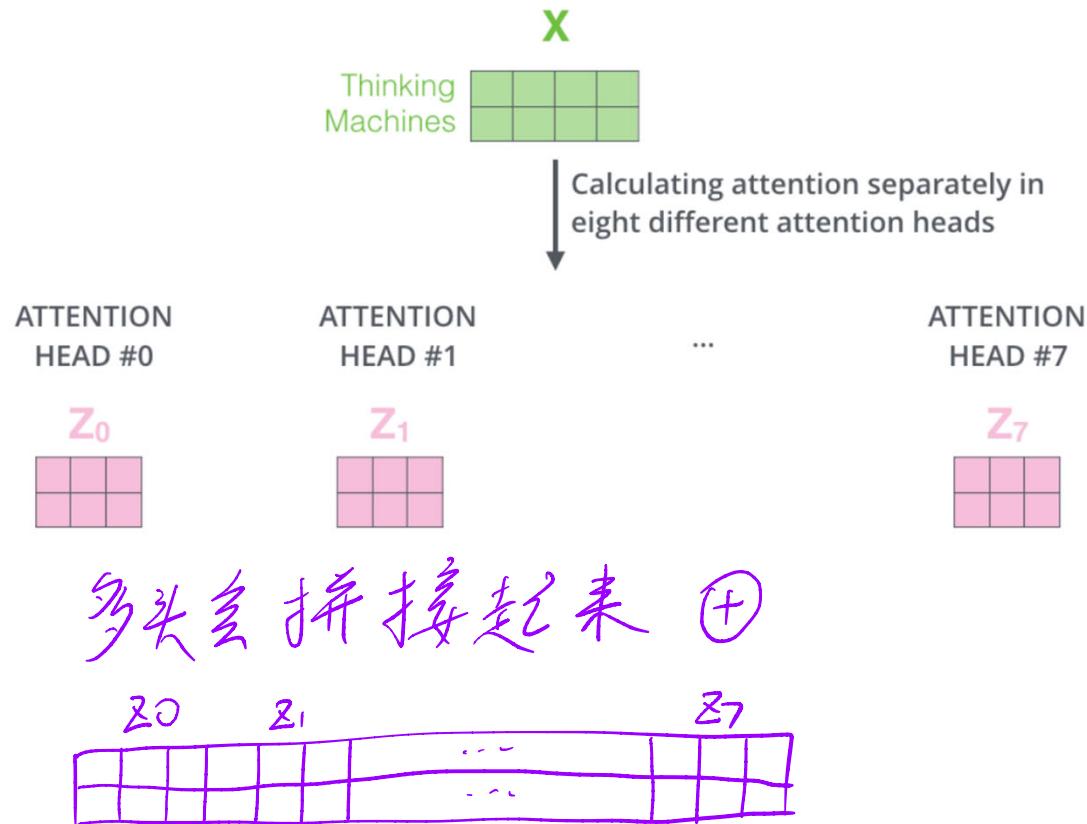
The self-attention calculation in matrix form

多头注意力机制

多头，就是多套 $\left\{ \begin{array}{l} W^Q \\ W^K \\ W^V \end{array} \right.$ 参数



多个头就会有多个输出，需要合在一起输出

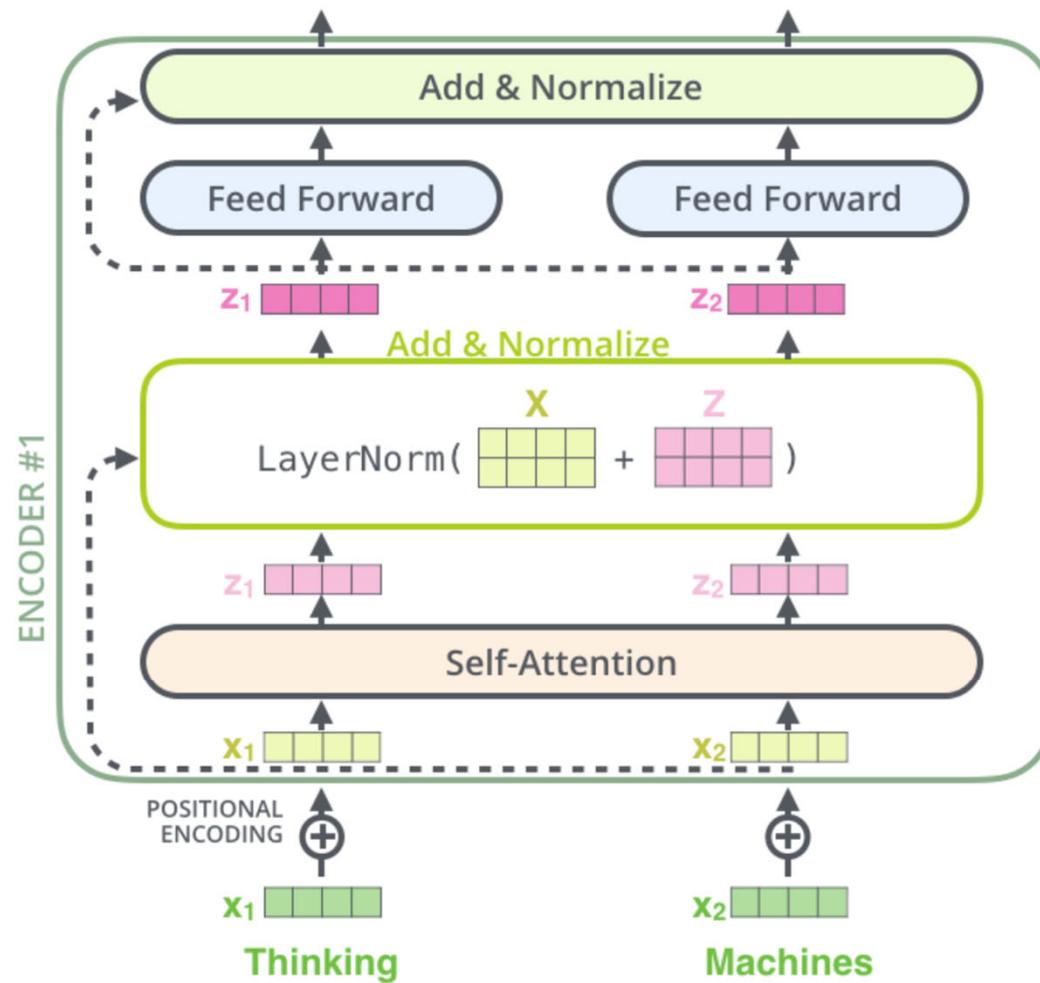


多头会拼接起来 ①

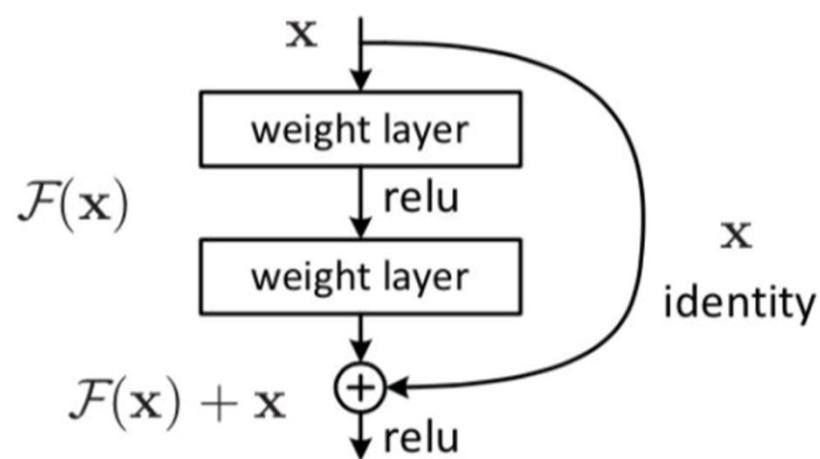


可以再乘一个 W 与多头序列做矩阵乘法
得到 Attention Layer 输出

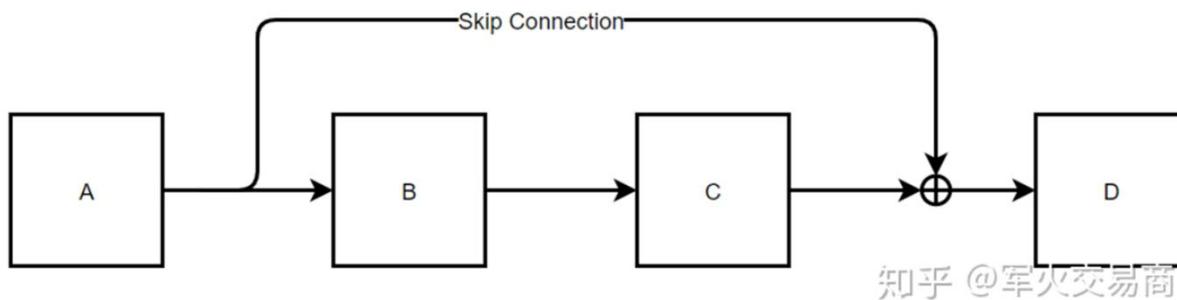
残差和LayNorm



残差

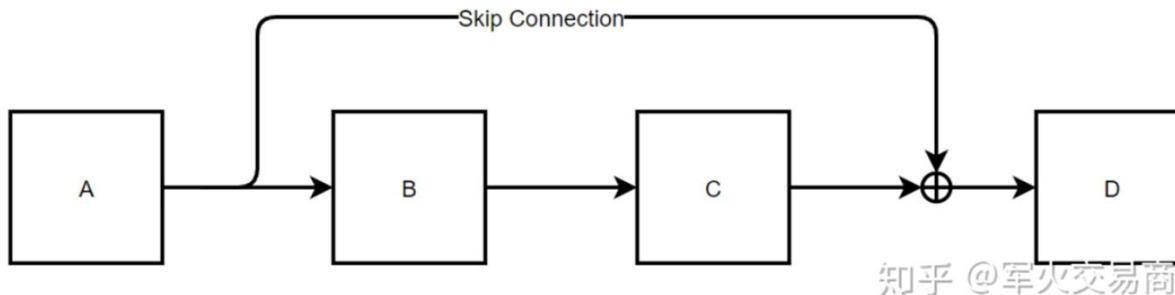


残差的作用



知乎 @军火交易商

A,B,C,D 为四个不同的网络块，箭头代表“数据流”；



A, B, C, D 为四个不同的网络块，箭头代表“数据流”；

根据后向传播的链式法则，

$$\frac{\partial L}{\partial X_{Aout}} = \frac{\partial L}{\partial X_{Din}} \frac{\partial X_{Din}}{\partial X_{Aout}}$$

而 $X_{Din} = X_{Aout} + C(B(X_{Aout}))$

所以：

$$\frac{\partial L}{\partial X_{Aout}} = \frac{\partial L}{\partial X_{Din}} [1 + \frac{\partial X_{Din}}{\partial X_C} \frac{\partial X_C}{\partial X_B} \frac{\partial X_B}{\partial X_{Aout}}]$$

保证在右侧项求导为零的情况下
 仍然有梯度
 缓解梯度消失

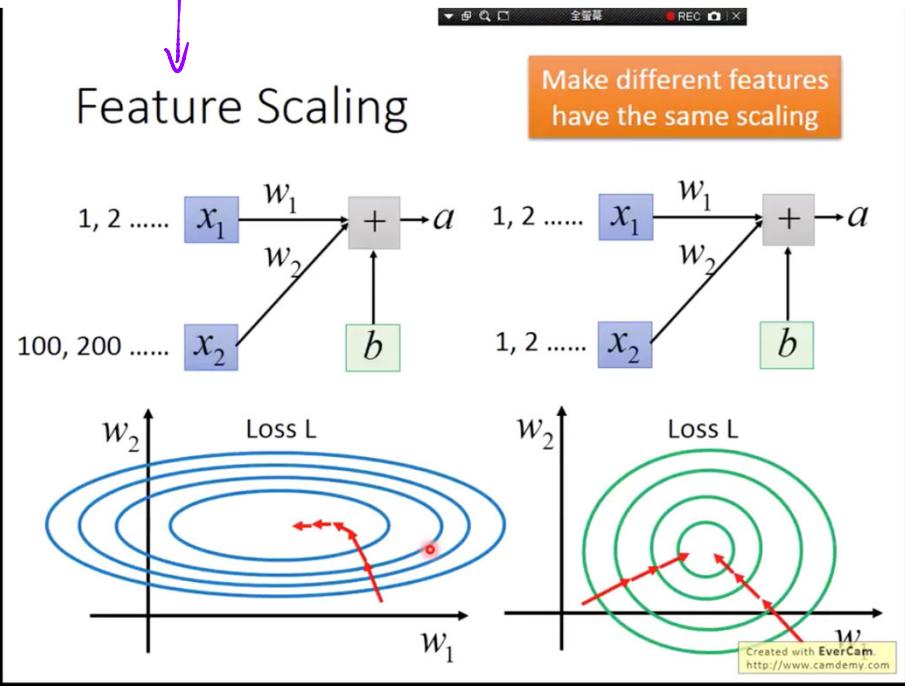
想一下RNN

Layer Normalization

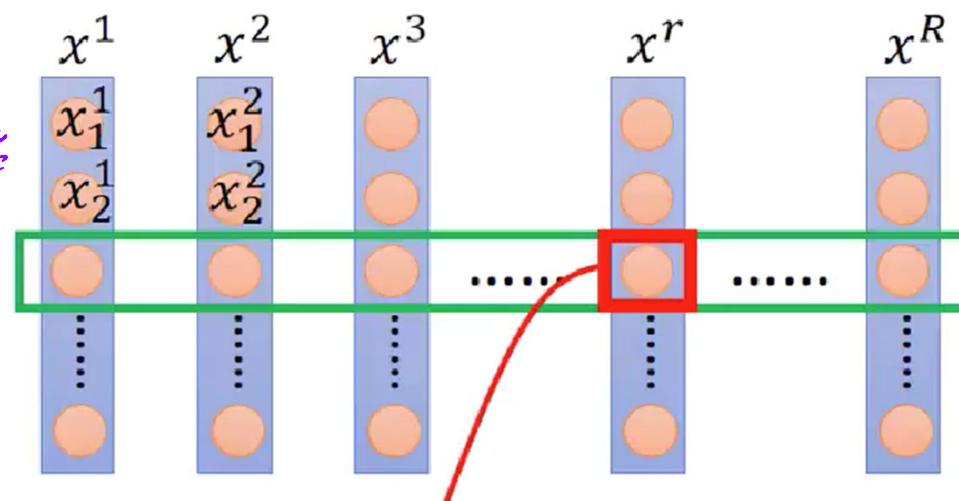
BN的效果差，所以不用

什么是BN，以及使用场景

消除量纲，让模型收敛更快



针对 Batch 中所有样本
的每一个特征做 BN



BN优点

第一个就是可以解决内部协变量偏移

第二个优点就是缓解了梯度饱和问题（如果使用sigmoid激活函数的话），加快收敛

BN的缺点

第一个，batch_size较小的时候，效果差。

用样本的分布模拟空间的分布

BN的缺点

第二个缺点就是 BN 在RNN中效果比较差。这一点和第一点原因很类似，不过我单挑出来说。

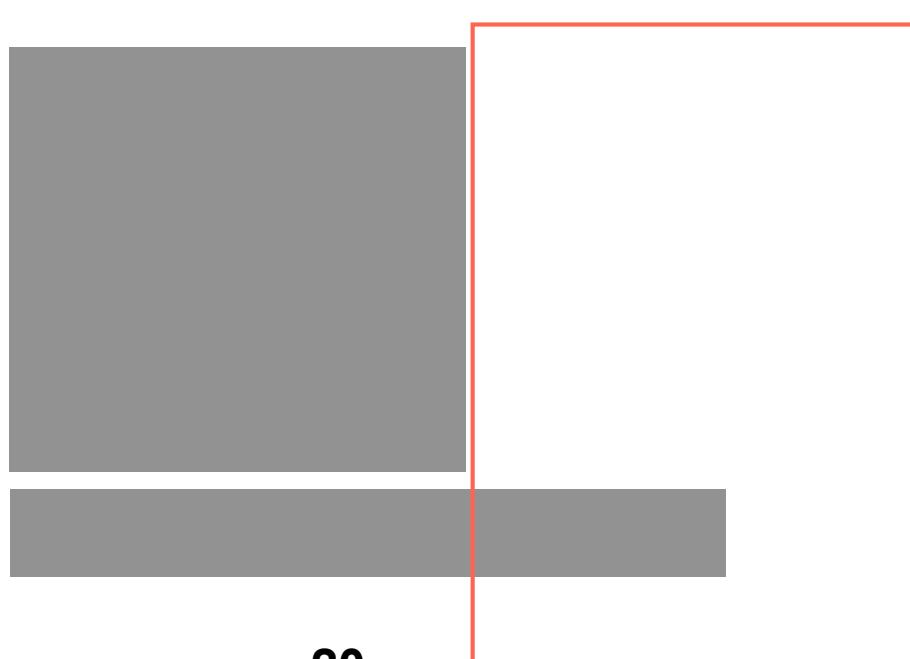
在这种情况下用 BN
就回到了第一个问题

9

5

1

20



为什么使用layer-norm

理解：为什么LayerNorm单独对一个样本的所有单词做缩放可以起到效果。

把BN引申到RNN

NLP中，对齐位置的语义相关性不大
因此 BN 效果不好

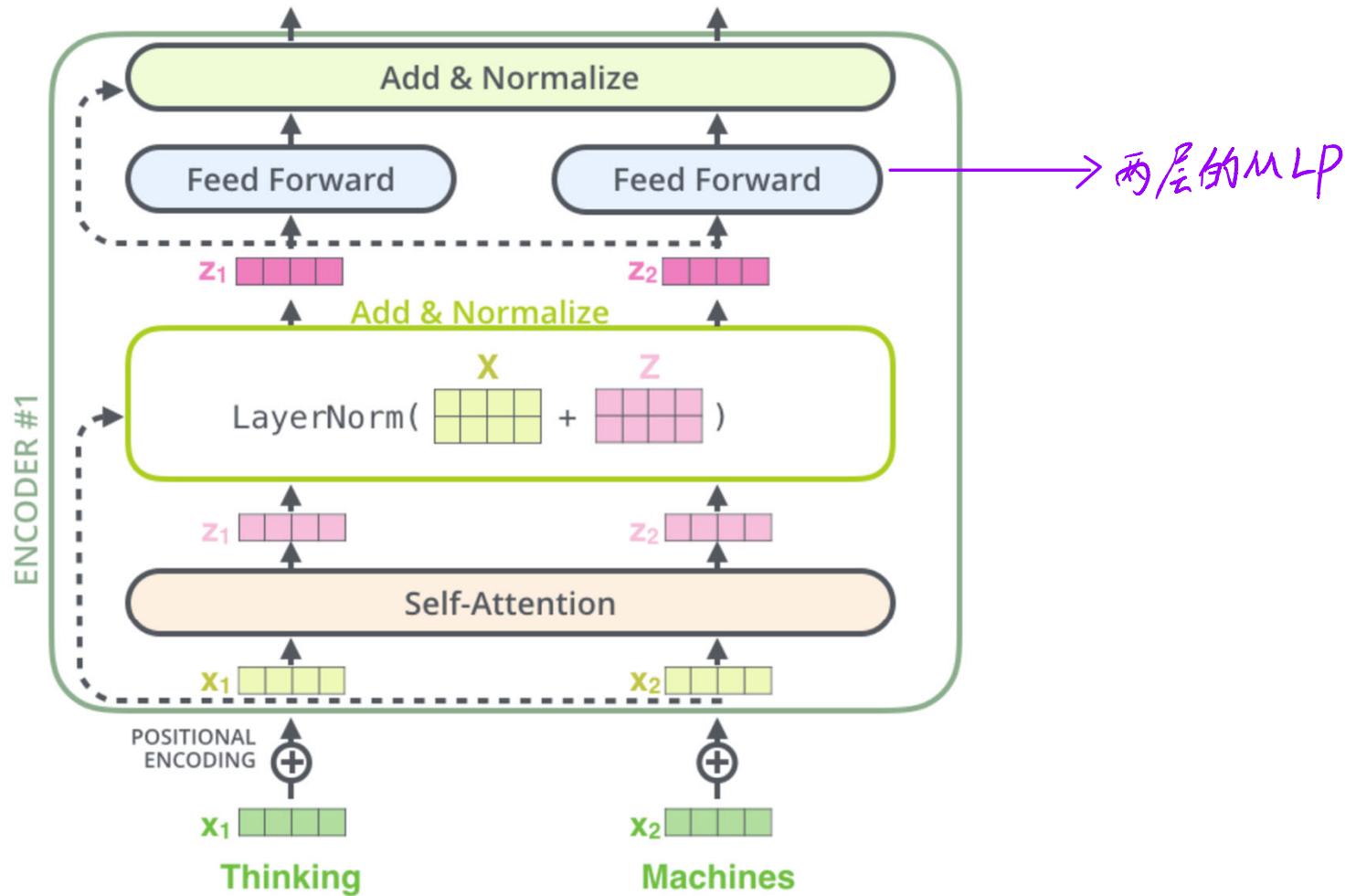
我爱中国共产党



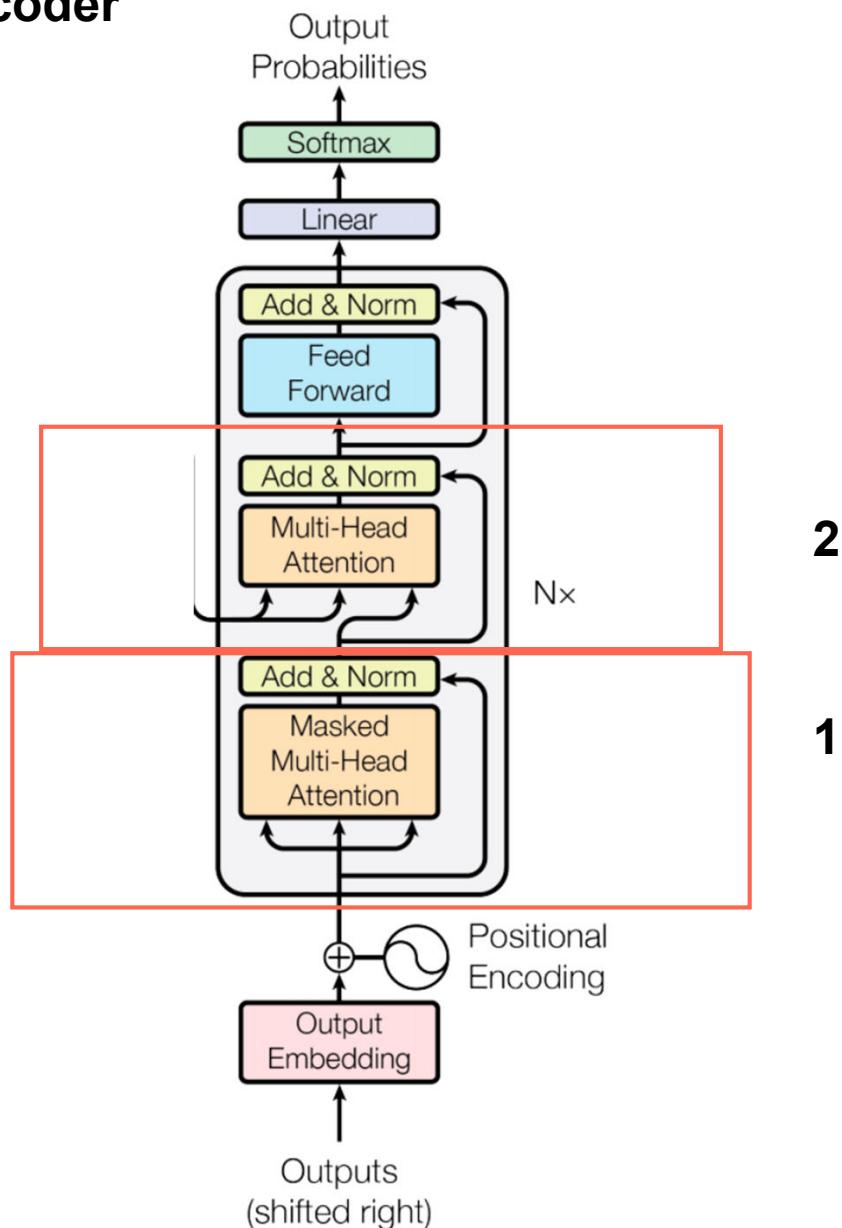
今天天气真不错

同一语句中，相关性更大

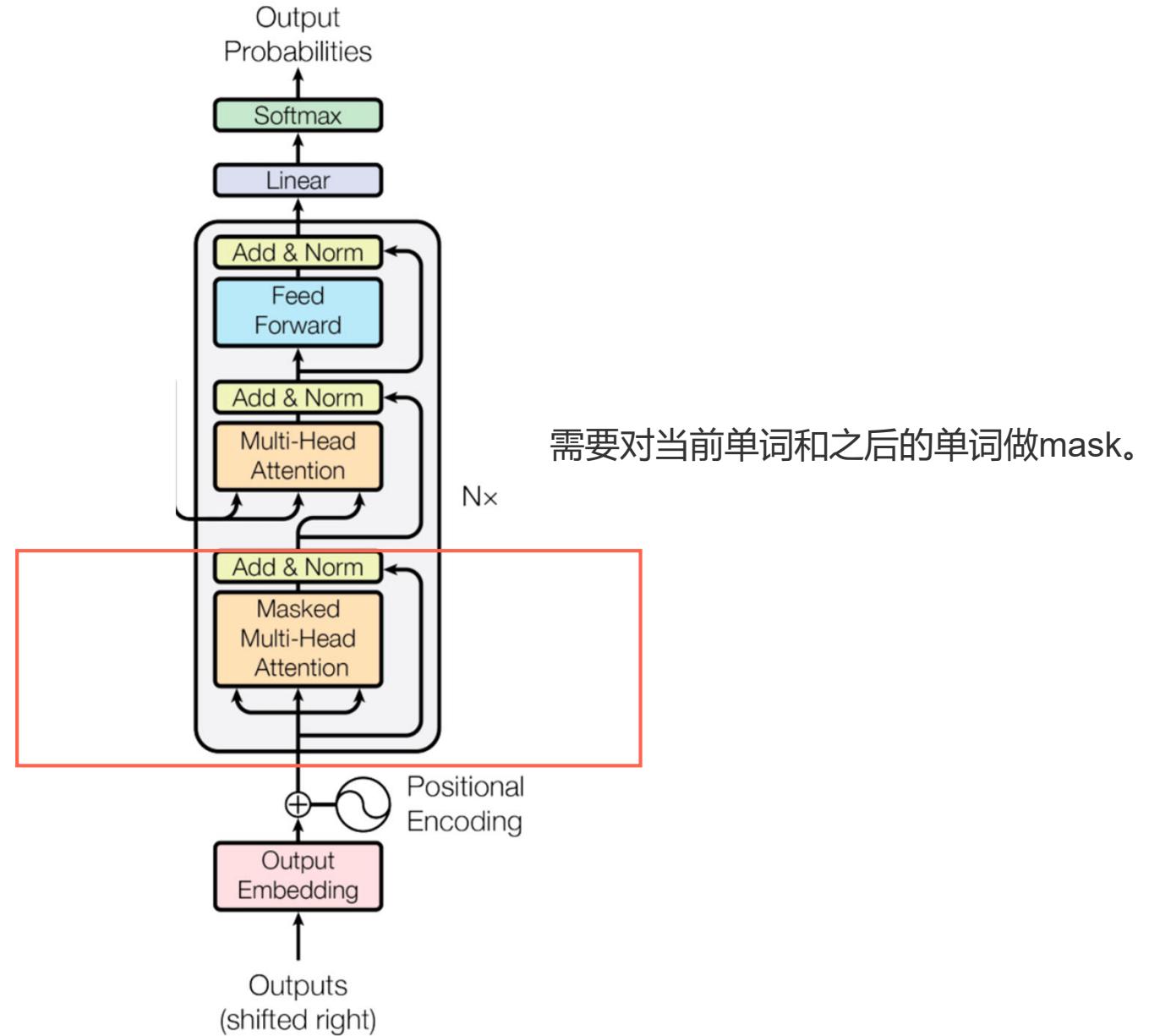
前馈神经网路



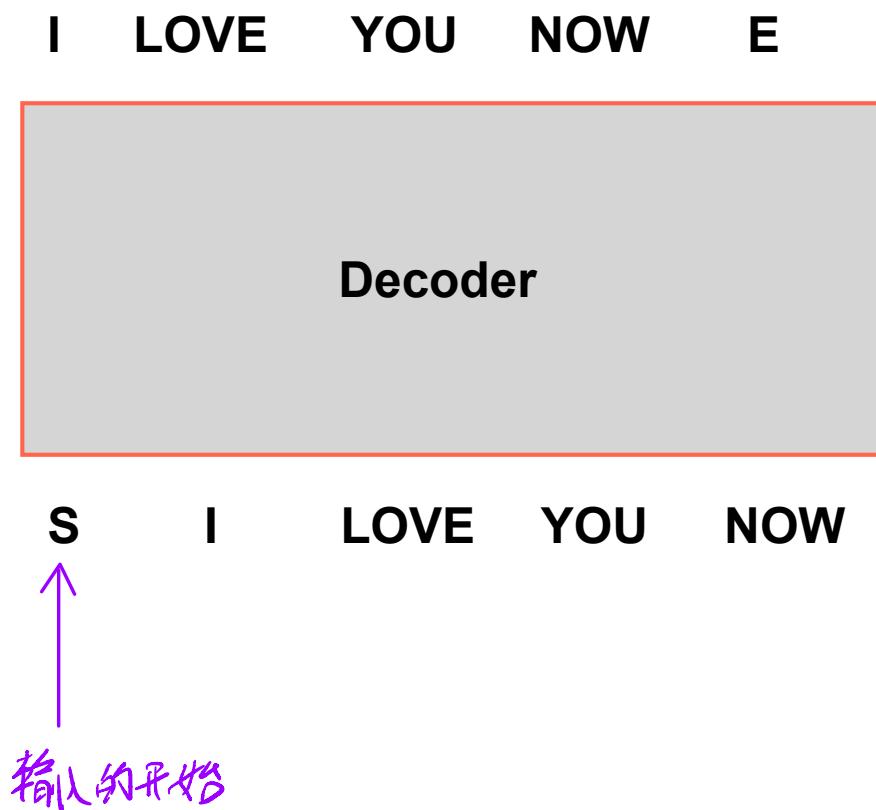
Decoder



1: 多头注意力机制

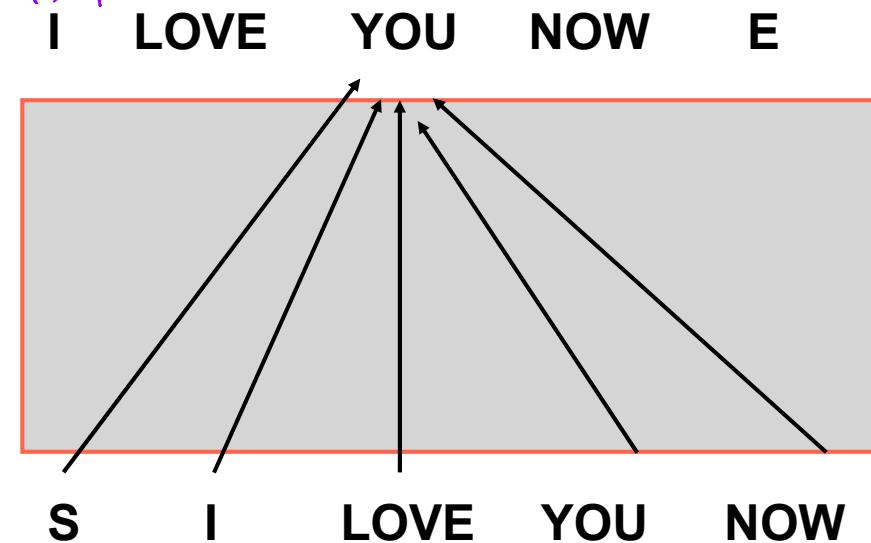


为什么需要mask

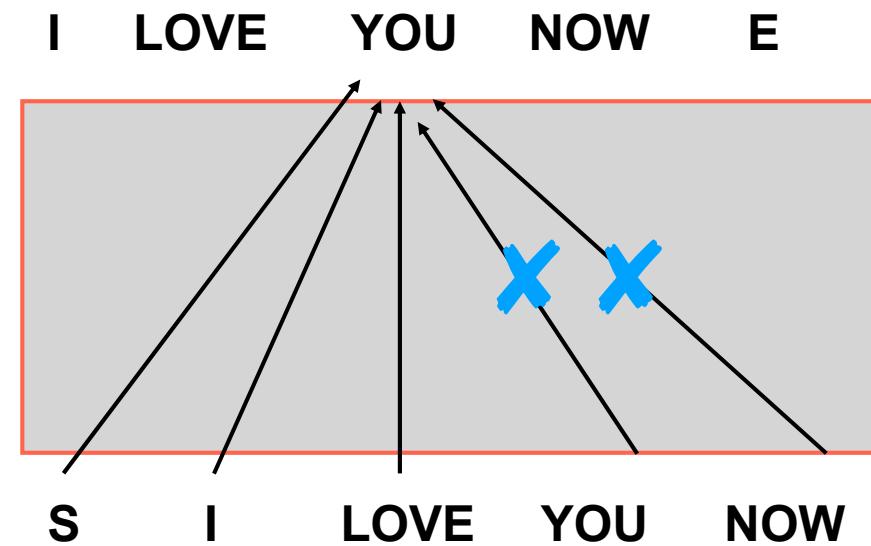


为什么需要mask

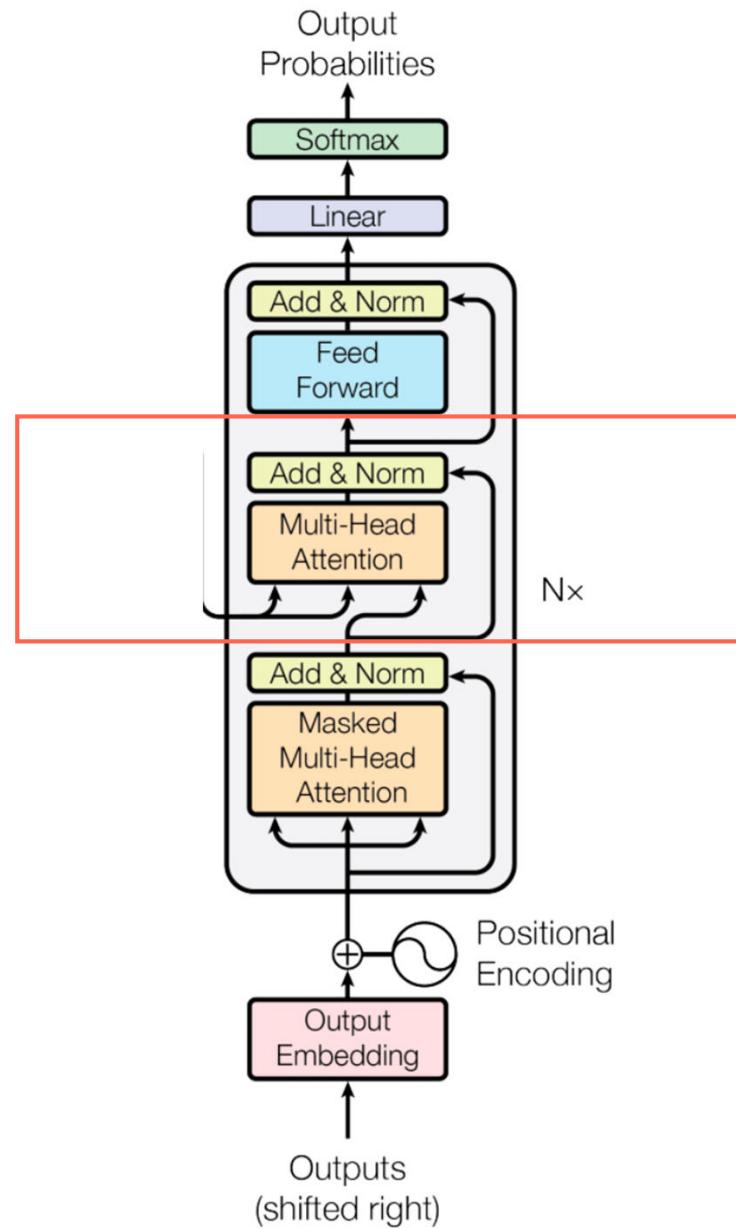
预测阶段是无法知道序列后半部分的信息。
需要用 Mask 来模拟

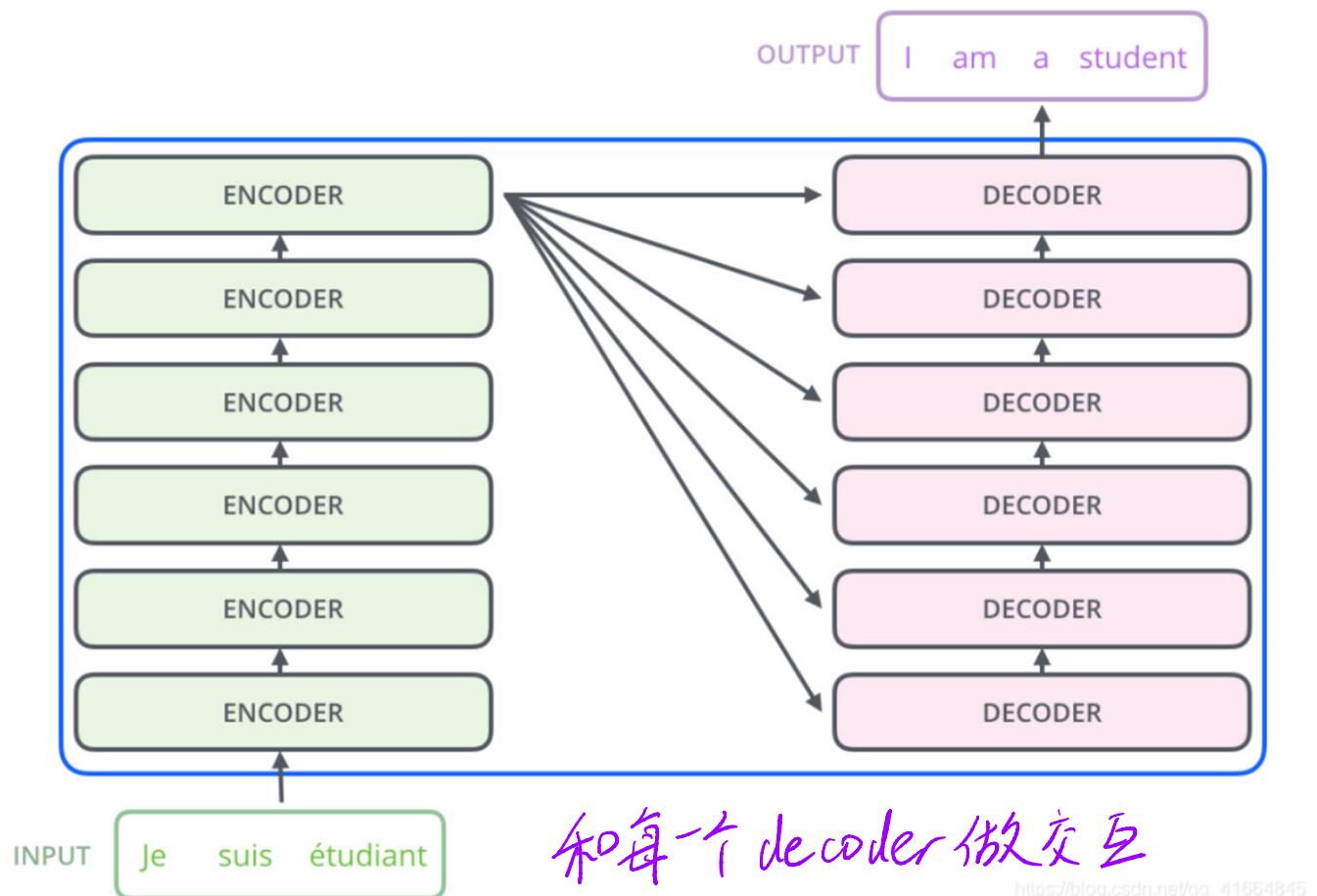


为什么需要mask



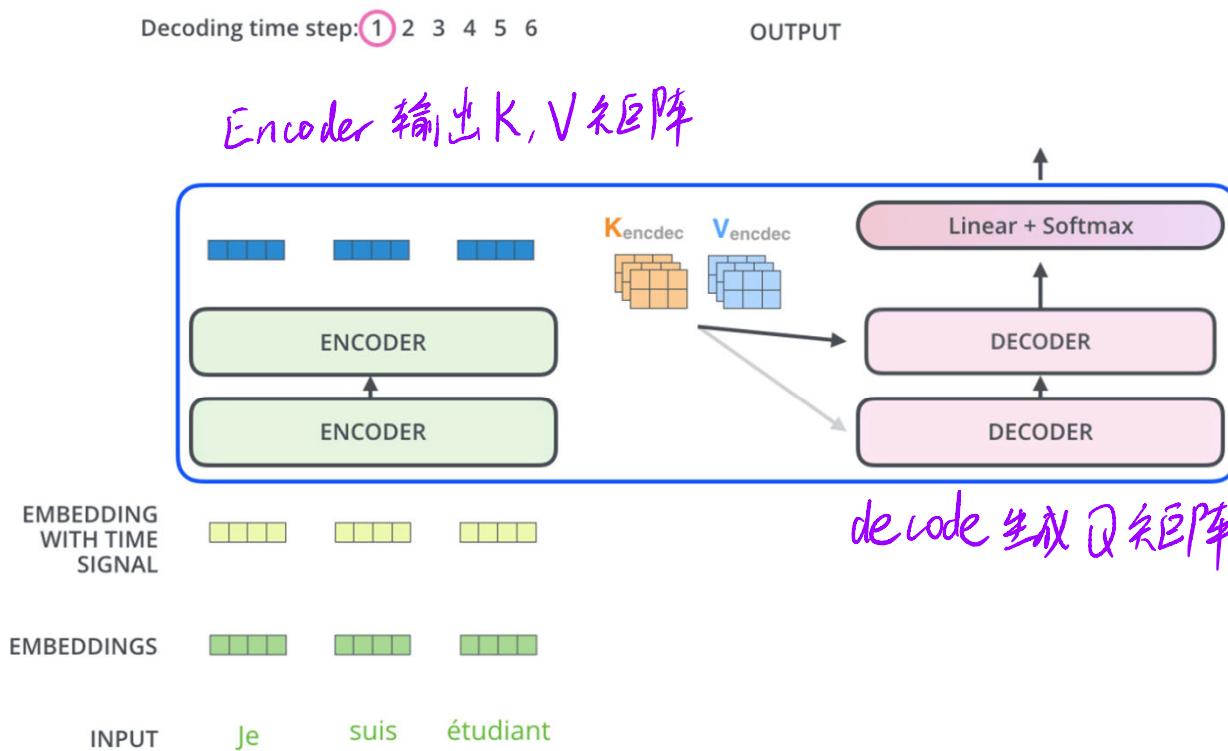
交互层

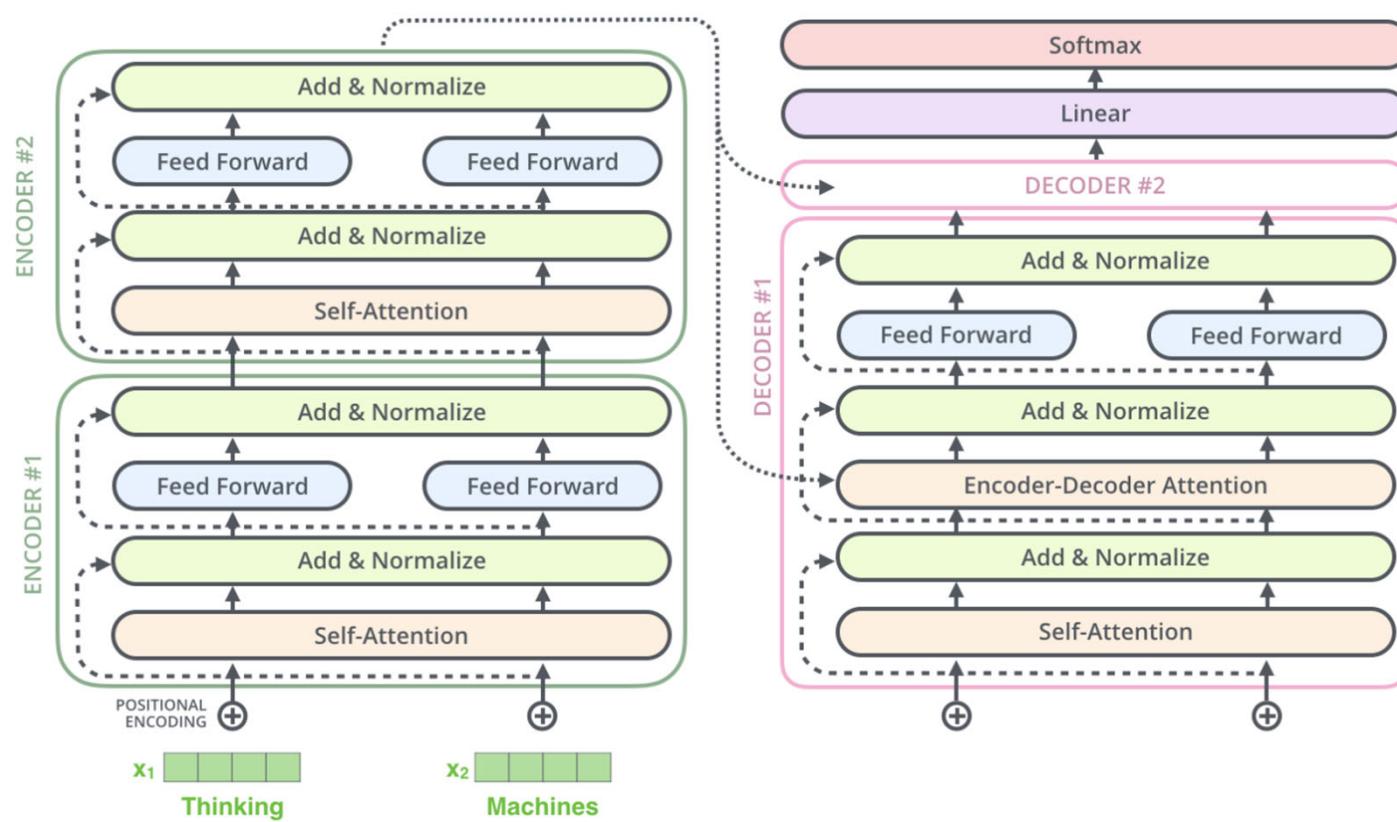




https://blog.csdn.net/qq_41664845

Decoder





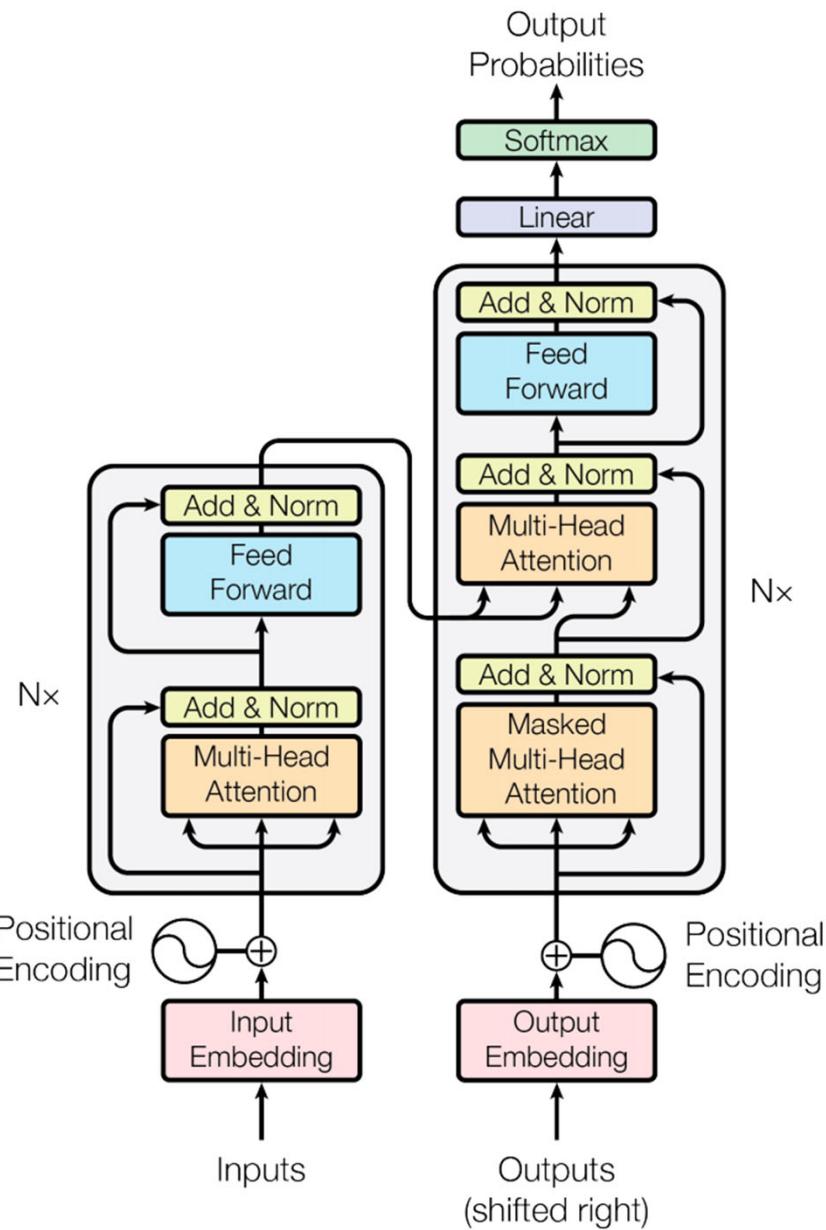
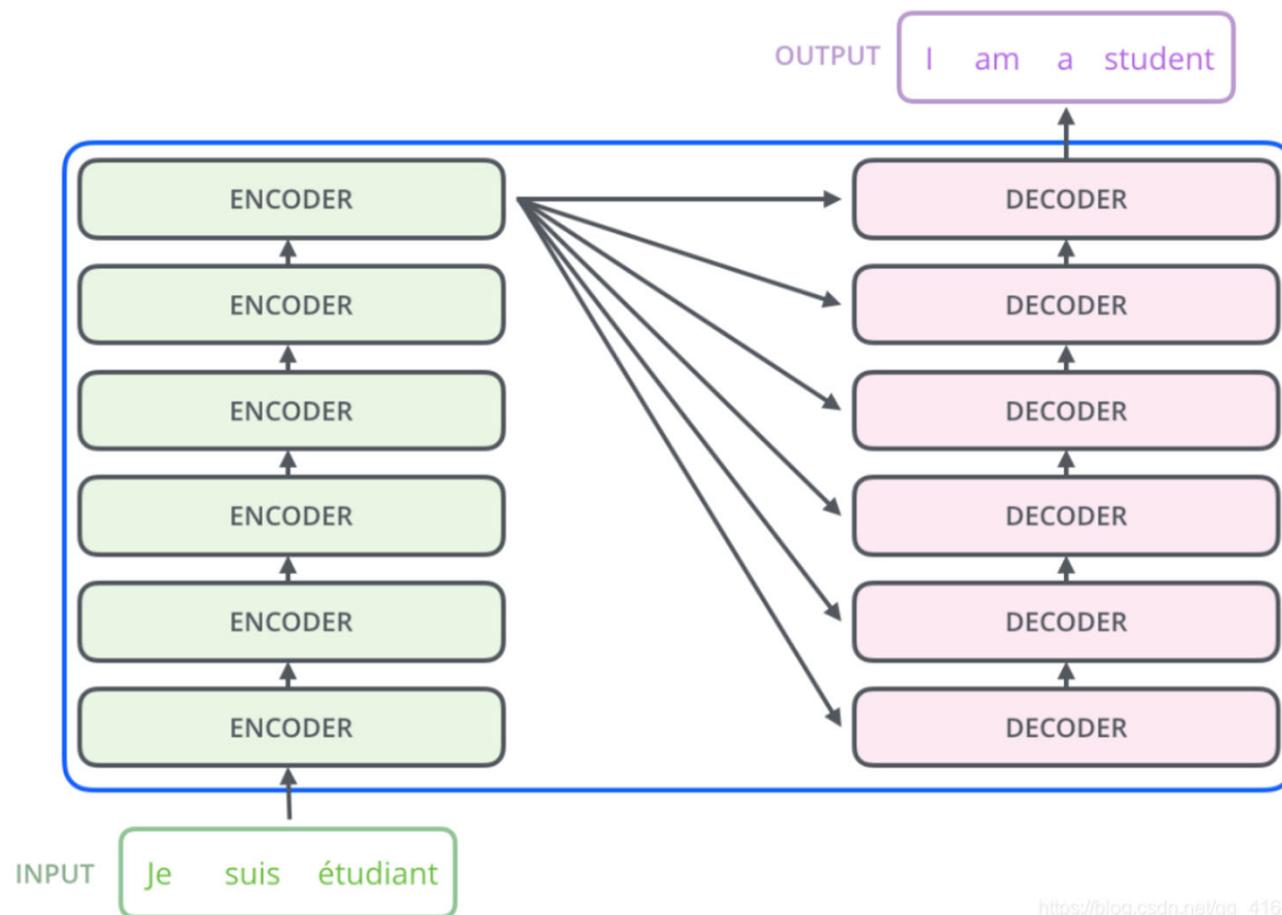
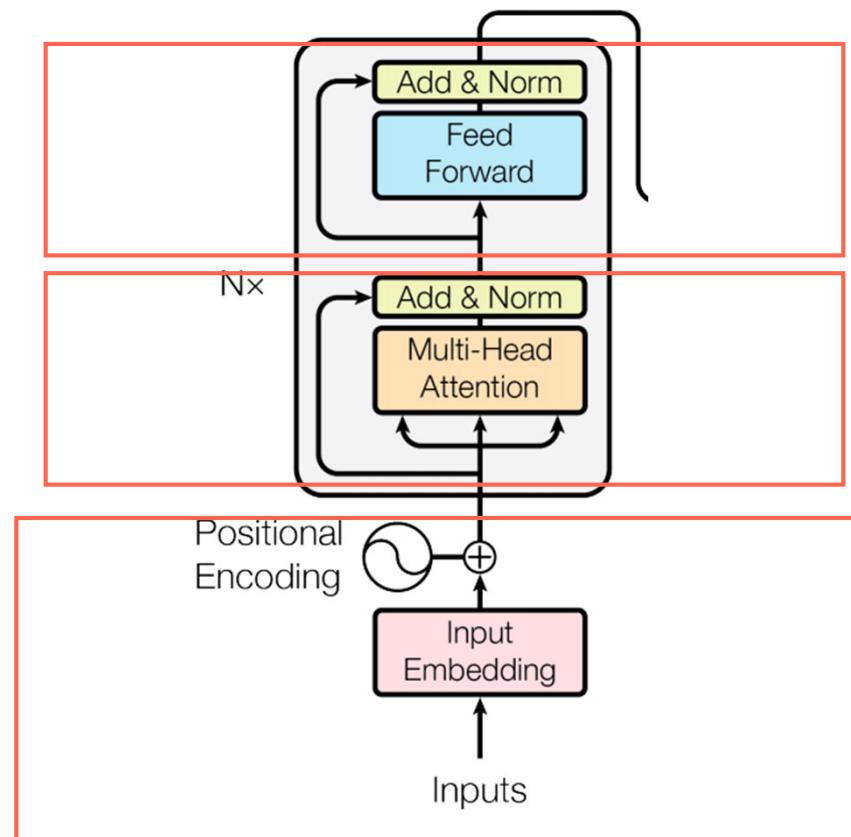


Figure 1: The Transformer - model architecture.



https://blog.csdn.net/qq_41664845



3 前馈神经网络

2 注意力机制

1 输入部分

完