



中国科学院大学

University of Chinese Academy of Sciences

自然语言处理

第4讲 注意力机制

王石 资康莉 刘瑜

2026年春季课程

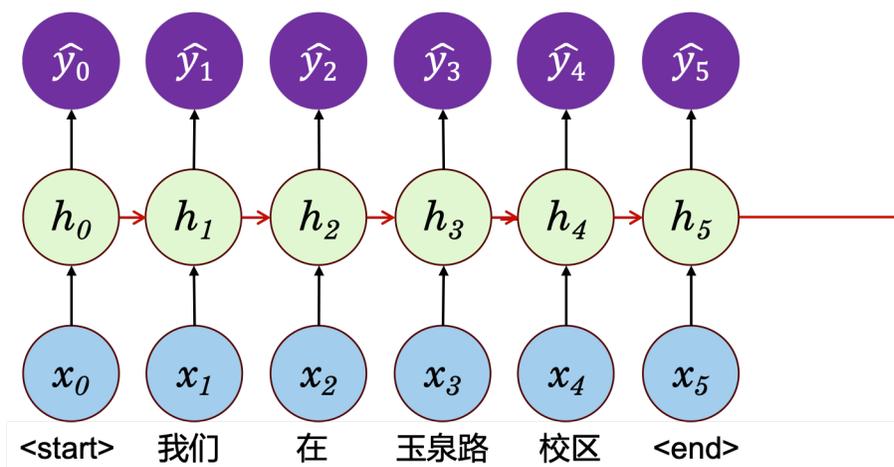
<https://ictkc.github.io/teaching/>



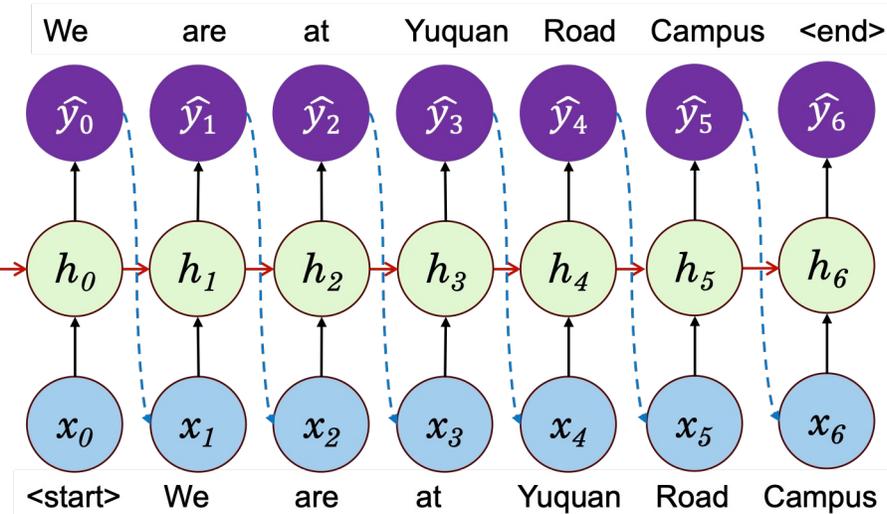
第四讲 注意力机制

Encoder-Decoder原理

□ 先理解，再表达

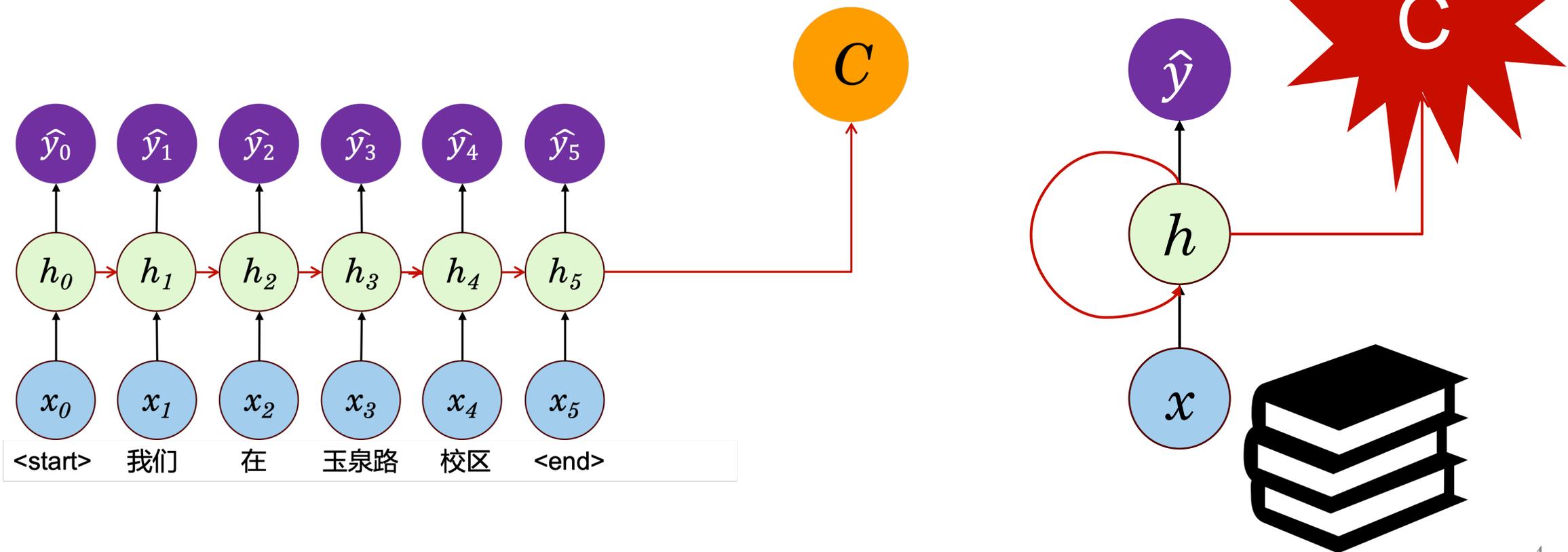


C



Encoder-Decoder缺点

□ 区区一个固定长度的 C ，能容下多少信息？



C的作用：把输入所有信息压缩存储

the late Emperor was taken from us before he could finish his life's work, the restoration of the Han. Today, the empire is still divided in three, and our very survival is threatened. Yet still, the officials at court and the soldiers throughout the realm remain loyal to you, your majesty. Because they remember the late emperor, all of them, and they wish to repay his kindness in service to you. This is the moment to extend your divine influence, to honor the memory of the late Emperor and strengthen the morale of your officers. It is not the time to listen to bad advice or close your ears to the suggestions of loyal men. The emperors of the Western Han chose their courtiers wisely, and their dynasty flourished.

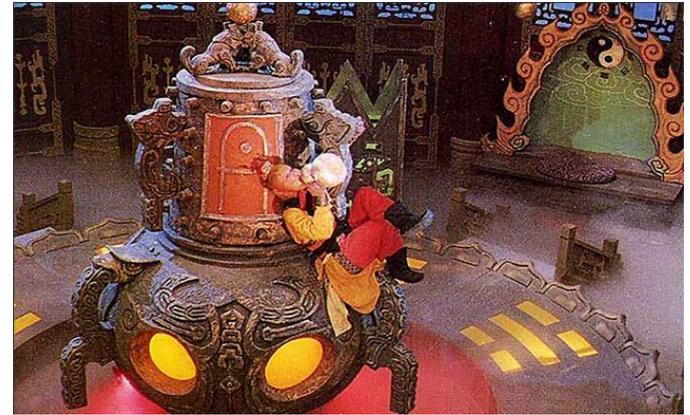


先帝创业未半而中道崩殒，今天下三分，益州疲弊，此诚危急存亡之秋也。然侍卫之臣不懈于内，忠志之士忘身于外者，盖追先帝之殊遇，欲报之于陛下也。诚宜开张圣听，以光先帝遗德，恢弘志士之气，不宜妄自菲薄，引喻失义，以塞忠谏之路也。

其他NLP任务还有更长的上下文

第七回 八卦炉中逃大圣 五行山下定心猿

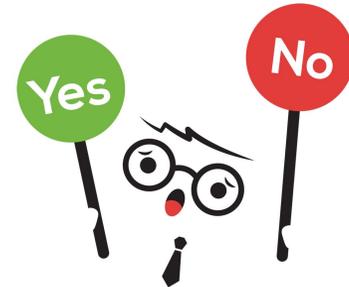
……忽一日，开炉取丹。那大圣双手捂着眼，正自搓揉流涕，只听得炉头声响，猛睁睛看见光明，他就忍不住将身一纵，跳出丹炉，唵喇一声，**蹬倒八卦炉**，往外就走……



C

第六十回 牛魔王罢战赴华筵 孙行者二调芭蕉扇

……土地道：“这火原是大圣放的。”行者怒道：“我在那里，你这等乱谈！**我可是放火之辈？**”



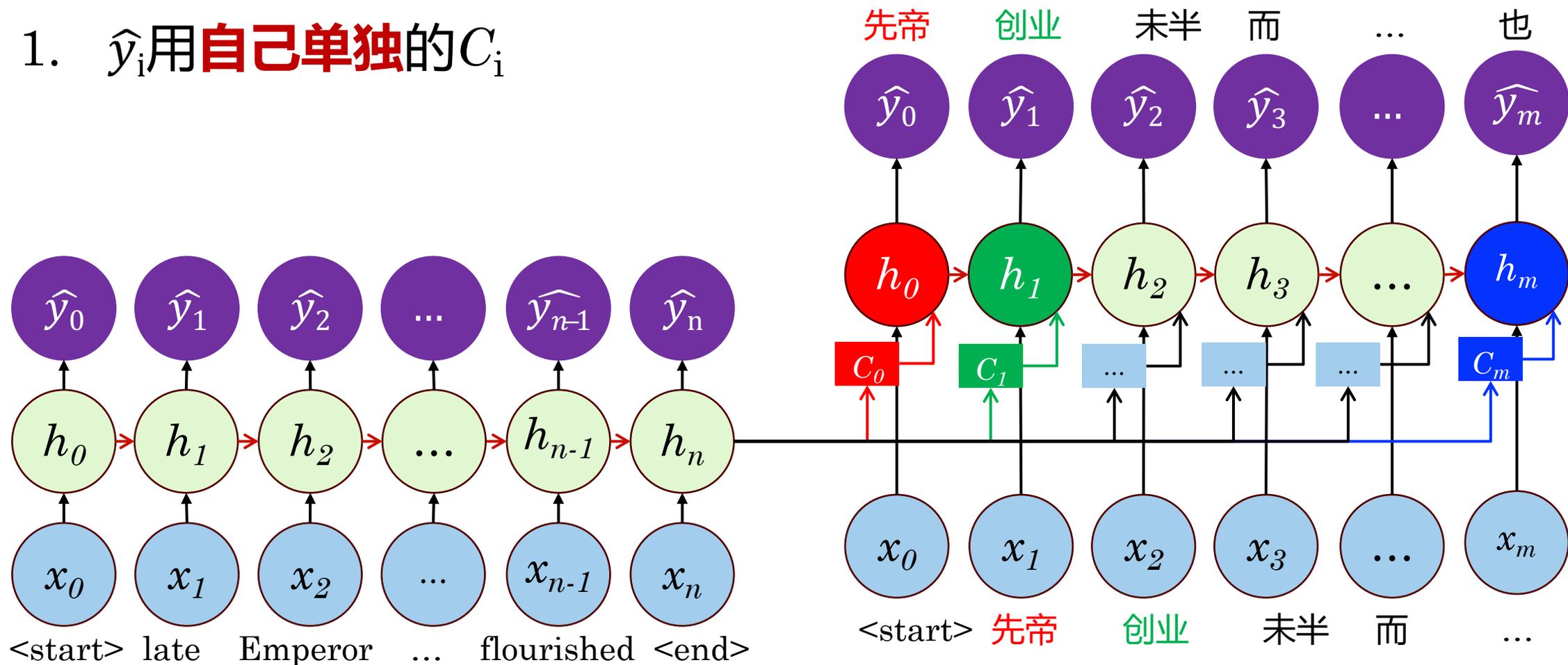


目 录

- 1 带注意力的Encoder-Decoder
- 2
- 3
- 4

如果一个C解决不了，那就2个

1. \hat{y}_i 用**自己单独**的 C_i



注意力机制： C_i 需多注意和翻译它有关的词

the late Emperor was taken from us before he could finish his life's work, the restoration of the Han. Today, the empire is still divided in three, and our very survival is threatened. Yet still, the officials at court and the soldiers throughout the realm remain loyal to you, your majesty. Because they remember the late emperor, all of them, and they wish to repay his kindness in service to you. This is the moment to extend your divine influence, to honor the memory of the late Emperor and strengthen the morale of your officers. It is not the time to listen to bad advice or close your ears to the suggestions of loyal men. The emperors of the Western Han chose their courtiers wisely, and their dynasty flourished.

先帝创业未半而中道崩殂，今天下三分，益州疲弊，此诚危急存亡之秋也。然侍卫之臣不懈于内，忠志之士忘身于外者，盖追先帝之殊遇，欲报之于陛下也。诚宜开张圣听，以光先帝遗德，恢弘志士之气，不宜妄自菲薄，引喻失义，以塞忠谏之路也。

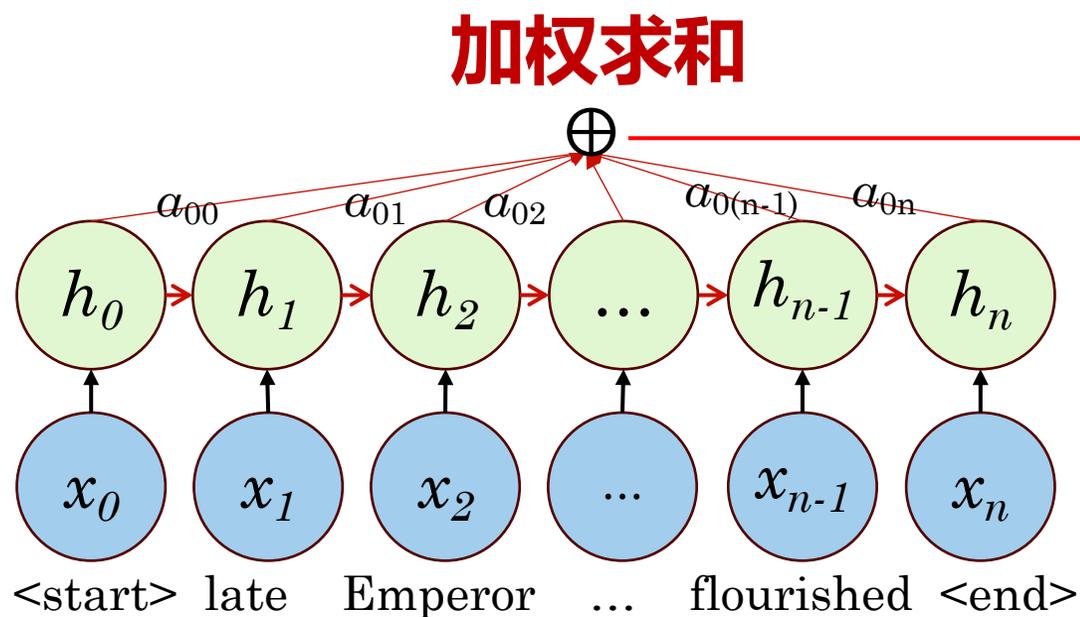
注意：不是“只注意”，是“多注意”

the late Emperor was taken from us before he could finish his life's work, the restoration of the Han. Today, the empire is still divided in three, and our very survival is threatened. Yet still, the officials at court and the soldiers throughout the realm remain loyal to you, your majesty. Because they remember the late emperor, all of them, and they wish to repay his kindness in service to you. This is the moment to extend your divine influence, to honor the memory of the late Emperor and strengthen the morale of your officers. It is not the time to listen to bad advice or close your ears to the suggestions of loyal men. The emperors of the Western Han chose their courtiers wisely, and their dynasty flourished.

先帝创业未半而中道崩殂，今天下三分，益州疲弊，此诚危急存亡之秋也。然侍卫之臣不懈于内，忠志之士忘身于外者，盖追先帝之殊遇，欲报之于陛下也。诚宜开张圣听，以光先帝遗德，恢弘志士之气，不宜妄自菲薄，引喻失义，以塞忠谏之路也。

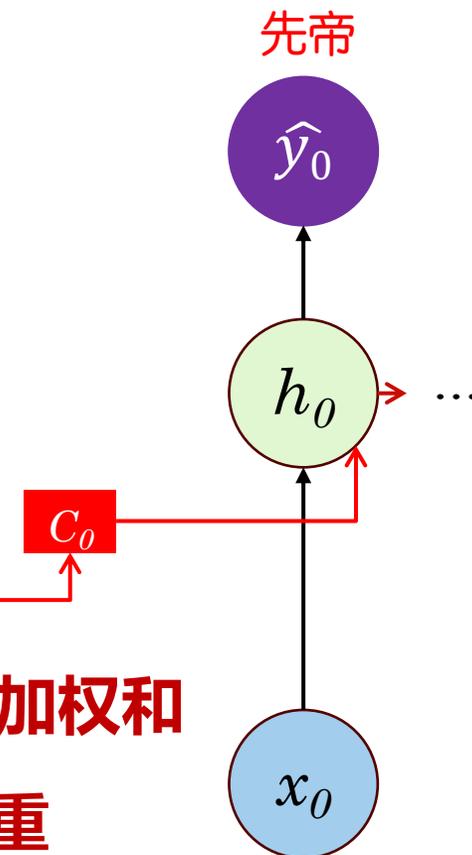
C_i 如何计算

1. 来源于所有的源语言隐变量
2. 不同语言隐变量有不同的权重



$$C_i = \sum_{j=0}^n a_{ij} h_j$$

- C_i 是编码器中隐状态的**加权和**
- a_{ij} 是 x_j 对 \hat{y}_i 的重要性**权重**



示例

Today empire is divided in three

$$\begin{array}{l}
 h_1 * a_{11} + h_2 * a_{12} + h_3 * a_{13} + h_4 * a_{14} + h_5 * a_{15} + h_6 * a_{16} = c_1 \rightarrow \text{今} \\
 h_1 * a_{21} + h_2 * a_{22} + h_3 * a_{23} + h_4 * a_{24} + h_5 * a_{25} + h_6 * a_{26} = c_2 \rightarrow \text{天下} \\
 h_1 * a_{31} + h_2 * a_{32} + h_3 * a_{33} + h_4 * a_{34} + h_5 * a_{15} + h_6 * a_{36} = c_1 \rightarrow \text{三分}
 \end{array}$$

问题： a_{ij} 如何计算？

a_{ij} 如何计算

a_{ij} 特点分析

1. 归一（所有输入对 \hat{y}_i 的权重之和为1， $\sum_{j=0}^n a_{ij} = 1$ ）
2. 与编码器 h_j 、解码器 h'_{i-1} 有关
3. 属于模型参数，需训练产生

$$a_{ij} = \frac{e^{w_{ij}}}{\sum_{k=0}^n e^{w_{ik}}}$$

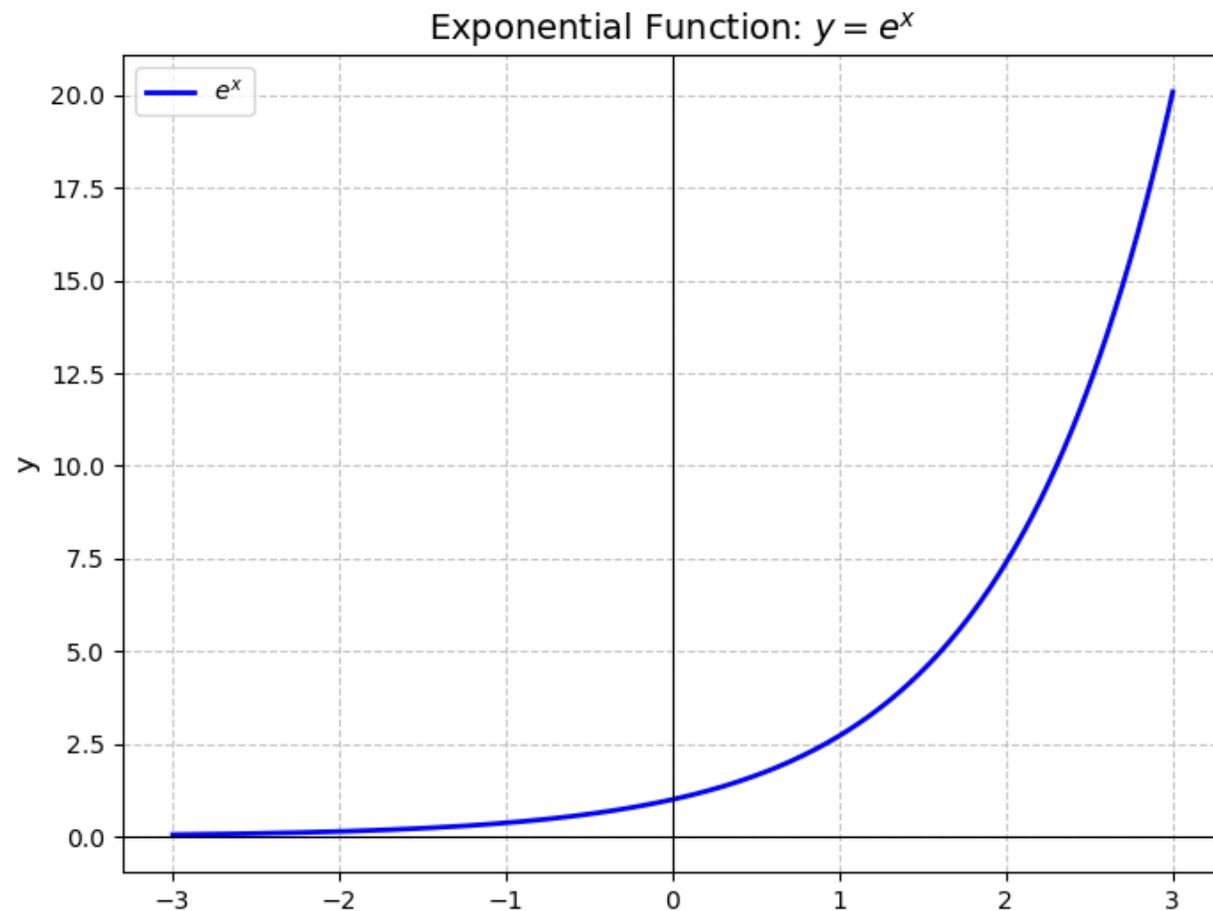
← Softmax归一化

$$w_{ij} = \varphi(h'_{i-1}, h_j)$$

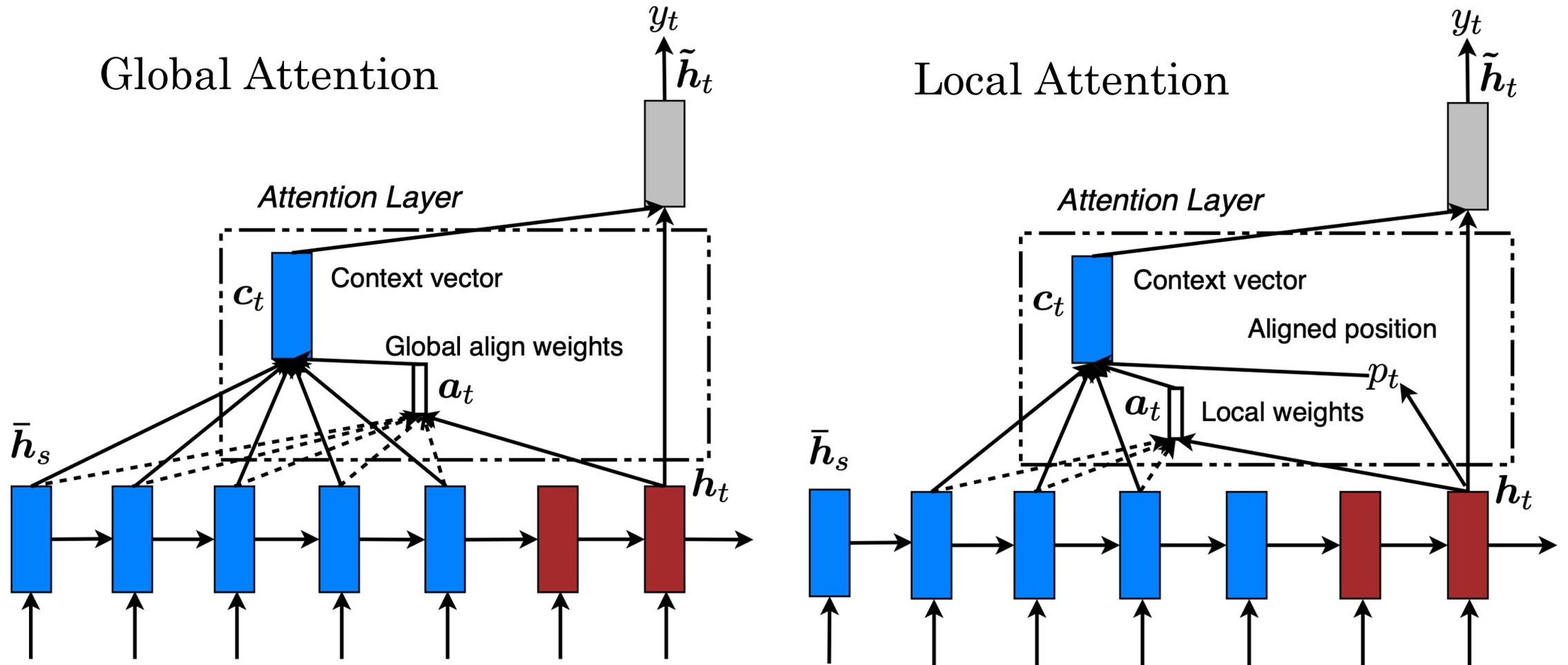
← 权重对齐模型，需训练

补充：Softmax归一化为什么用 e^x ？

□ 保证正值、放大差异、便于求导



Global Attention *v.s.* Local Attention



注意力机制开山之作

[PDF] [Neural machine translation by jointly learning to align and translate](#)

[D Bahdanau](#), [K Cho](#), [Y Bengio](#)

arXiv preprint [arXiv:1409.0473](#), 2014 • [peerj.com](#)

Abstract

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of

展开 ∨

transformer 的前传

☆ 保存  引用 被引用次数: 42458 相关文章 所有 28 个版本 

Dzmitry Bahdanau, Kyunghyun Cho, and **Yoshua Bengio**. Neural machine translation by jointly learning to align and translate, 2014.

基于Attention+双向RNN的机器翻译

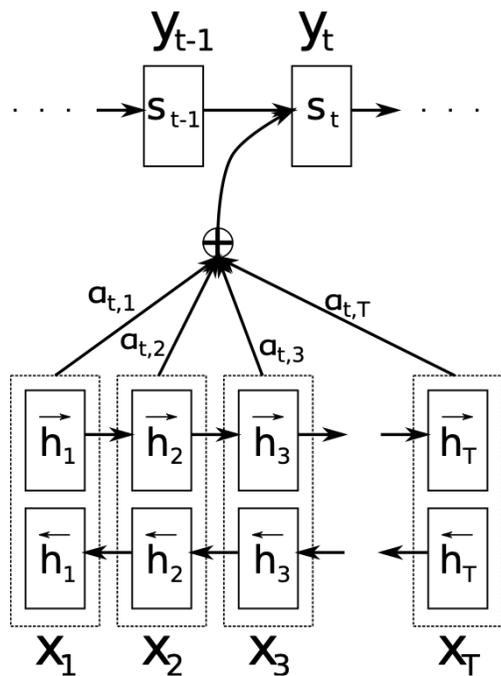
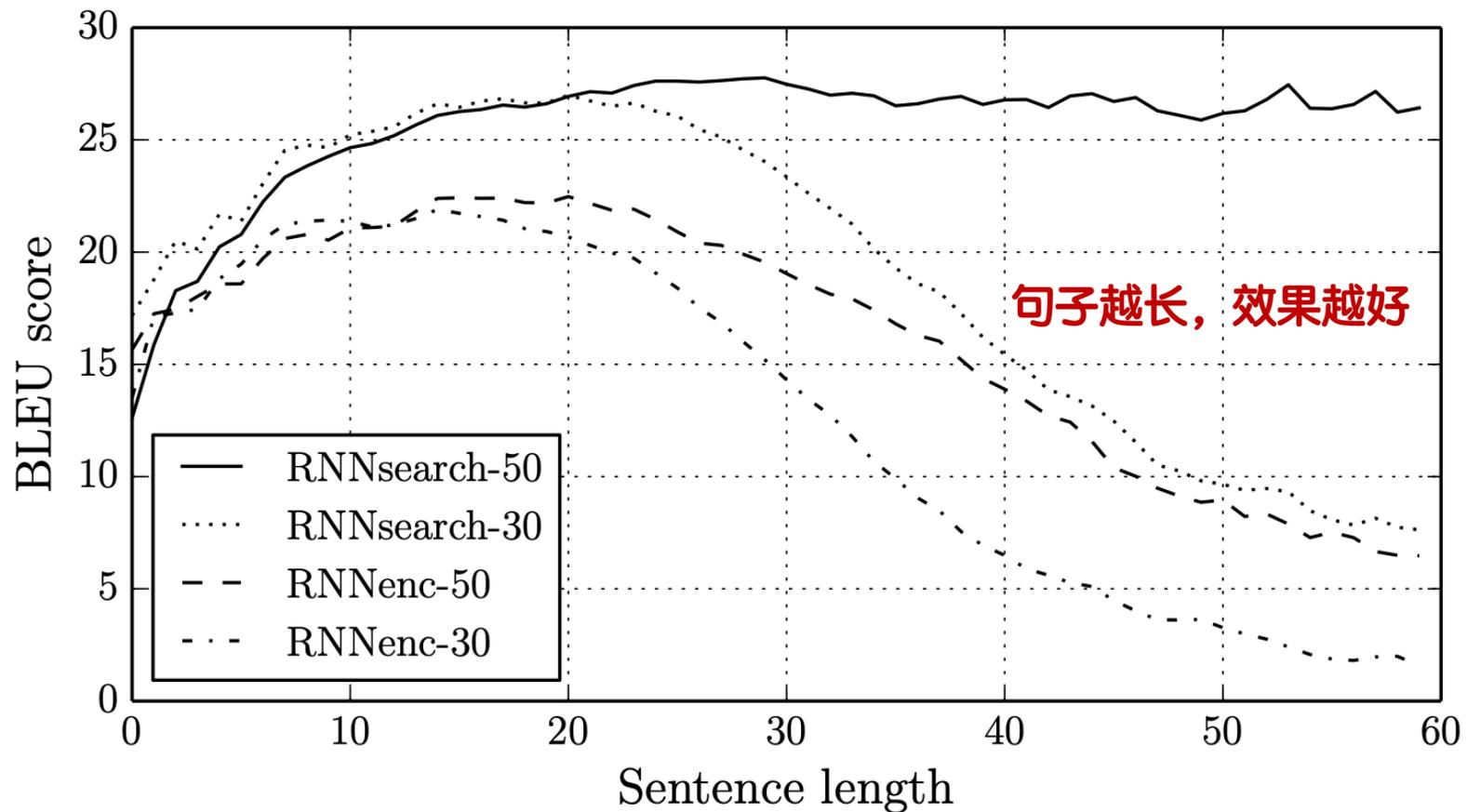
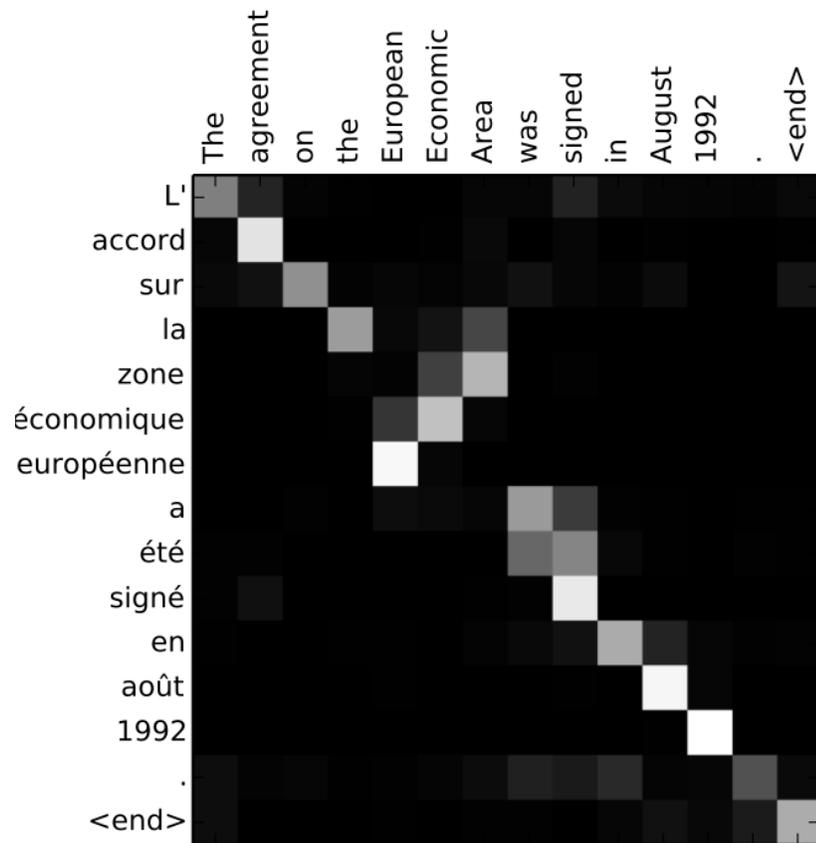


Figure 1: The graphical illustration of the proposed model trying to generate the t -th target word y_t given a source sentence (x_1, x_2, \dots, x_T) .

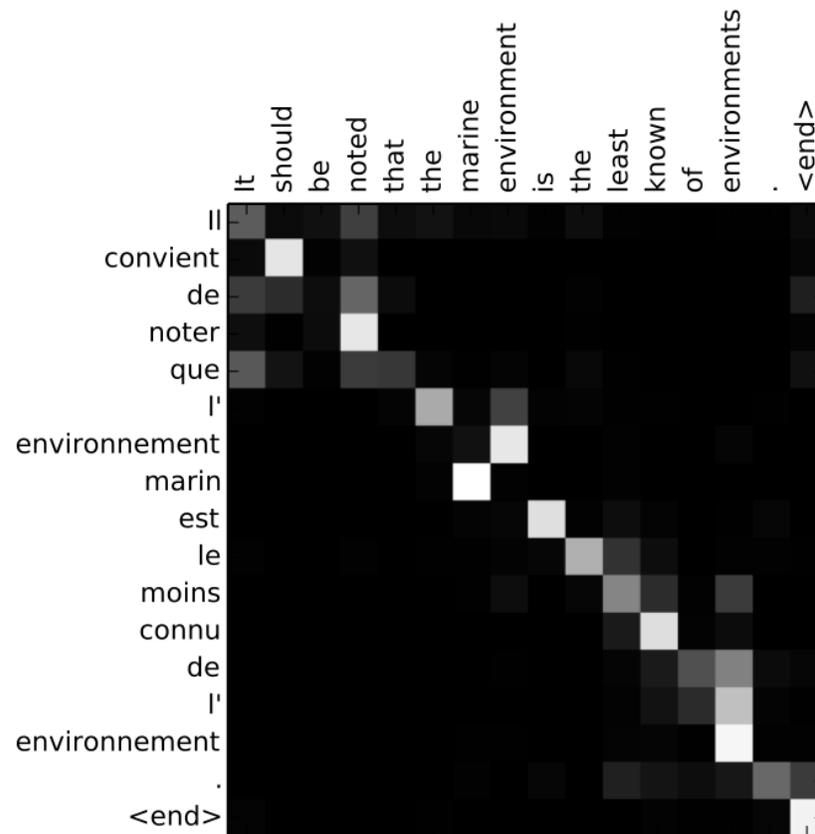


基于Attention+双向RNN的机器翻译

展示注意力机制的双语对齐（英语-法语）效果，像素灰度对齐权重



(a)



(b)

再回到翻译的另一个问题：长距离依赖

the late Emperor was taken from us before he could finish his life's work, the restoration of the Han. Today, the empire is still divided in three, and our very survival is threatened. Yet still, the officials at court and the soldiers throughout the realm remain loyal to you, your majesty. Because they remember the late emperor, all of them, and they wish to repay his kindness in service to you. This is the moment to extend your divine influence, to honor the memory of the late Emperor and strengthen the morale of your officers. It is not the time to listen to bad advice or close your ears to the suggestions of loyal men. The emperors of the Western Han chose their courtiers wisely, and their dynasty flourished.

先帝创业未半而中道崩殒，今天下三分，益州疲弊，此诚危急存亡之秋也。然侍卫之臣不懈于内，忠志之士忘身于外者，盖追先帝之殊遇，欲报之于陛下也。诚宜开张圣听，以光先帝遗德，恢弘志士之气，不宜妄自菲薄，引喻失义，以塞忠谏之路也。

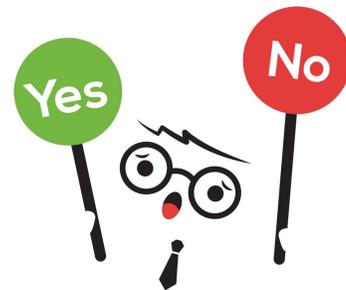
再看其他更长上下文的任务

第七回 八卦炉中逃大圣 五行山下定心猿

……忽一日，开炉取丹。那大圣双手捂着眼，正自搓揉流涕，只听得炉头声响，猛睁睛看见光明，他就忍不住将身一纵，跳出丹炉，唵喇一声，**蹬倒八卦炉**，往外就走……

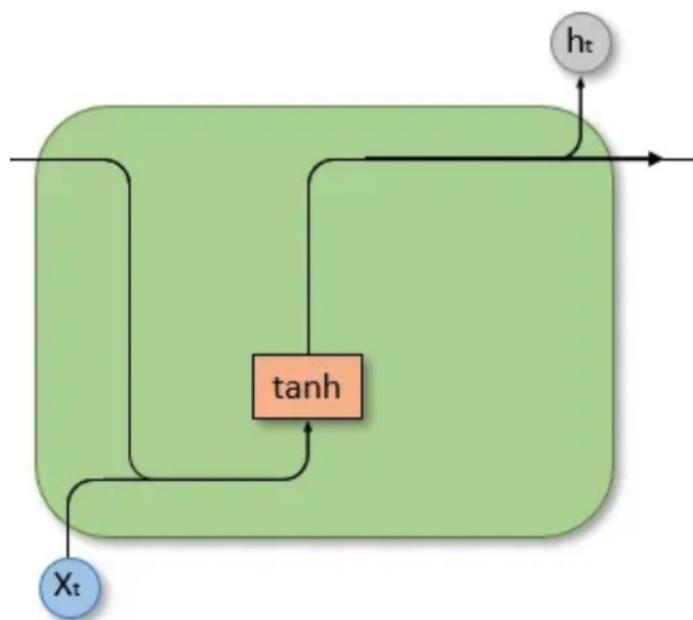


第六十回 牛魔王罢战赴华筵 孙行者二调芭蕉扇
……土地道：“这火原是大圣放的。”行者怒道：“我在那里，你这等乱谈！**我可是放火之辈？**”

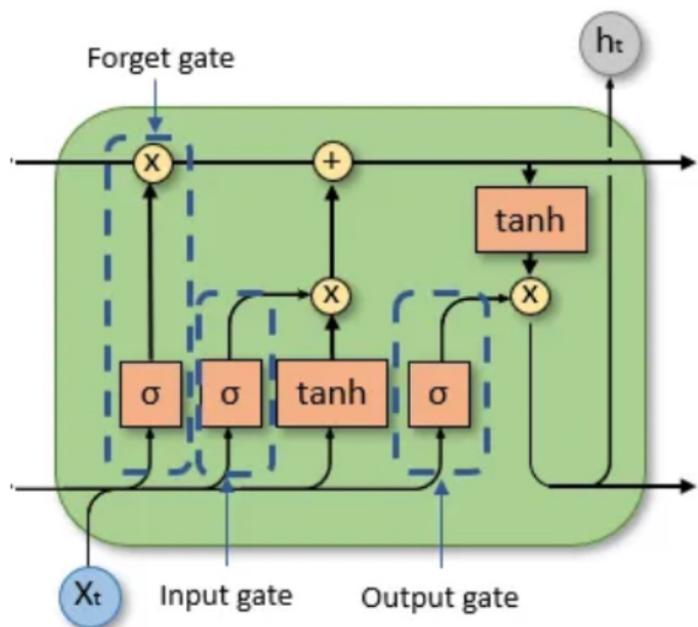


解决办法回顾

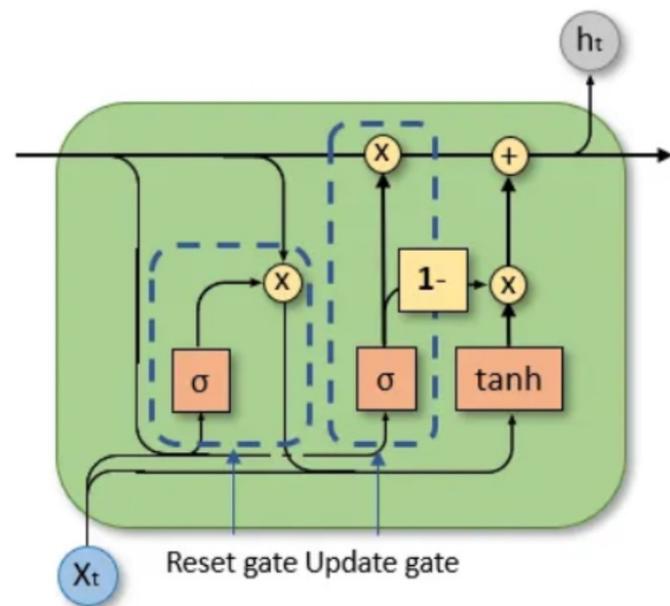
RNN



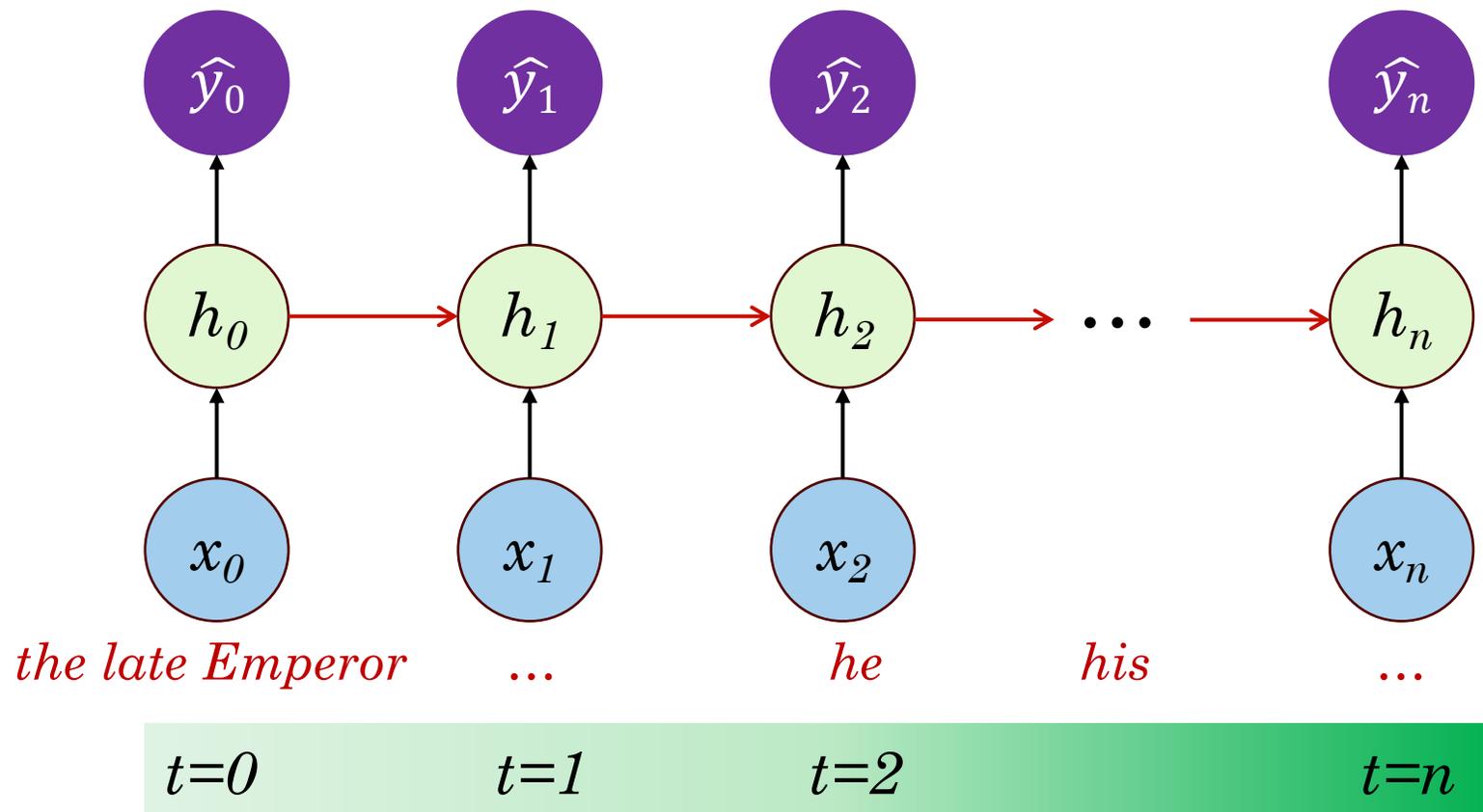
LSTM



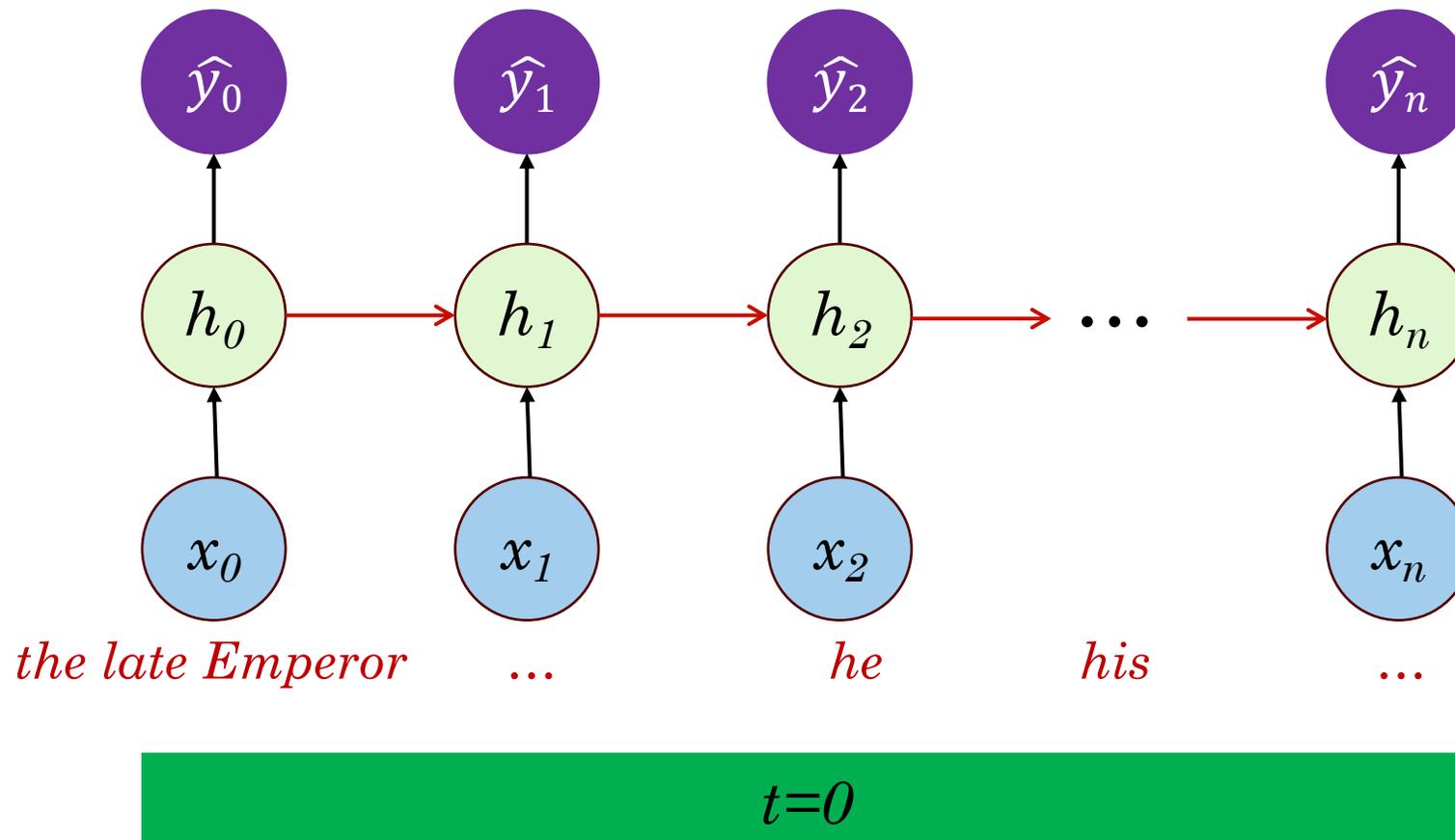
GRU



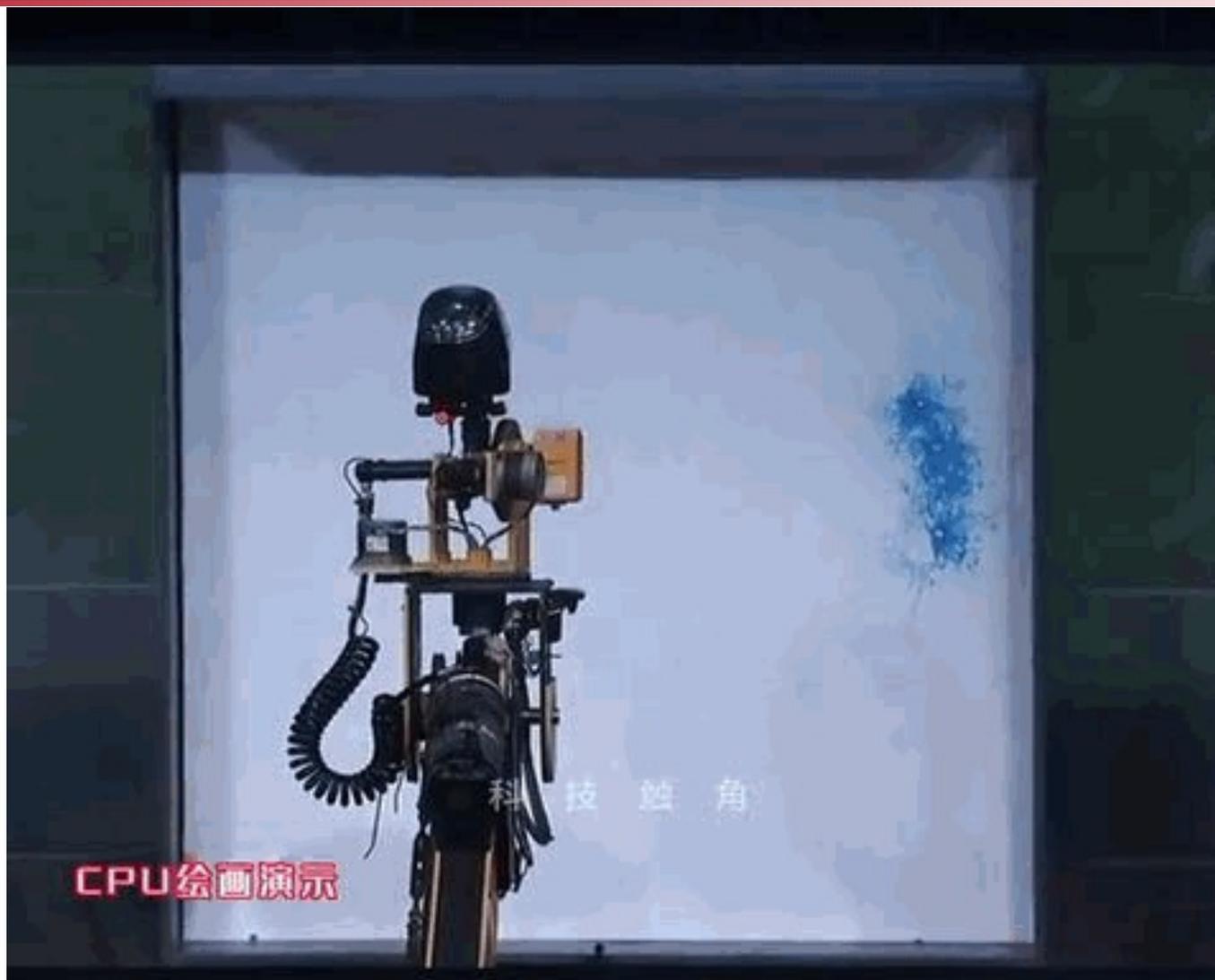
共同存在的问题：串行化，效率低



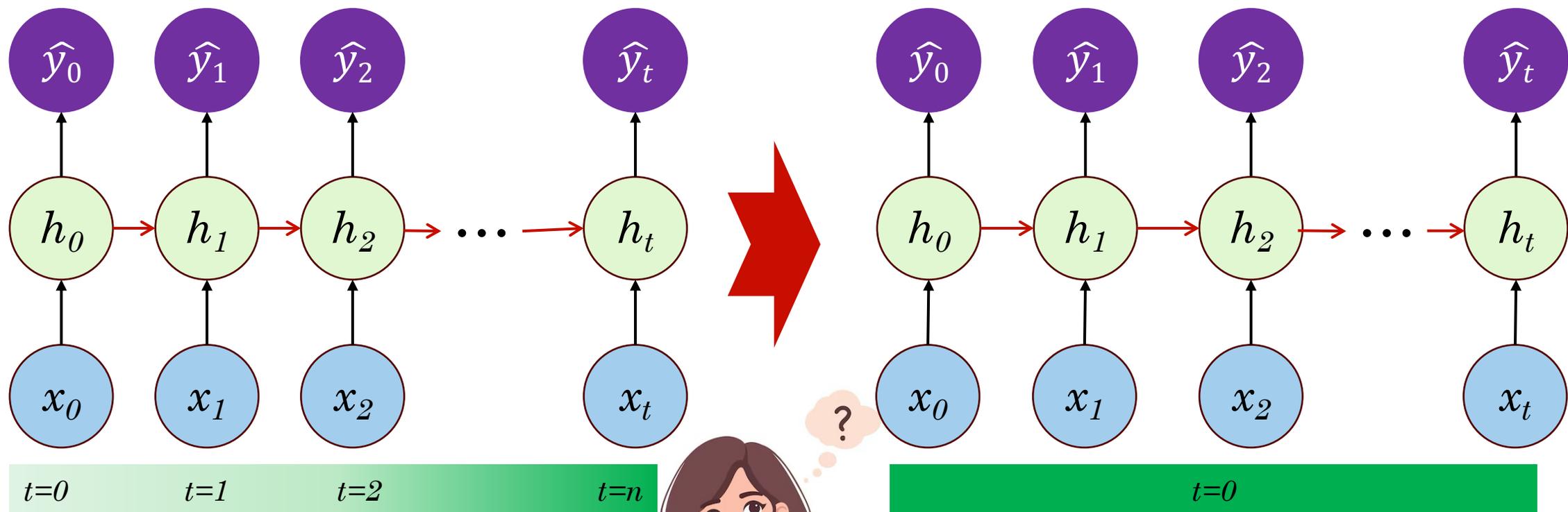
期望的并行化处理



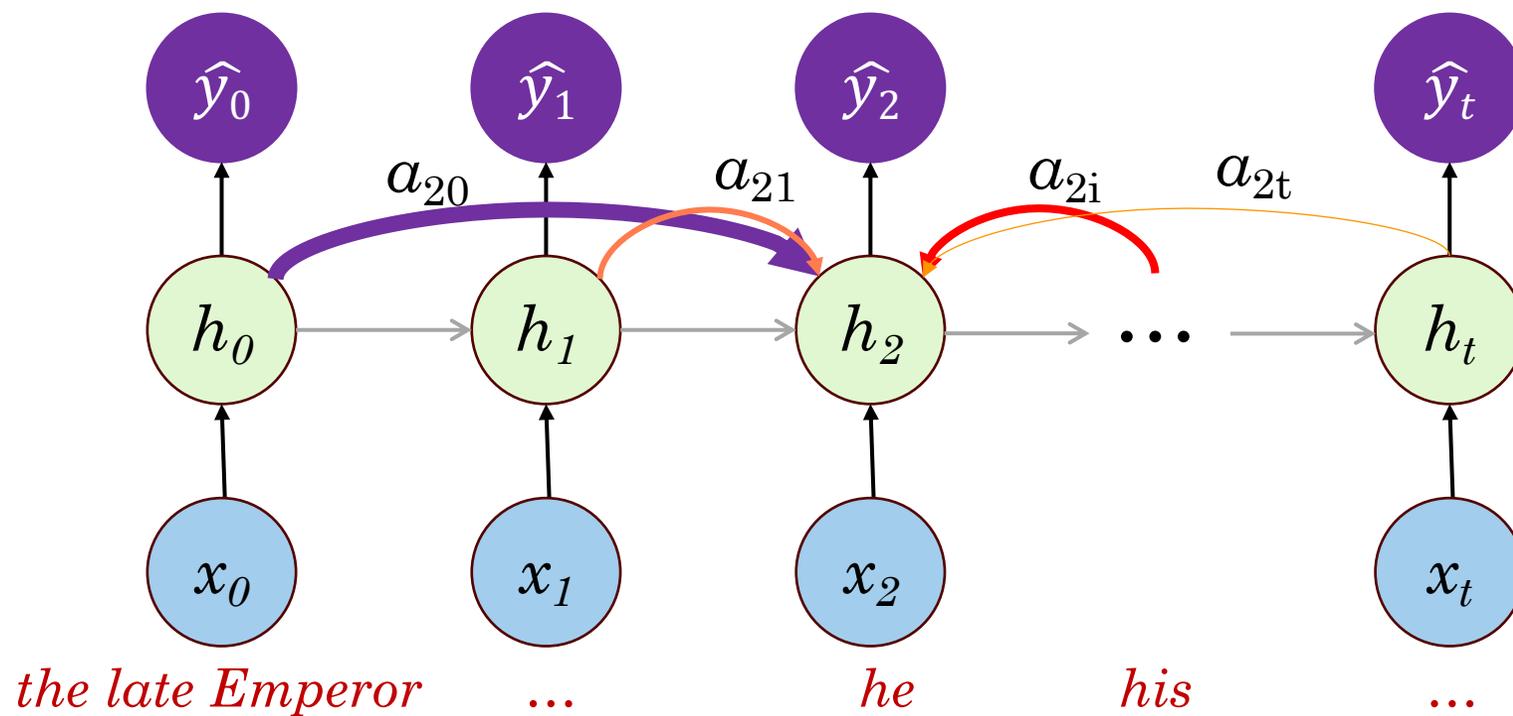
串行 v.s. 并行



串行 v.s. 并行



用注意力视角看上下文



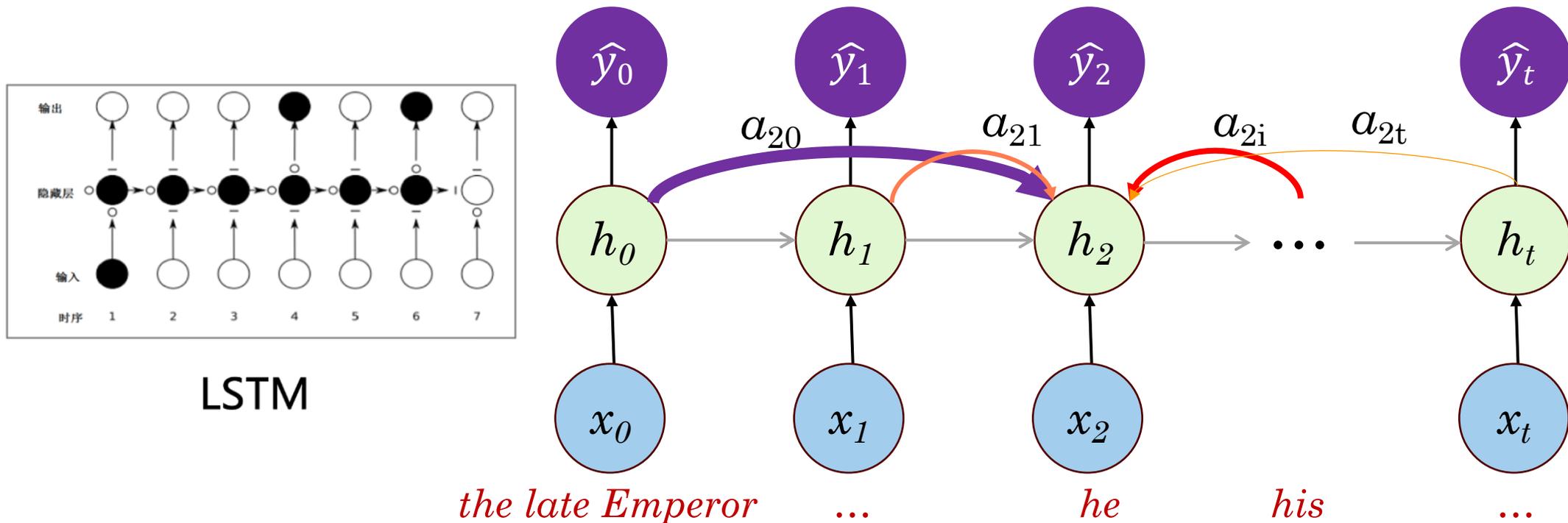


目 录

- 1 带注意力的Encoder-Decoder
- 2 自注意力
- 3
- 4

消除代代相传记忆的一个方法：查读

- 如果每个词都知道“我该注意谁”，那么就不再需要RNN中不断代代相传历史记忆，而是需要时去使用该注意的词的信息即可



LSTM

怎么让每个词都知道“我该注意谁”？



若要每个人都知道“我该注意谁”，那么所有人都要回答

- 我要注意什么？（我**该注意哪些方面，我要用来去找人**）
- 别人要想注意我，该怎么找到我？（**我的标签，好让别人也能要回答**）

学生毕业
身体健康
NLP课程
项目、论文、经费

中科院计算所
知识计算、具身智能
自然语言处理
国科大教师

在NLP中，每个词也要回答

- 我要注意什么？ (Query)
- 别人要想注意我，该怎么找到我？ (Key)

那时候，柔嘉在家里等鸿渐回家来吃晚饭，希望他会跟姑母和好，到她厂里做事

W	Q	K	能获得的信息: V
柔嘉	我要被谁指代?	我是一个女性名字 我是主语	孙柔嘉，女，方鸿渐妻子；长圆脸，微有雀斑，两眼分得太开，使我常带着惊异的表情……
她	我指代谁?	我是一个女性代词	我就是个代词，用来指代女性……

在NLP中，每个词也要回答

- 我要注意什么? (Query)
- 别人要想注意我, 该怎么找到我? (Key)
- 我的信息 (Value)

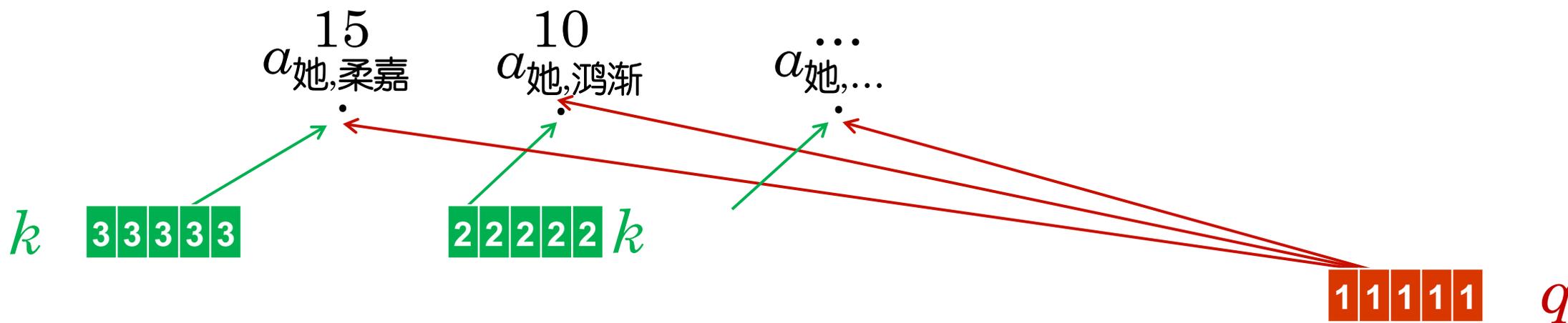
那时候, 柔嘉在家里等鸿渐回家来吃晚饭, 希望他会跟
 姑母和好, 到她厂里做事

均为1*N向量 (N为超参数)

W	Q	K	V
w_i	$q^i =$ 	$k^i =$ 	$v^i =$ 
w_j	$q^j =$ 	$k^j =$ 	$v^j =$ 

寻找该注意的词

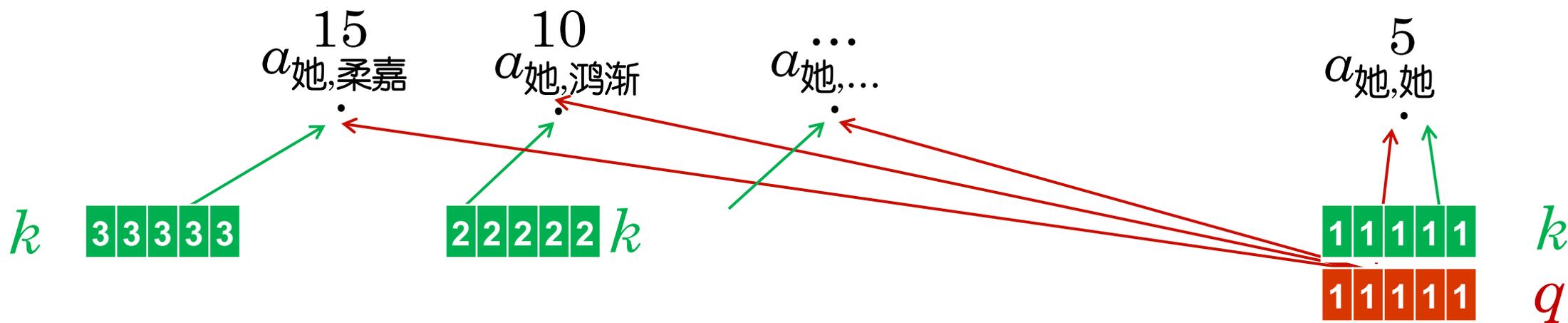
1. w_i 对 w_j 的注意力: $\alpha_{i,j} = q^i \cdot k^j$



那时候, 柔嘉在家里等鸿渐回家来吃晚饭, 希望他会跟姑母和好, 到她厂里做事

寻找该注意的词

2. 自己也不放过: $\alpha_{i,i} = q^i \cdot k^i$

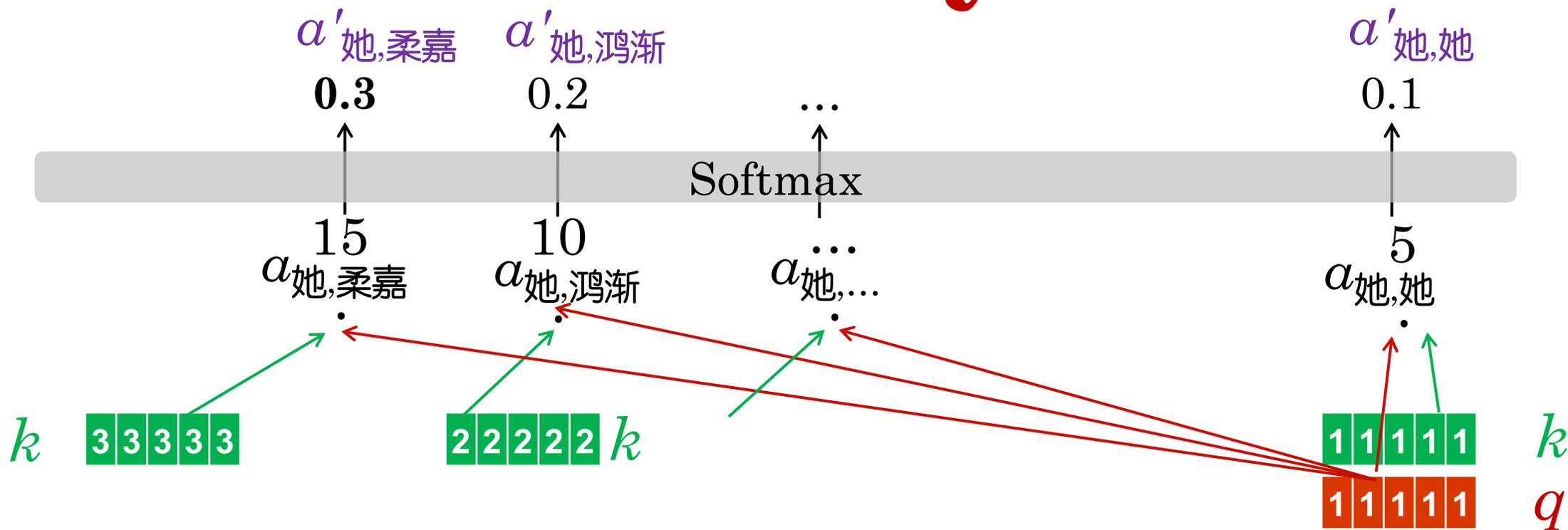


那时候，柔嘉在家里等鸿渐回家来吃晚饭，希望他会跟姑母和好，到她厂里做事

寻找该注意的词

OVER

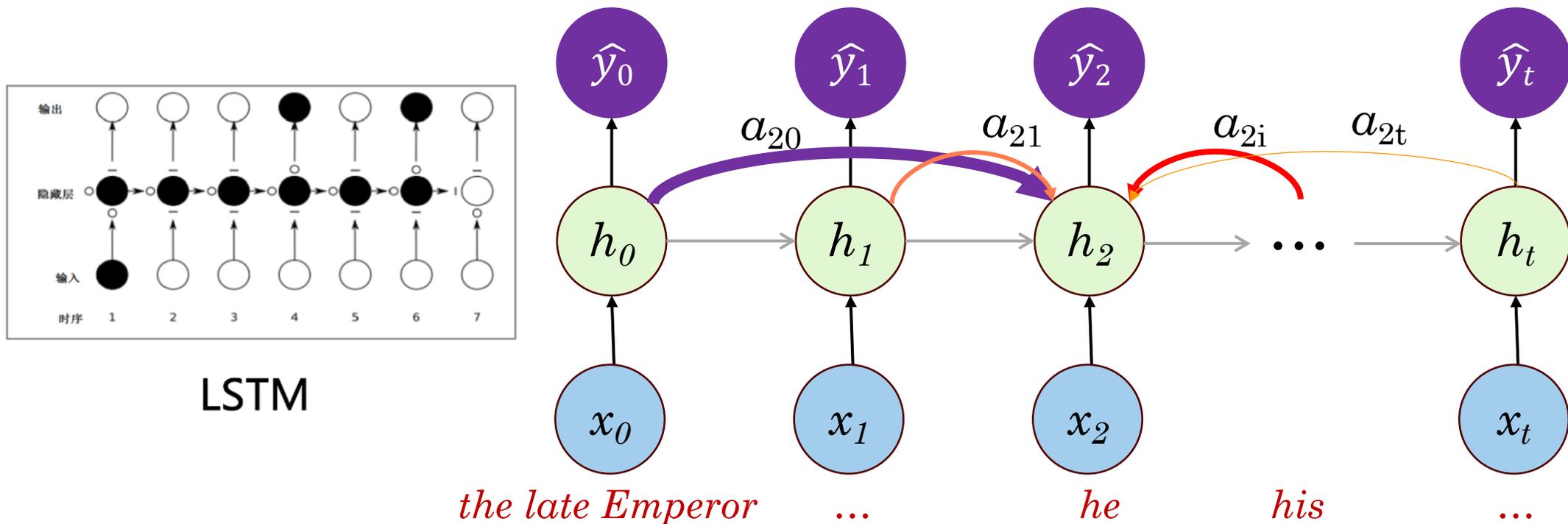
3. 归一化



那时候，柔嘉在家里等鸿渐回家来吃晚饭，希望他会跟姑母和好，到她厂里做事

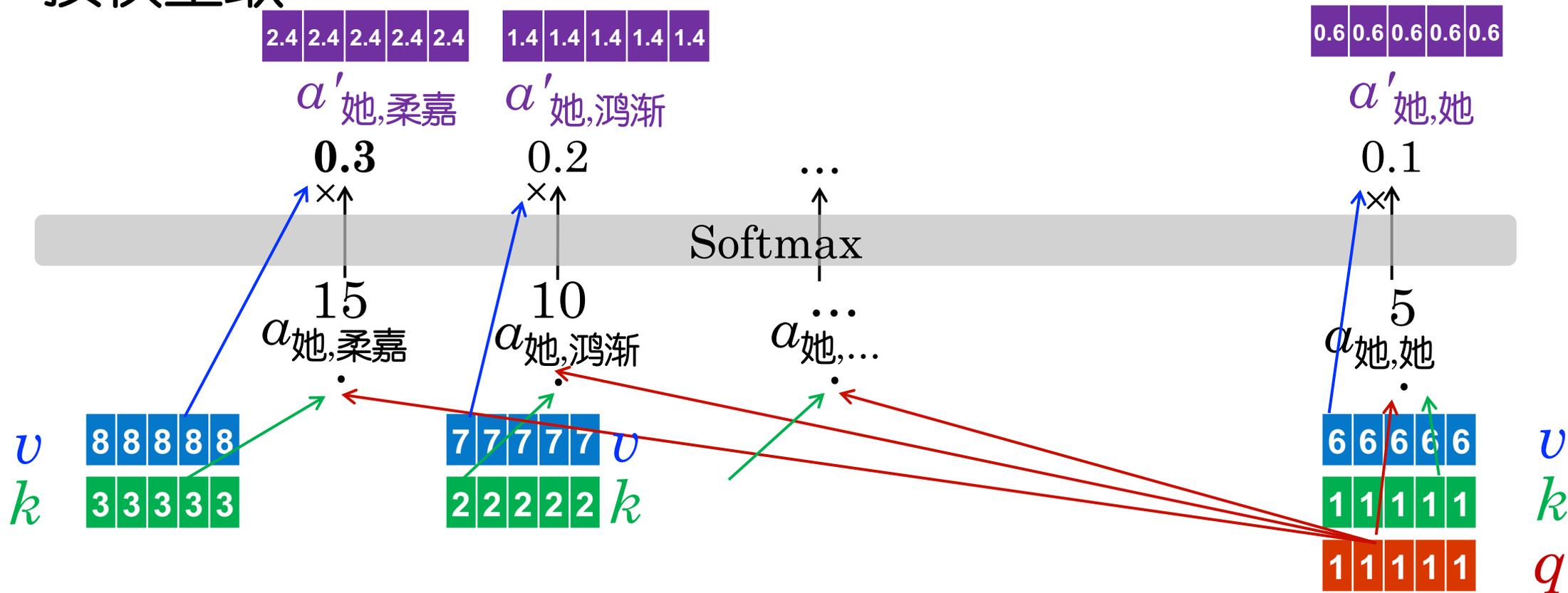
消除代代相传记忆的一个方法：查读

- 如果每个词都知道“我该注意谁”，那么就不再需要RNN中不断代代相传历史记忆，而是需要时去使用该注意的**词的信息**即可



如何快速读取所注意词的信息

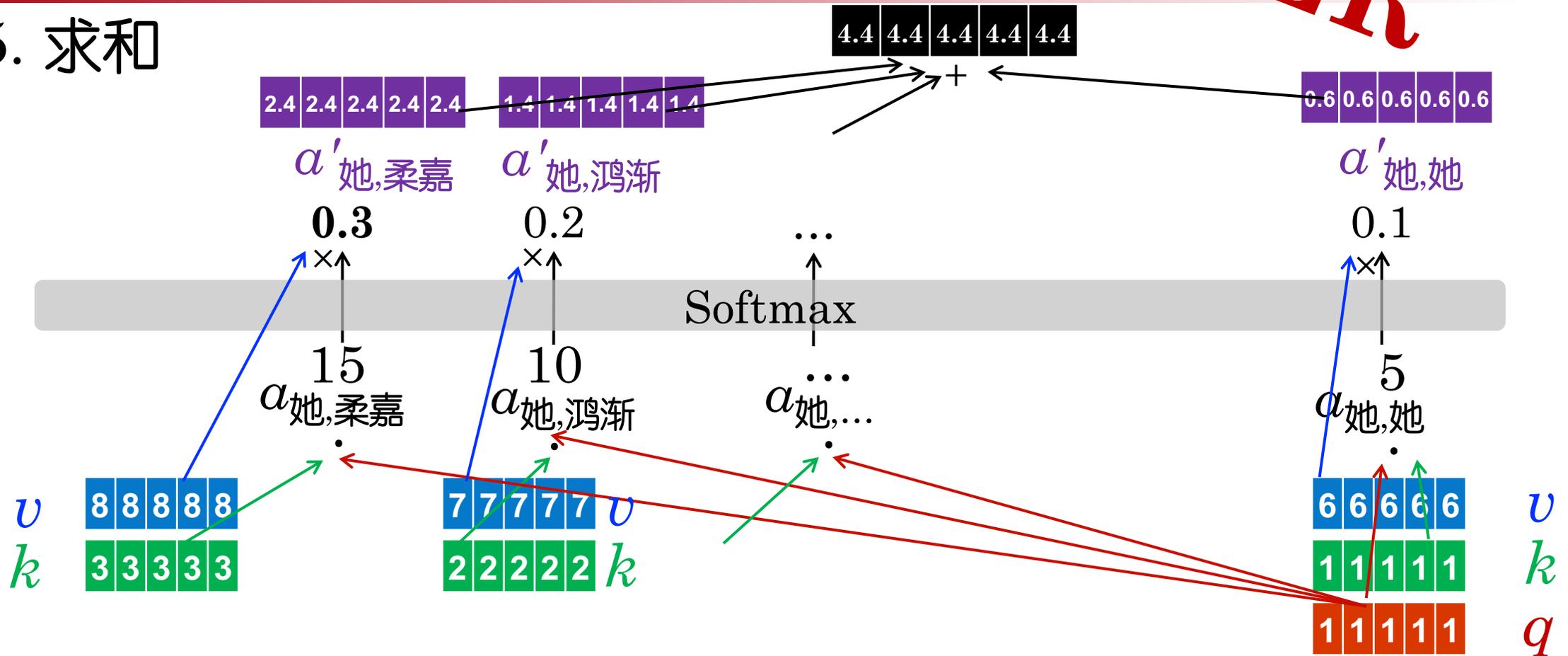
4. 按权重取Value



那时候，柔嘉在家里等鸿渐回家来吃晚饭，希望他会跟姑母和好，到她厂里做事

如何快速读取所注意词的信息 **OVER**

5. 求和



那时候，柔嘉在家里等鸿渐回家来吃晚饭，希望他会跟姑母和好，到她厂里做事

如何使用所注意词的信息

无它

4.4 4.4 4.4 4.4 4.4

1. 拼接:

5	5	5	5	5	4.4	4.4	4.4	4.4	4.4
---	---	---	---	---	-----	-----	-----	-----	-----

2. 相加:

9.4	9.4	9.4	9.4	9.4
-----	-----	-----	-----	-----

那时候，柔嘉在家里等鸿渐回家来吃晚饭，希望他会跟姑母和好，到她厂里做事

5 5 5 5 5

还有第三条路



我是效率控，多一次加法也不行！

4.4 4.4 4.4 4.4 4.4

演化
在寻找、读取过程中，
同步更新了“她”的原始编码！

5 5 5 5 5

那时候，柔嘉在家里等鸿渐回家来吃晚饭，希望他会跟姑母和好，到她厂里做事

精巧的方法

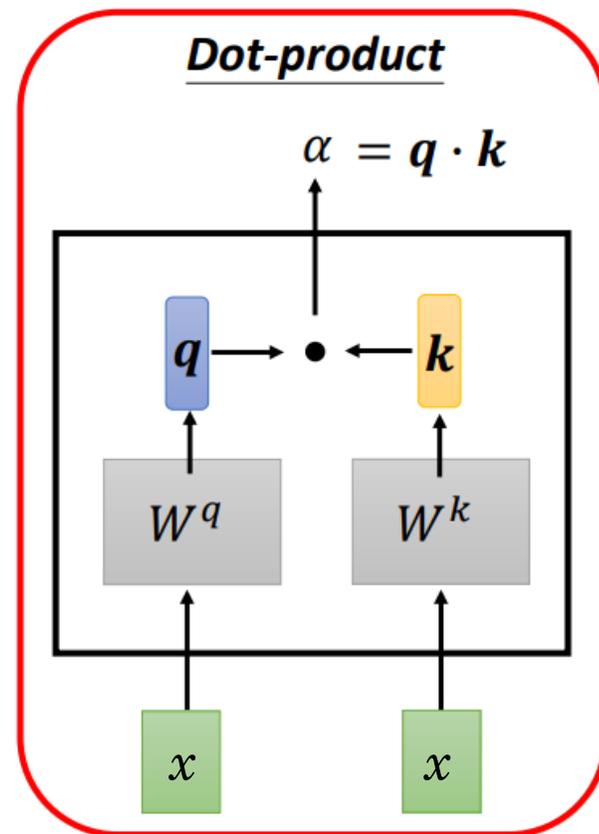
q, k, v 从 x 孵化而来

$$q = x \cdot W^q$$

$$k = x \cdot W^k$$

$$v = x \cdot W^v$$

模型参数



支持Q孵化有 W^Q 矩阵

柔嘉
鸿渐
他
她

x_1
 x_2
 x_3
 x_4

词数4 × 词编码长度5

W^Q

0.1	0.0	0.0	0.2	0.0
0.0	0.2	0.1	0.0	0.0
0.1	0.0	0.0	0.0	0.2
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.1	0.0	0.0

词编码长度5 × q维度5

(q维度此处等于词编码长度,

但也可以不相等)

q_1
 q_2
 q_3
 q_4

1	1	1	1	1

词数4 × q维度

还有 W^K, W^V “孵化”矩阵，都是模型参数

柔嘉
鸿渐
他
她

x_1					
x_2					
x_3					
x_4	5	5	5	5	5

0.1	0.0	0.0	0.2	0.0
0.0	0.2	0.1	0.0	0.0
0.1	0.0	0.0	0.0	0.2
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.1	0.0	0.0

W^Q

W^K

W^V

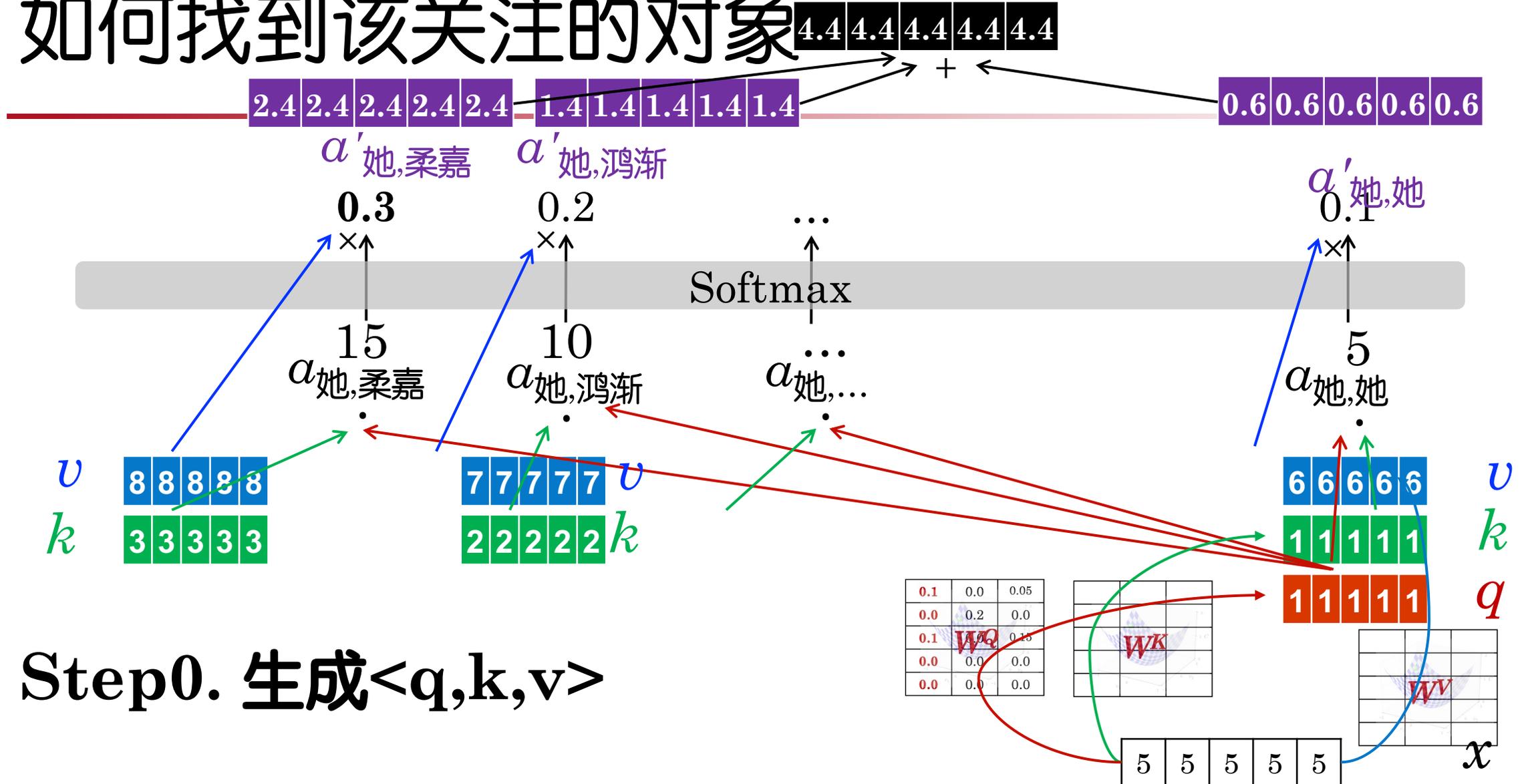


q_1				
q_2				
q_3				
q_4	1	1	1	1

k_1				
k_2				
k_3				
k_4				

u_1				
u_2				
u_3				
u_4				

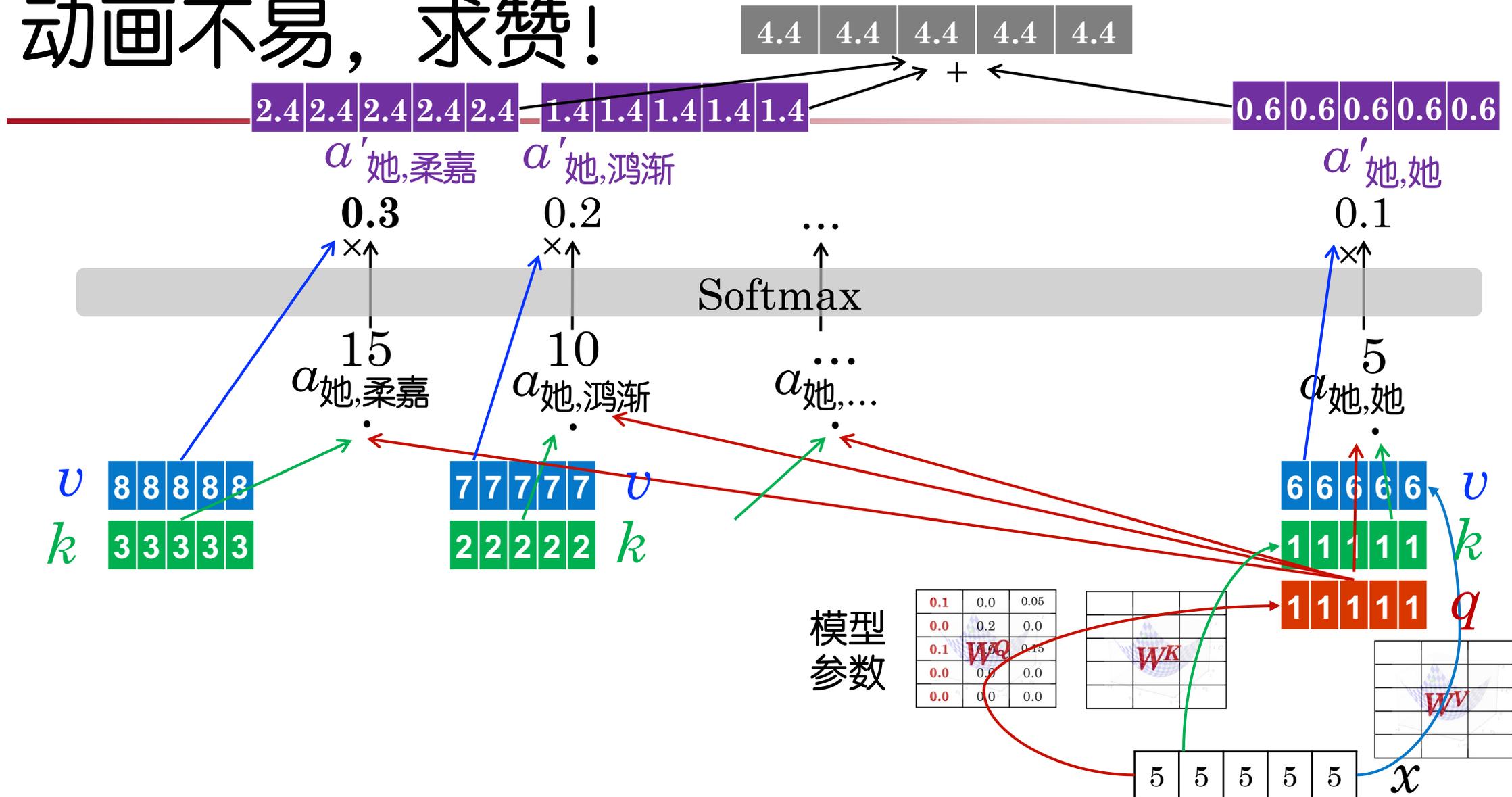
如何找到该关注的对象



Step0. 生成 $\langle q, k, v \rangle$

那时候，柔嘉在家里等鸿渐回家来吃晚饭，希望他会跟姑母和好，到她厂里做事

动画不易，求赞！



那时候，柔嘉在家里等鸿渐回家来吃晚饭，希望他会跟姑母和好，到她厂里做事

4.4 4.4 4.4 4.4 4.4



演化

在寻找、读取过程中，“她”同步更新了自己的原始编码，将自己关心的人融入自己的编码！

她不再是一个人在战斗！

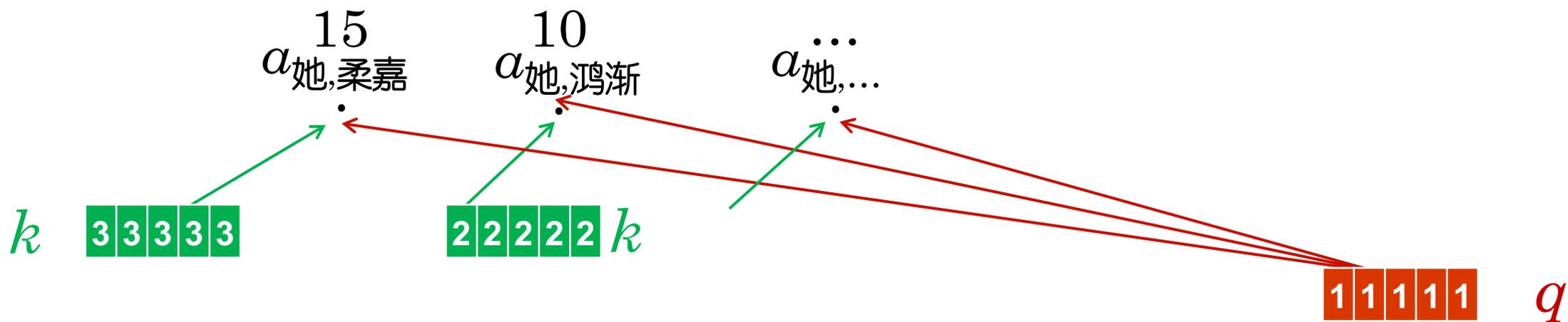
5 5 5 5 5 x

那时候，柔嘉在家里等鸿渐回家来吃晚饭，希望他会跟姑母和好，到她厂里做事

Trick: 缩放 – 消除 $\langle q, k, v \rangle$ 向量长度影响

1. 计算 w_i 对所有 w_j 的注意力分数: $\alpha_{i,j} = q^i \cdot k^j$

q, k 内积的方差与维度线性正比, 太大时Softmax分布极其“尖锐”

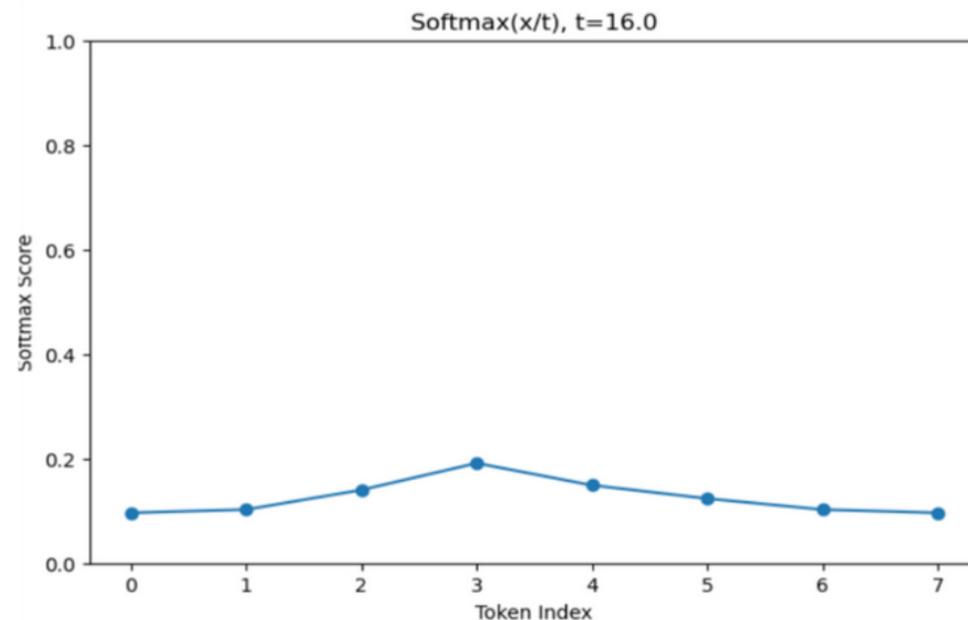
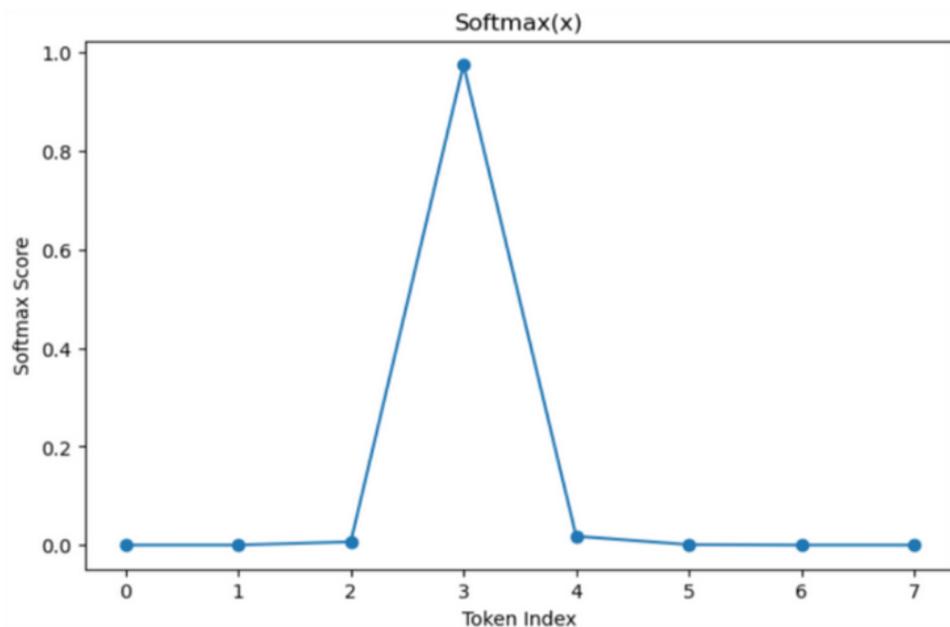


那时候, 柔嘉在家里等鸿渐回家来吃晚饭, 希望他会跟姑母和好, 到她厂里做事

Trick: 缩放 – 消除 $\langle q, k, v \rangle$ 向量长度影响

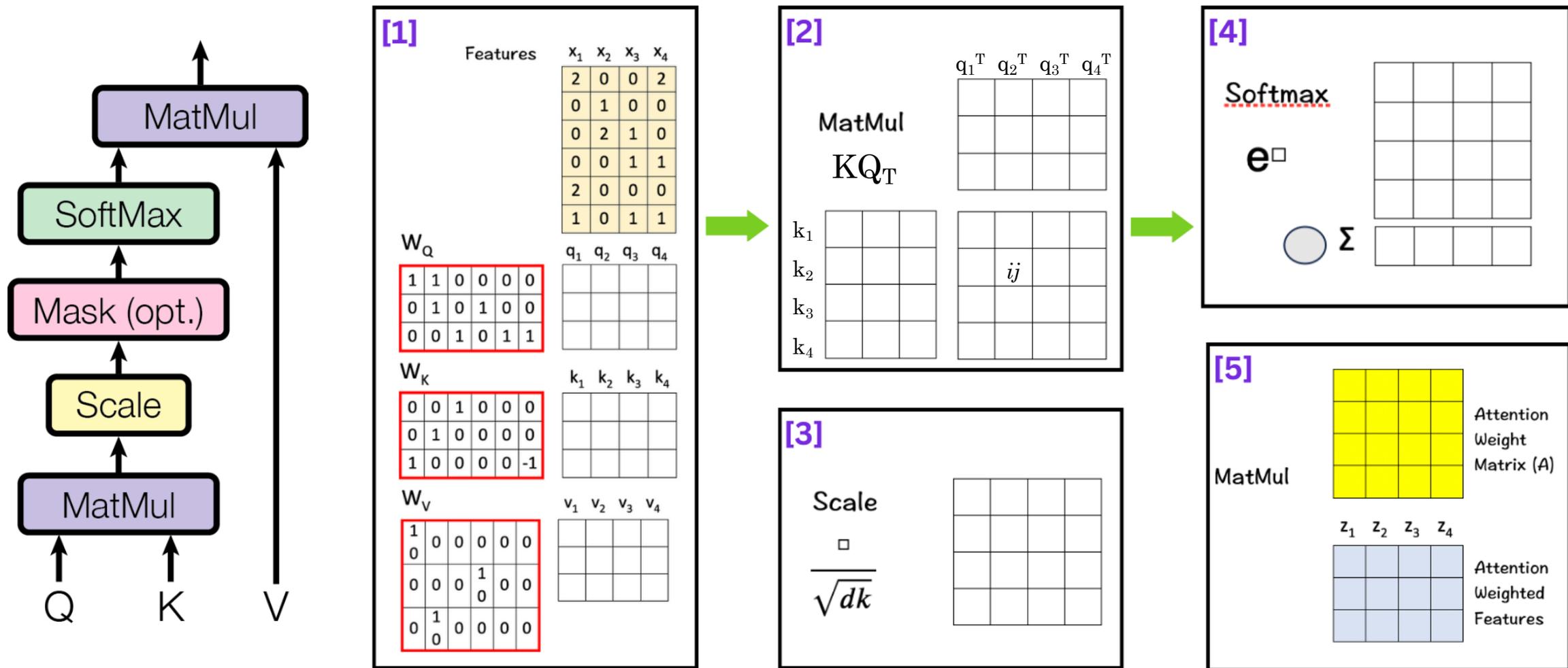
1. 计算 w_i 对所有 w_j 的注意力分数: $\alpha_{i,j} = q^i \cdot k^j$

q, k 内积的方差与维度线性正比, 太大时Softmax分布极其“尖锐”



[1, 2, 7, 12, 8, 5, 2, 1]

Self-Attention过程一览

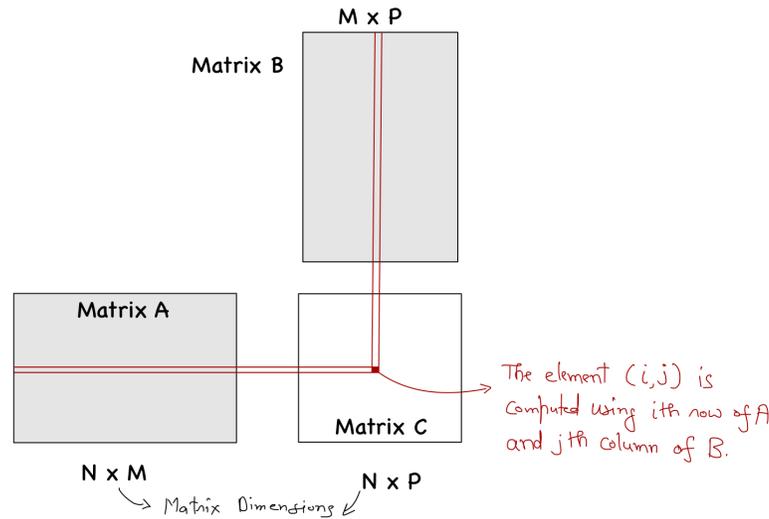
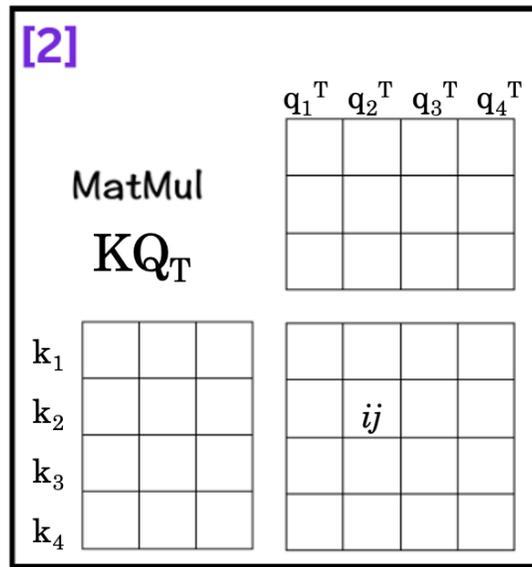


Self-Attention的数学表示

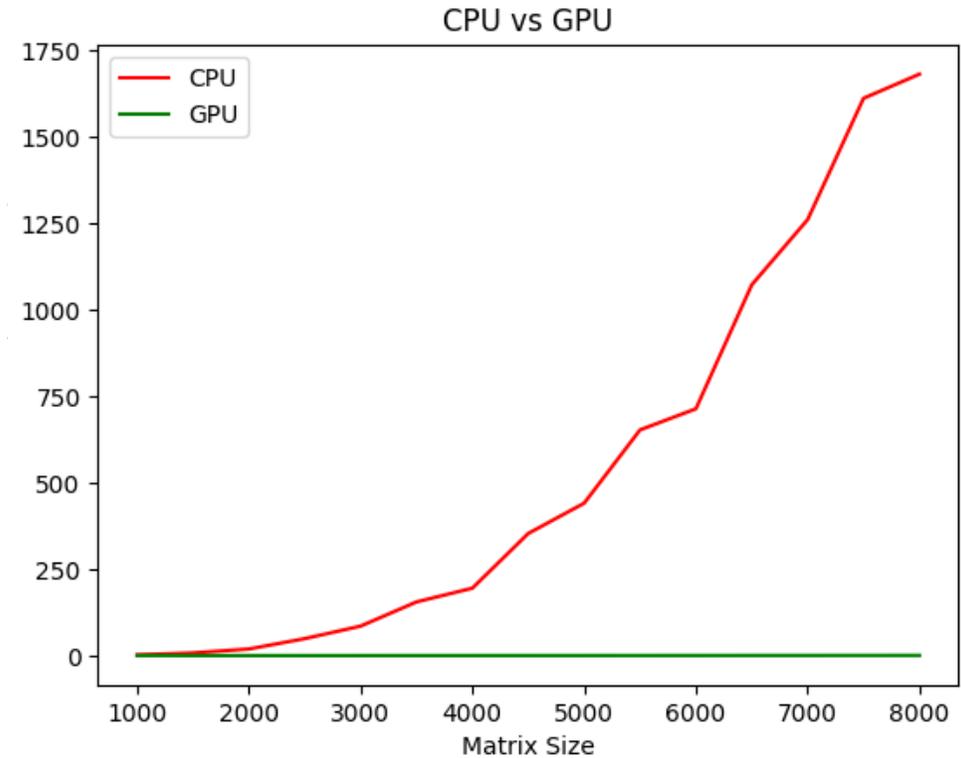
$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

并发能力：GPU善于处理矩阵运算

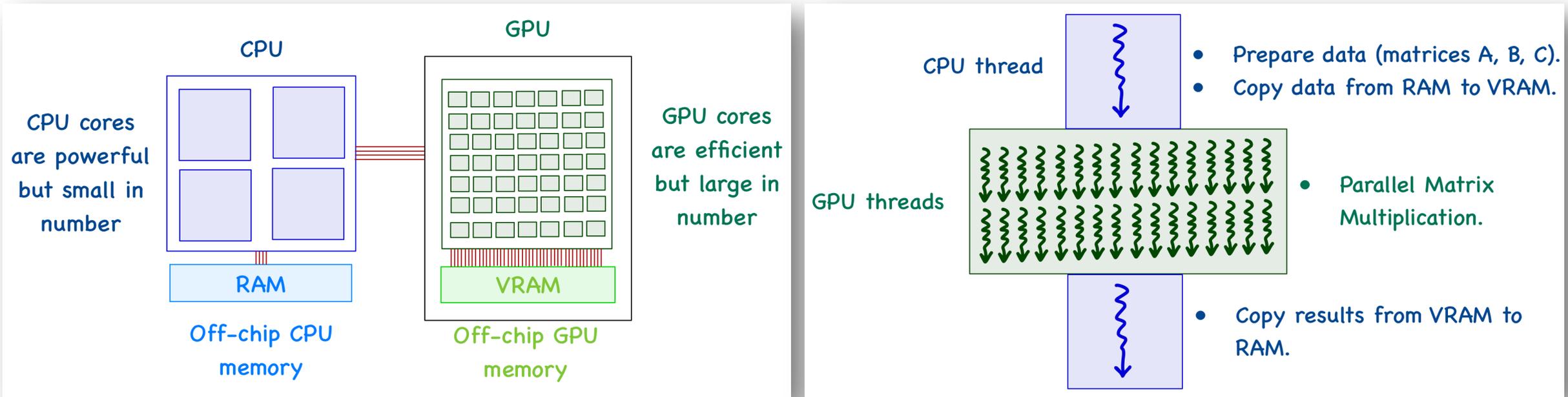


一次矩阵乘法，算出所有词之间的两两注意力



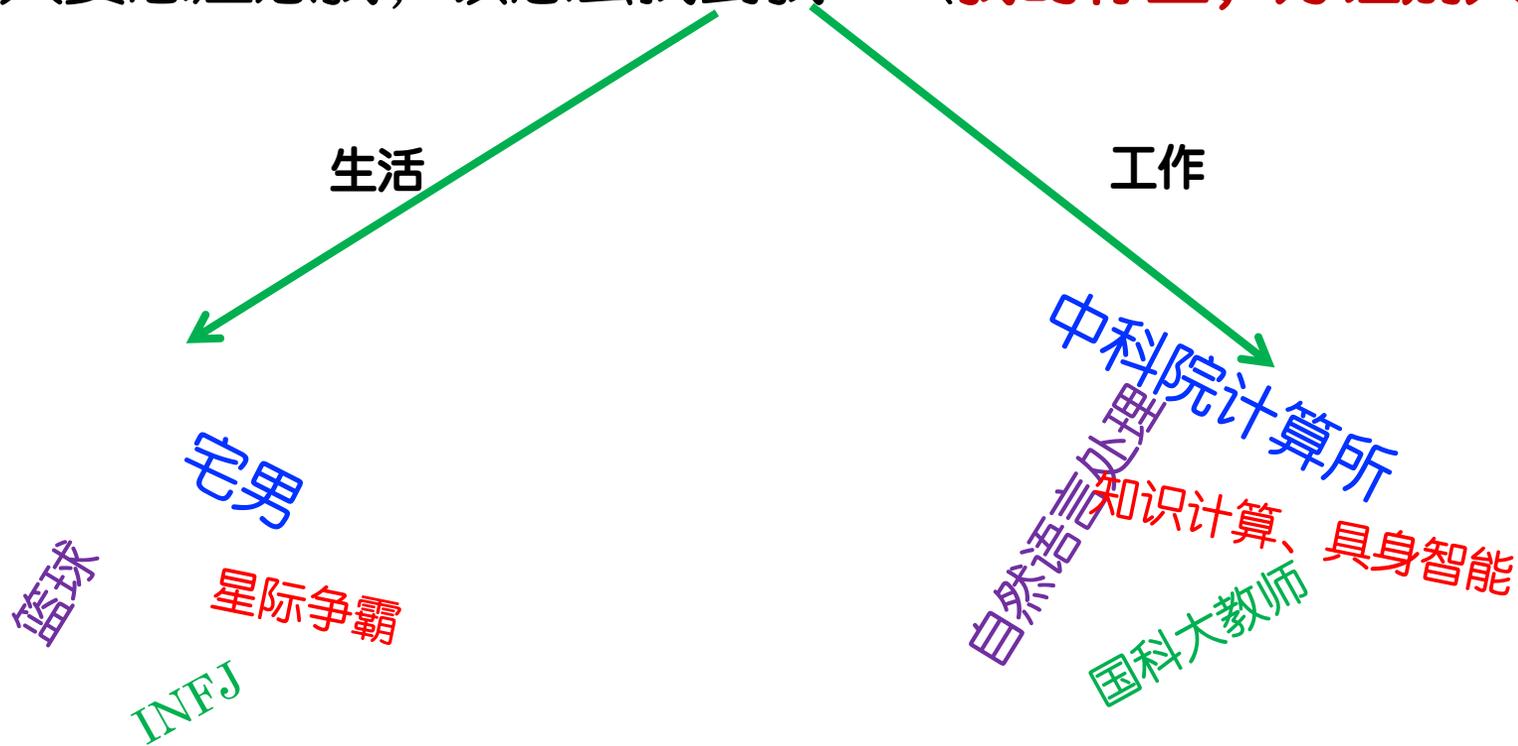
而GPU非常善于矩阵运算

并发能力：GPU善于处理矩阵运算



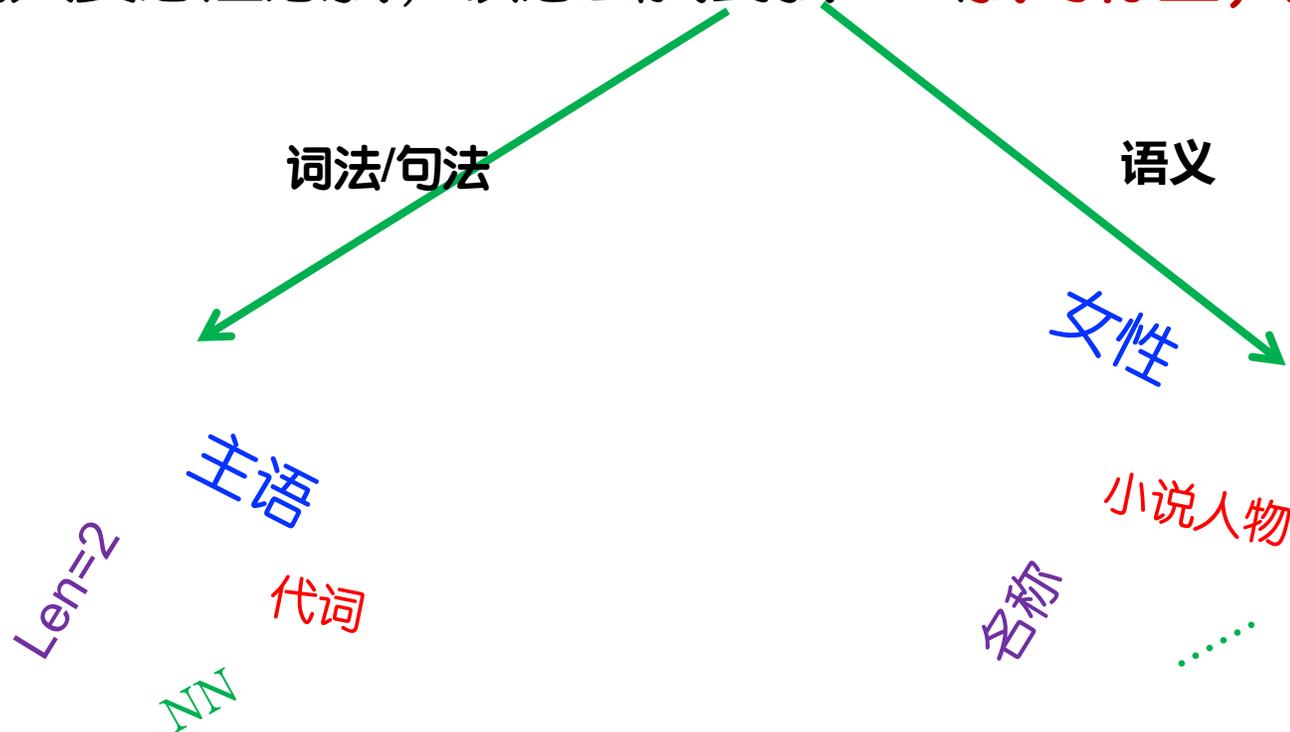
问题1：当我有两套标签，别人怎么找我？

□ 别人要想注意我，该怎么找到我？（我的标签，好让别人也能要回答）

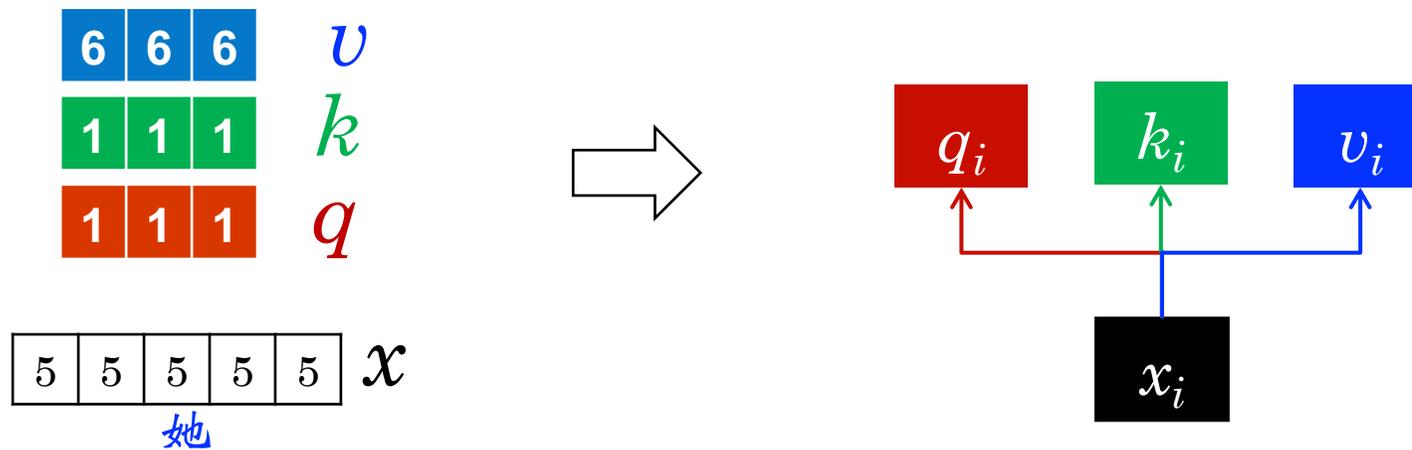


问题1：当我有两套标签，别人怎么找我？

- 别人要想注意我，该怎么找到我？（我的标签，好让别人也能要回答）

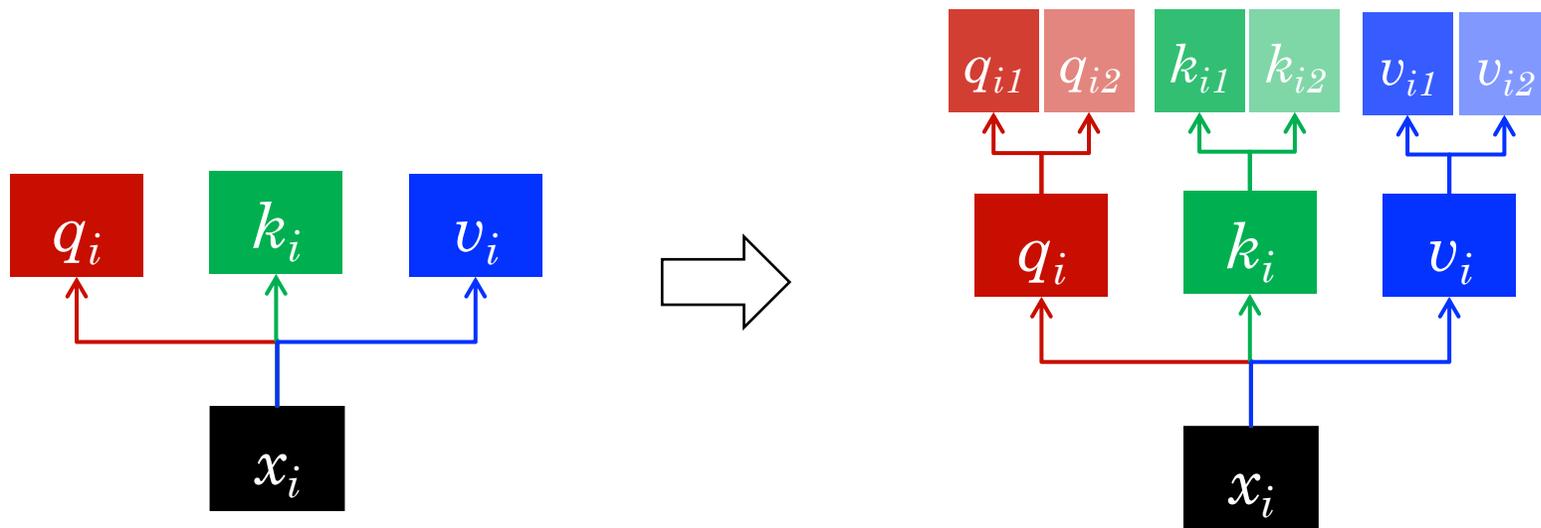


Self-attention 简化表示



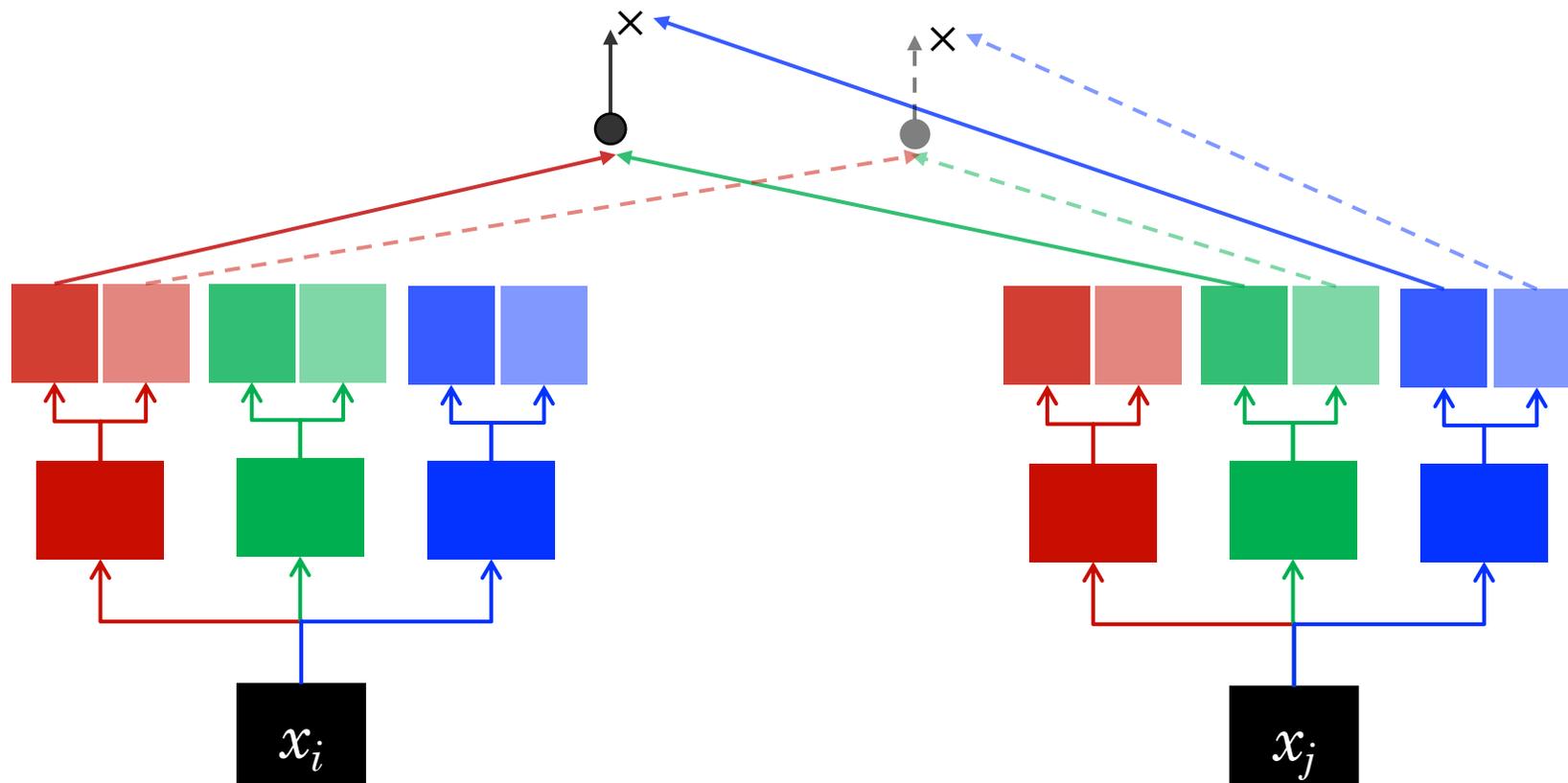
Multi-head Self-attention

一个注意力的维度不够，那就多个维度，各自单独使用



Multi-head Self-attention

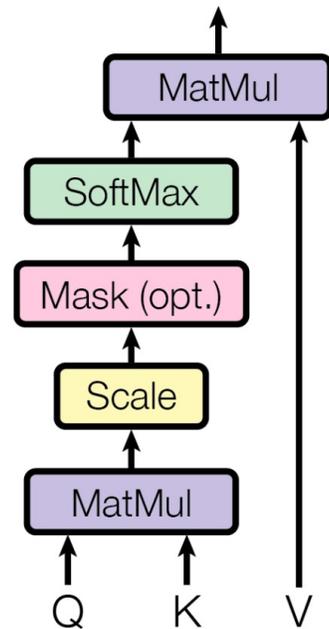
一个注意力的维度不够，那就多个维度，各自单独使用



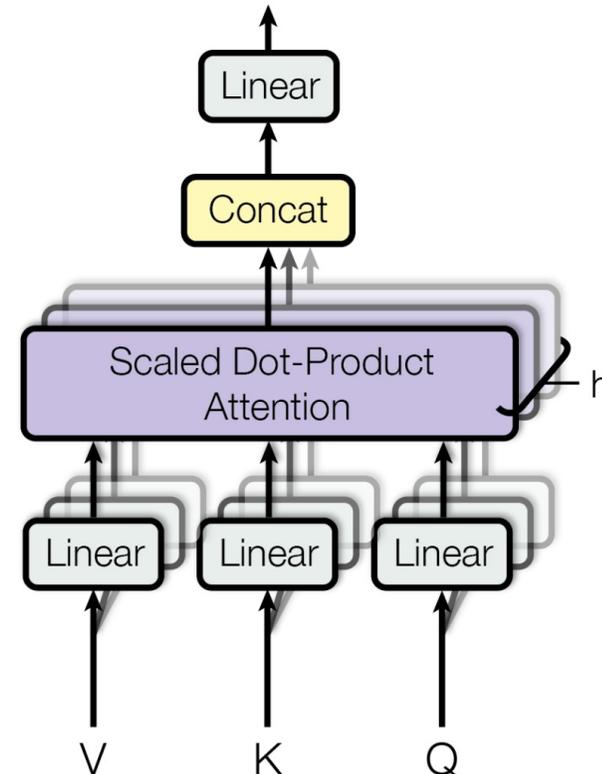
Multi-head Self-attention

一个注意力的维度不够，那就多个维度，**各自单独使用**

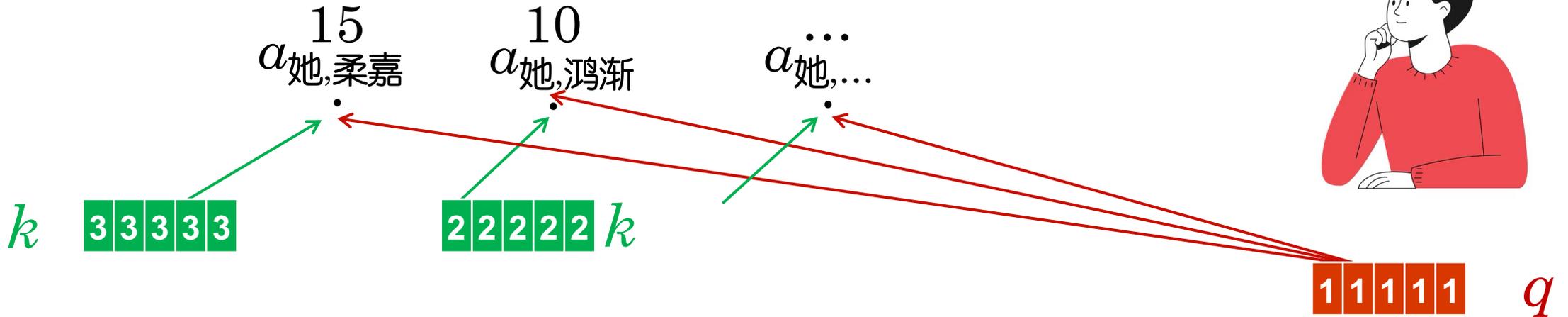
Scaled Dot-Product Attention



Multi-Head Attention



自注意力有什么缺点?



那时候，柔嘉在家里等鸿渐回家来吃晚饭，希望他会跟姑母和好，到她厂里做事

那时候，柔嘉在家里等鸿渐回家来吃晚饭，希望他到她厂里做事

注意力权重跟词之间的距离无关?

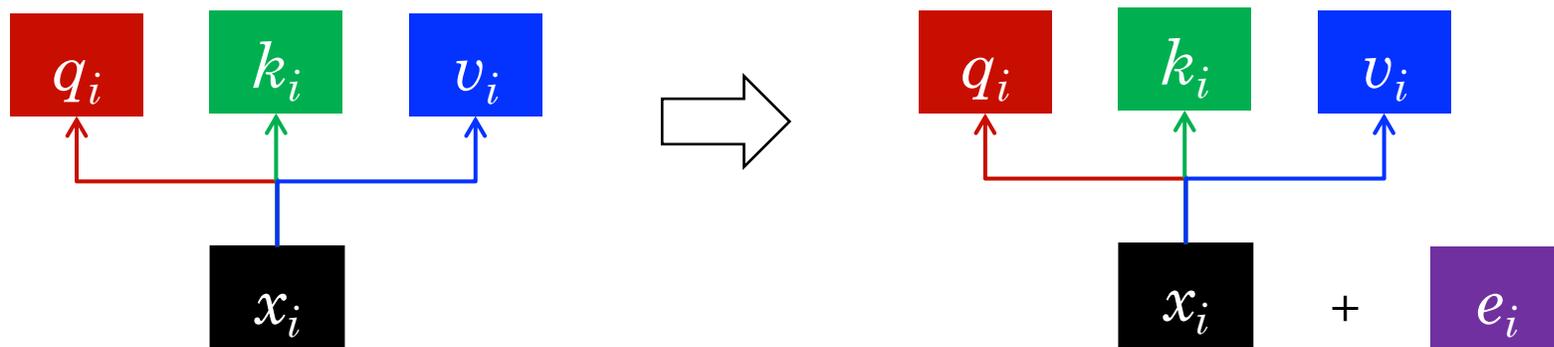


目 录

- 1 带注意力的Encoder-Decoder
- 2 自注意力
- 3 位置编码
- 4

方法1：可训练的绝对位置编码

- 自注意力对每个词增加了一个它独特的“位置编码”
- 典型模型：BERT



$$\text{input}_i = \text{token_embed}_i + \text{pos_embed}_i$$

方法1：可训练的绝对位置编码

- 缺点：无“外推”能力
 - 训练时句子最长5个字，来了7个字的不会了？
 - 只寻到了距离 <5 的字对注意力权重的影响

?



风劲角弓鸣，
将军猎渭城。

?



风急天高猿啸哀，
渚清沙白鸟飞回。

方法2: Sinusoidal (三角函数式) 位置编码

- 将位置编码为正弦余弦函数
- 原始Transformer使用

Position Embedding 是一个向量

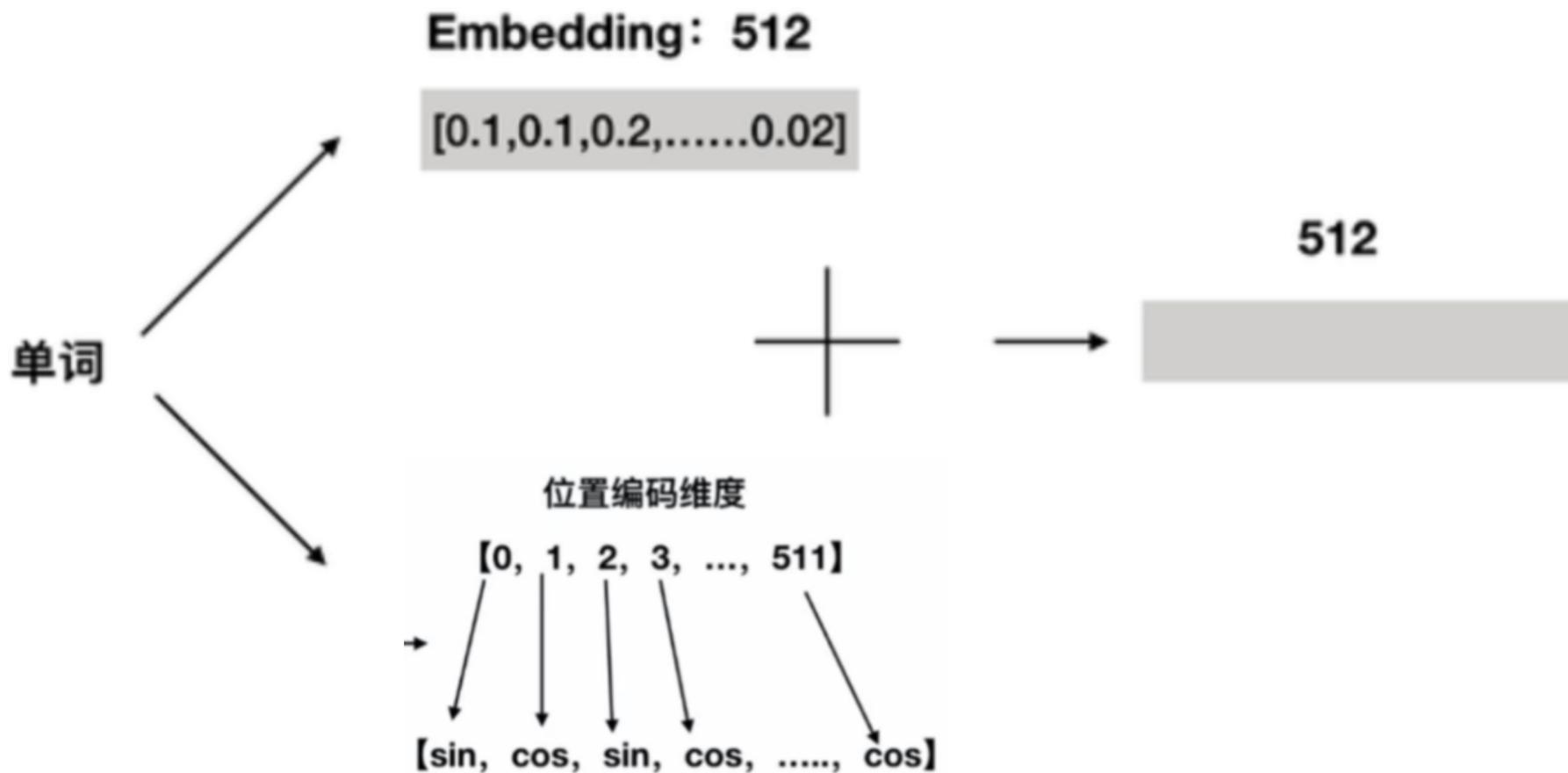
Position 词的位置

向量的元素下标

token编码长度

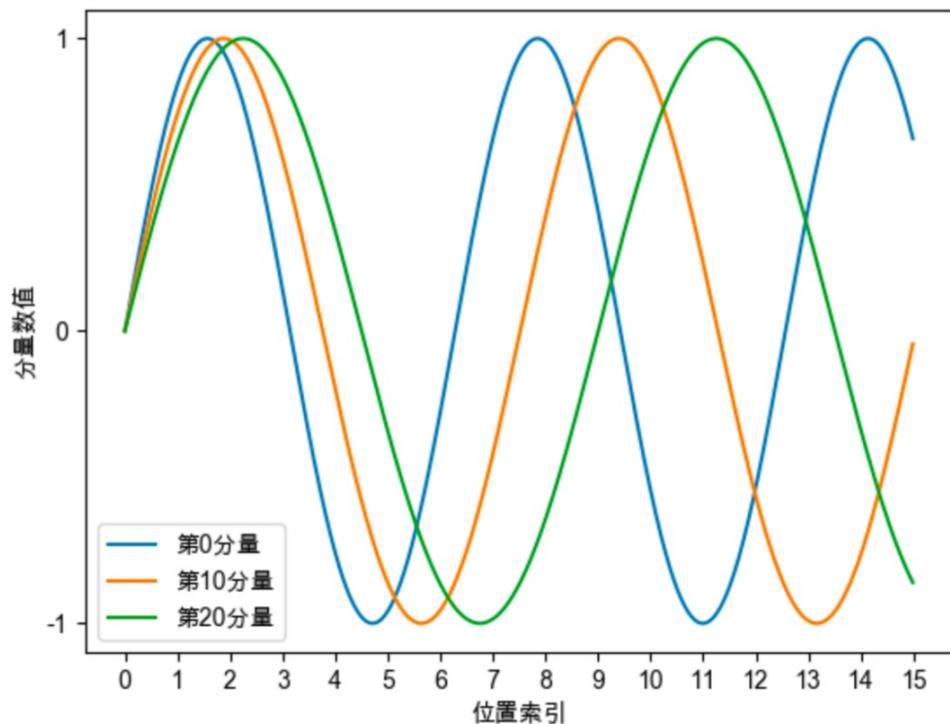
$$PE_{(pos, 2i)} = \sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos, 2i+1)} = \cos(pos/10000^{2i/d_{\text{model}}})$$

方法2: Sinusoidal (三角函数式) 位置编码

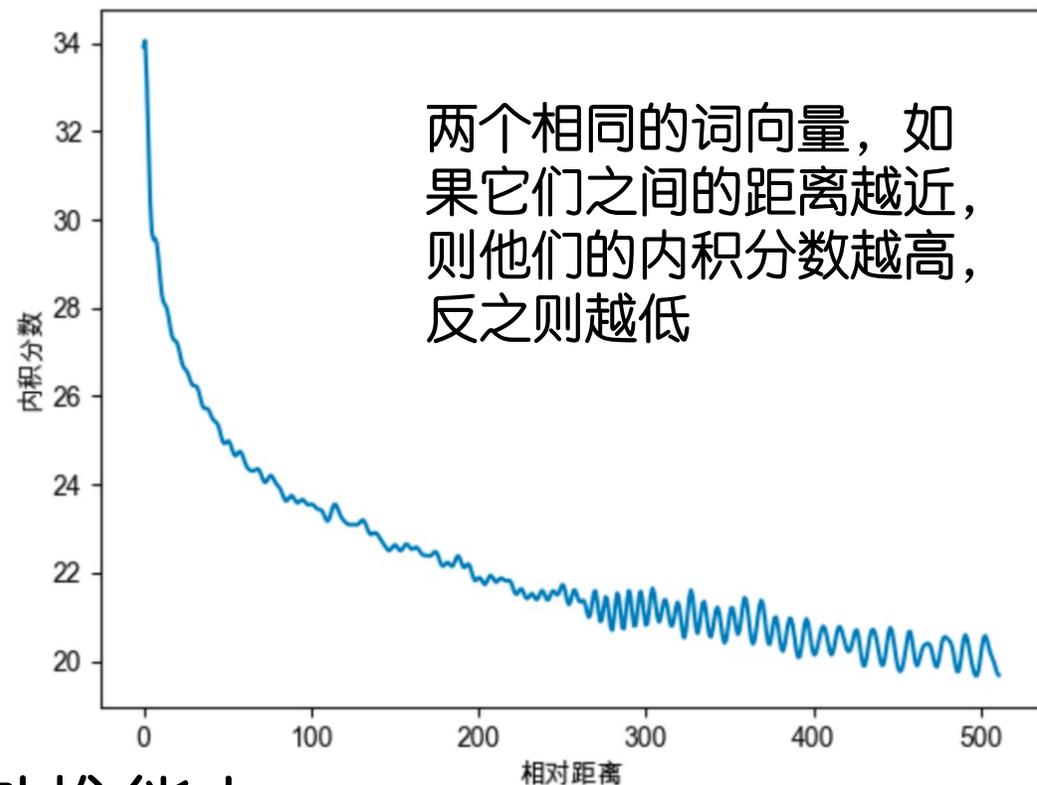


方法2: Sinusoidal (三角函数式) 位置编码

□ 性质1: 周期性



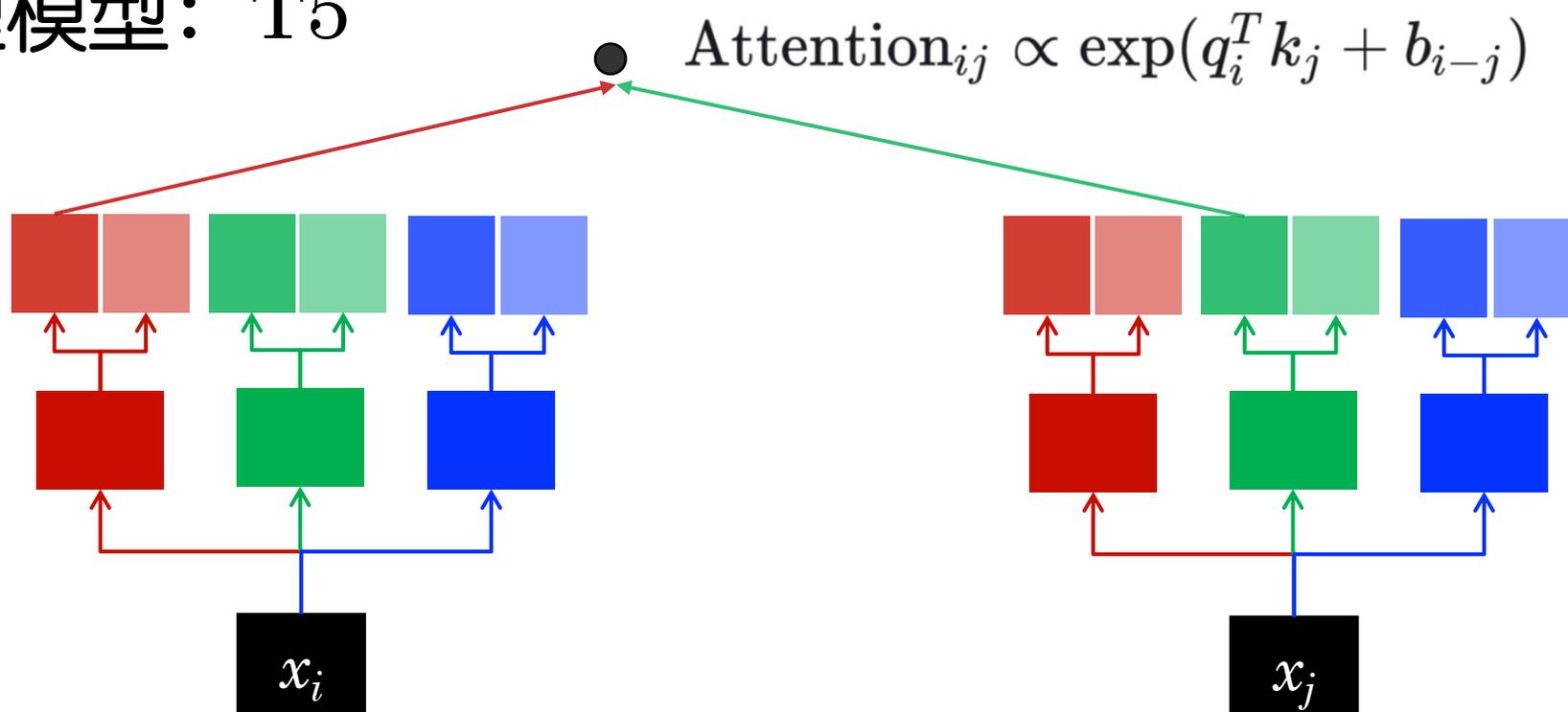
□ 性质2: 衰减性



产生一定的外推能力

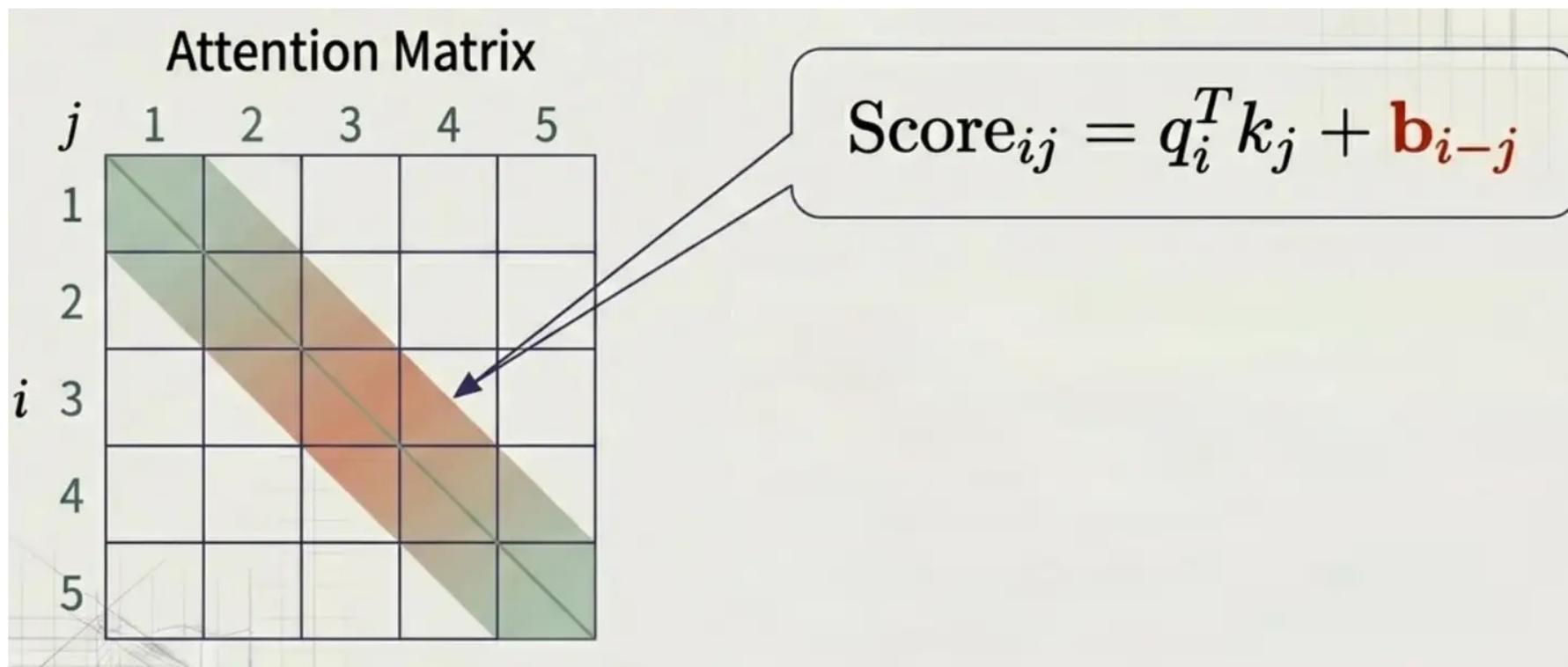
方法3：相对位置编码

- 在计算attention时，加入一个只依赖于相对位置的偏置
- 典型模型：T5



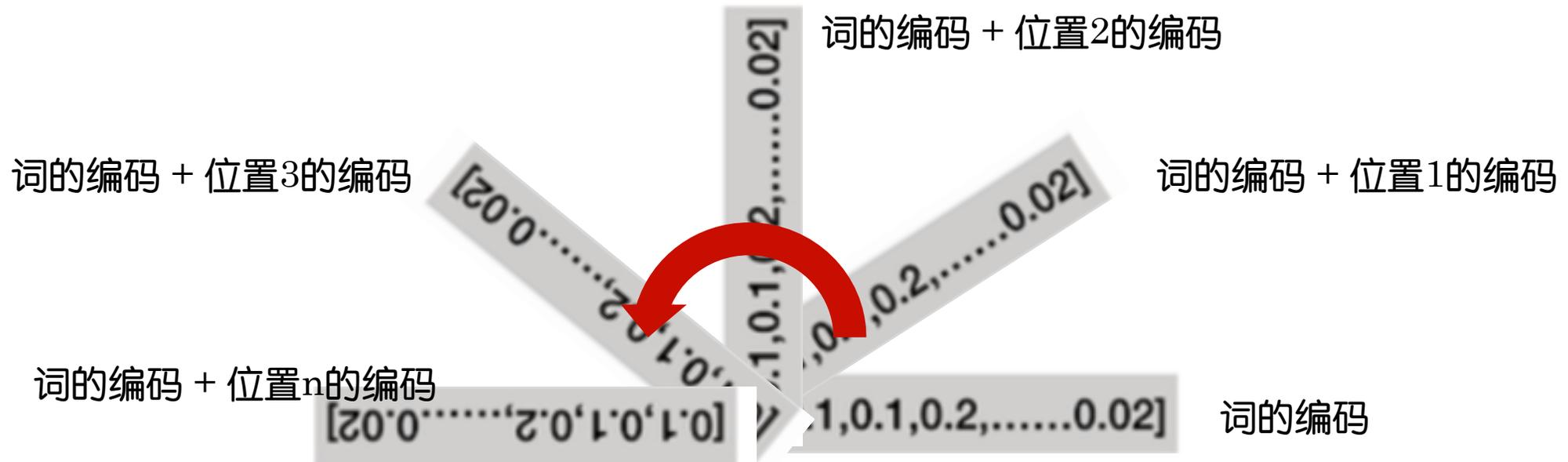
方法3：相对位置编码

- 在计算attention时，加入一个只依赖于相对位置的偏置
- 典型模型：T5



方法4: RoPE (Rotary Position Embedding 旋转位置编码)

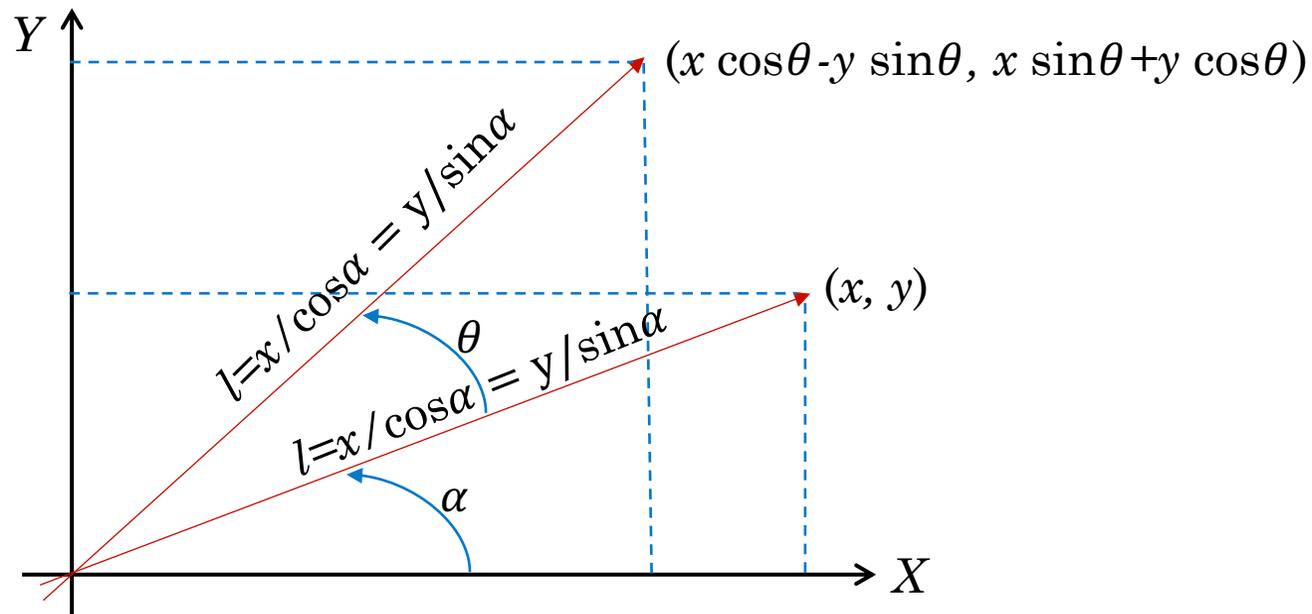
- ❑ 思想: 通过绝对位置编码的方式实现相对位置编码
- ❑ 方法: 将一个向量旋转某个角度代表位置信息
- ❑ 目前LLM的主流选择: LLaMA、GPT、ChatGLM



方法4: RoPE (Rotary Position Embedding 旋转位置编码)

□ 将一个向量旋转某个角度代表位置信息

1. 词编码为 (x, y) , 在笛卡尔空间表示为向量 (一个带有方向和长度的量)
2. (x, y) 逆时针旋转 θ 度后坐标如图

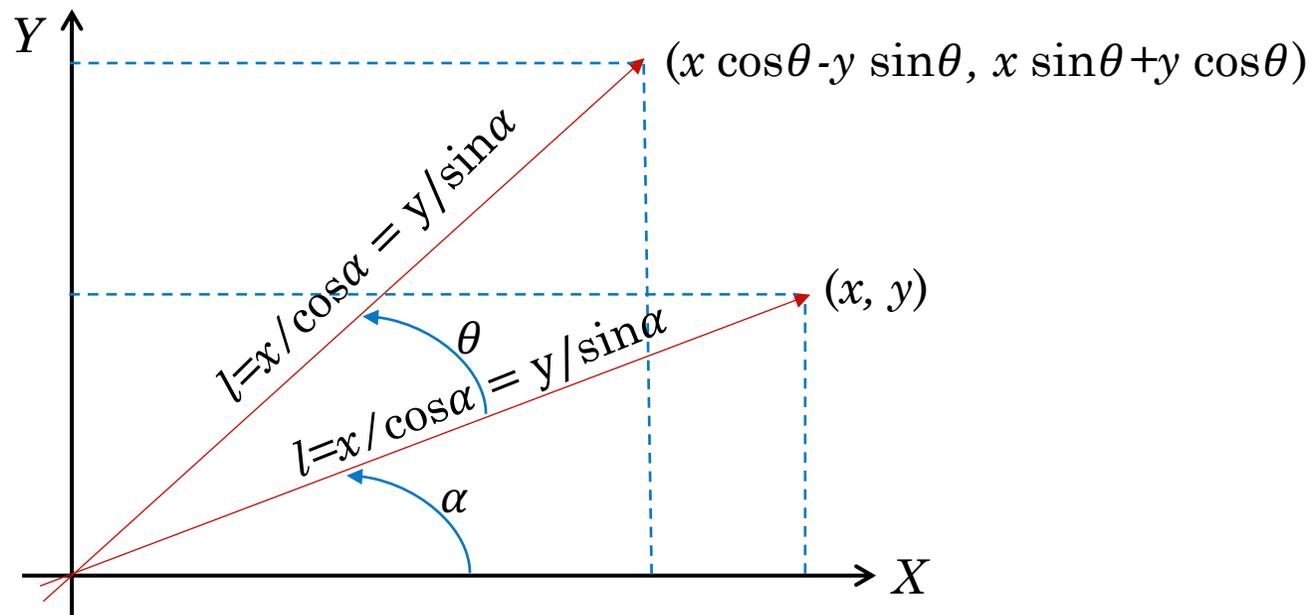


- 向量长度 $l = x / \cos \alpha = y / \sin \alpha$
- $x' = l \cos(\alpha + \theta)$
 $= l (\cos \alpha \cdot \cos \theta - \sin \alpha \cdot \sin \theta)$
 $= x \cos \theta - y \sin \theta$
- $y' = x \sin \theta + y \cos \theta$

方法4: RoPE (Rotary Position Embedding 旋转位置编码)

□ 将一个向量旋转某个角度代表位置信息

1. 词编码为 (x, y) , 在笛卡尔空间表示为向量 (一个带有方向和长度的量)
2. (x, y) 逆时针旋转 θ 度后坐标如图



$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

旋转矩阵

方法4: RoPE (Rotary Position Embedding 旋转位置编码)

□ 将一个向量旋转某个角度代表位置信息

Tips:
 $a \cdot b = a^T * b$

旋转 $m\theta$ 和 $n\theta$ 后两个向量的点积 (相似度), 等于旋转了 $(n-m)\theta$ 的旋转矩阵 (相对位置)

$$\begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \cdot \begin{pmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{pmatrix}$$

$$= \begin{bmatrix} \cos m\theta & \sin m\theta \\ -\sin m\theta & \cos m\theta \end{bmatrix} \times \begin{bmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{bmatrix}$$

$$= \begin{bmatrix} \cos m\theta \cos n\theta + \sin m\theta \sin n\theta & \cos m\theta \cos n\theta - \cos m\theta \sin n\theta \\ \cos m\theta \cos n\theta - \sin m\theta \cos n\theta & \cos m\theta \cos n\theta + \sin m\theta \sin n\theta \end{bmatrix}$$

$$= \begin{bmatrix} \cos(n-m)\theta & -\sin(n-m)\theta \\ \sin(n-m)\theta & \cos(n-m)\theta \end{bmatrix}$$

两个词离得越近

= 相似度越高

= 相对旋转角度越小

方法4: RoPE (Rotary Position Embedding 旋转位置编码)

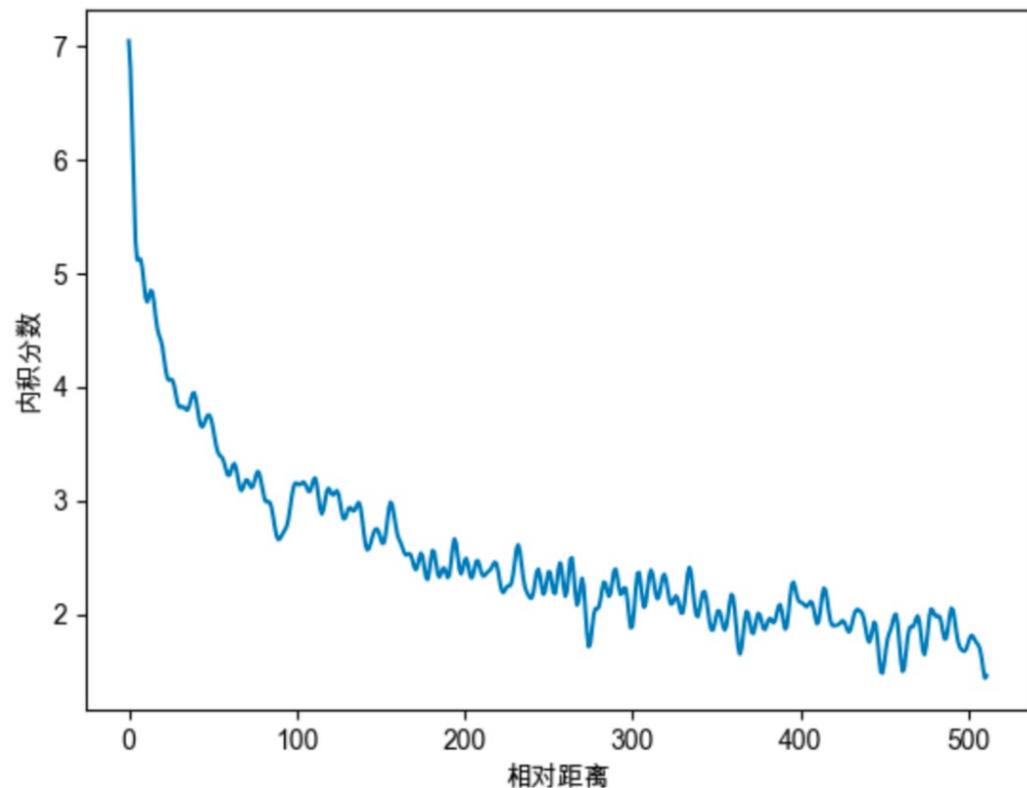
□ 将一个向量旋转某个角度代表位置信息

词编码是多维时，分解为两两一组旋转即可

$$\begin{pmatrix} \cos m\theta & -\sin m\theta & 0 & 0 & \dots & 0 & 0 \\ \sin m\theta & \cos m\theta & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \cos m\theta & -\sin m\theta & \dots & 0 & 0 \\ 0 & 0 & \sin m\theta & \cos m\theta & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \cos m\theta & -\sin m\theta \\ 0 & 0 & 0 & 0 & \dots & \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \\ \vdots \\ q_{d-2} \\ q_{d-1} \end{pmatrix}$$

方法4: RoPE (Rotary Position Embedding 旋转位置编码)

□ 衰减性



随机初始化两个向量 q 和 k ，将 q 固定在位置0上， k 的位置从0开始逐步变大，依次计算 q 和 k 之间的内积。我们发现随着 q 和 k 的相对距离的增加，它们之间的内积分数呈现出远程衰减的性质



1

带注意力的Encoder-Decoder

2

自注意力

3

位置编码

4

Transformers

Attention Is All You Need

Attention is all you need

[A Vaswani, N Shazeer, N Parmar...](#) - Advances in neural ..., 2017 - proceedings.neurips.cc

... to attend to **all** positions in the decoder up to and including that position. **We need** to prevent ... **We** implement this inside of scaled dot-product **attention** by masking out (setting to $-\infty$) ...

☆ 保存 引用 被引用次数: **236405** 相关文章 所有 71 个版本

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

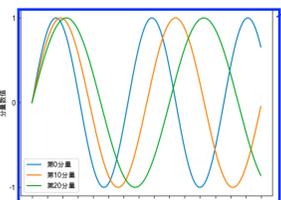
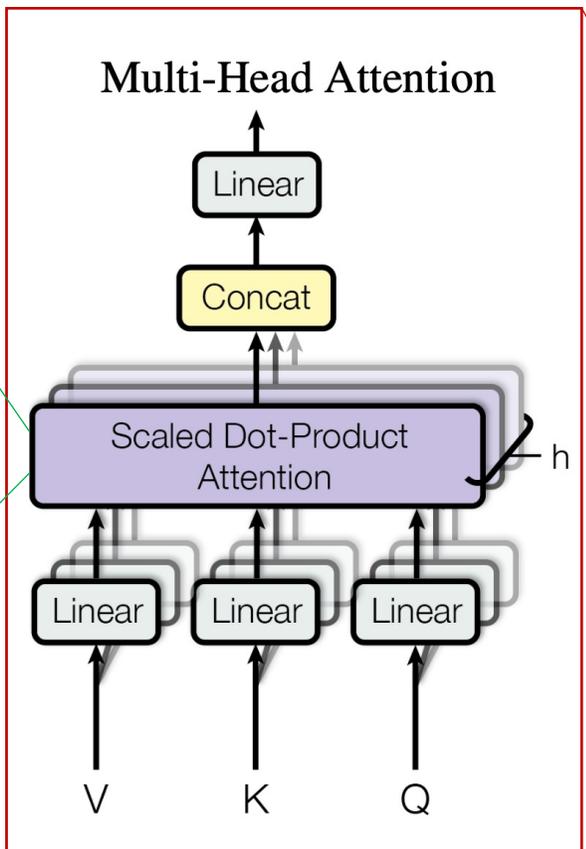
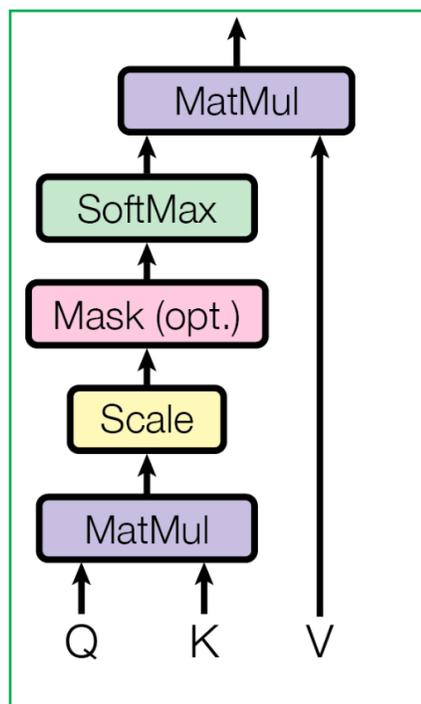
Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Illia Polosukhin* †
illia.polosukhin@gmail.com



标准模型架构



Sinusoidal
位置编码

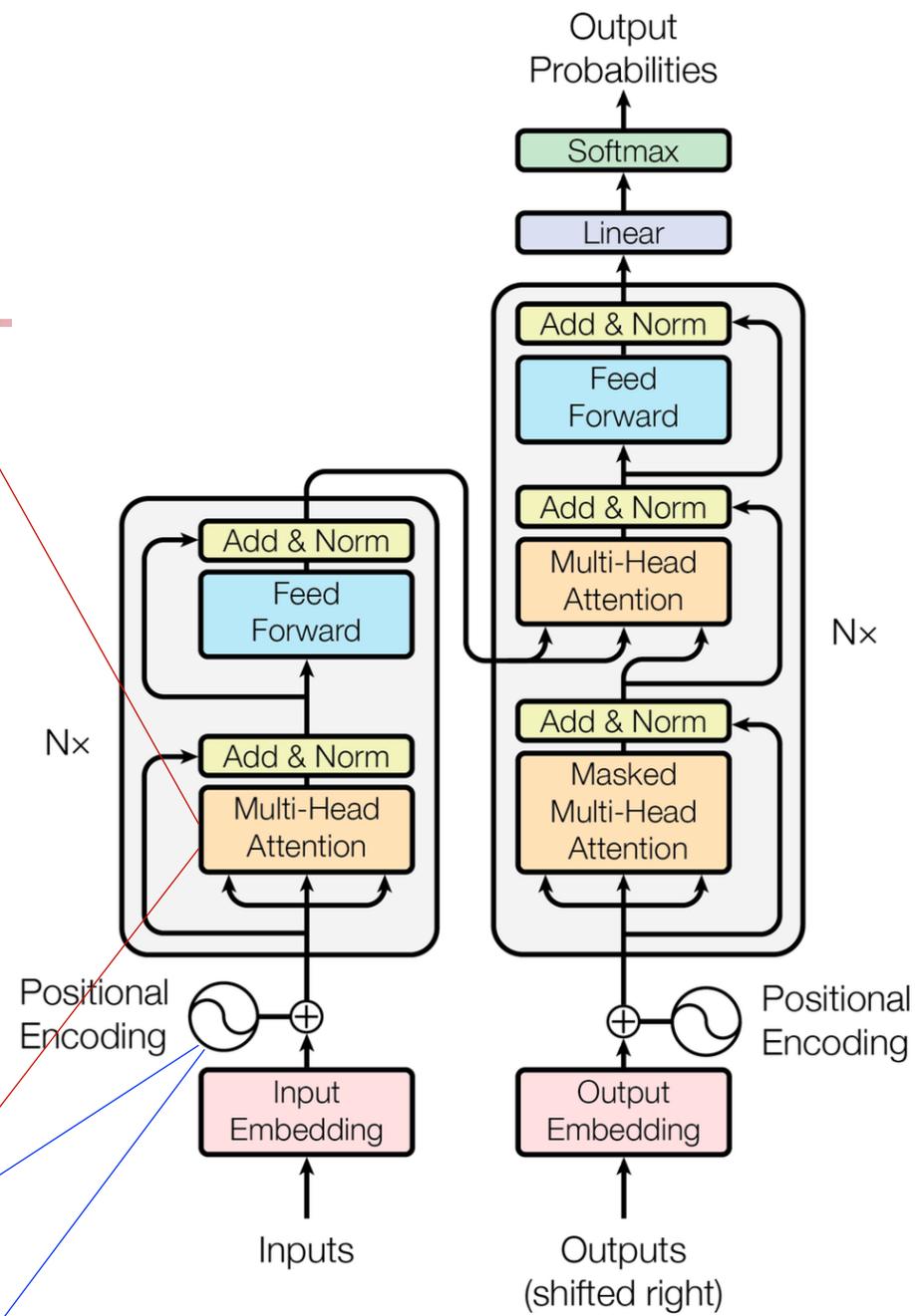


Figure 1: The Transformer - model architecture.

Add

将token原始编码与其
增加注意力后编码相加

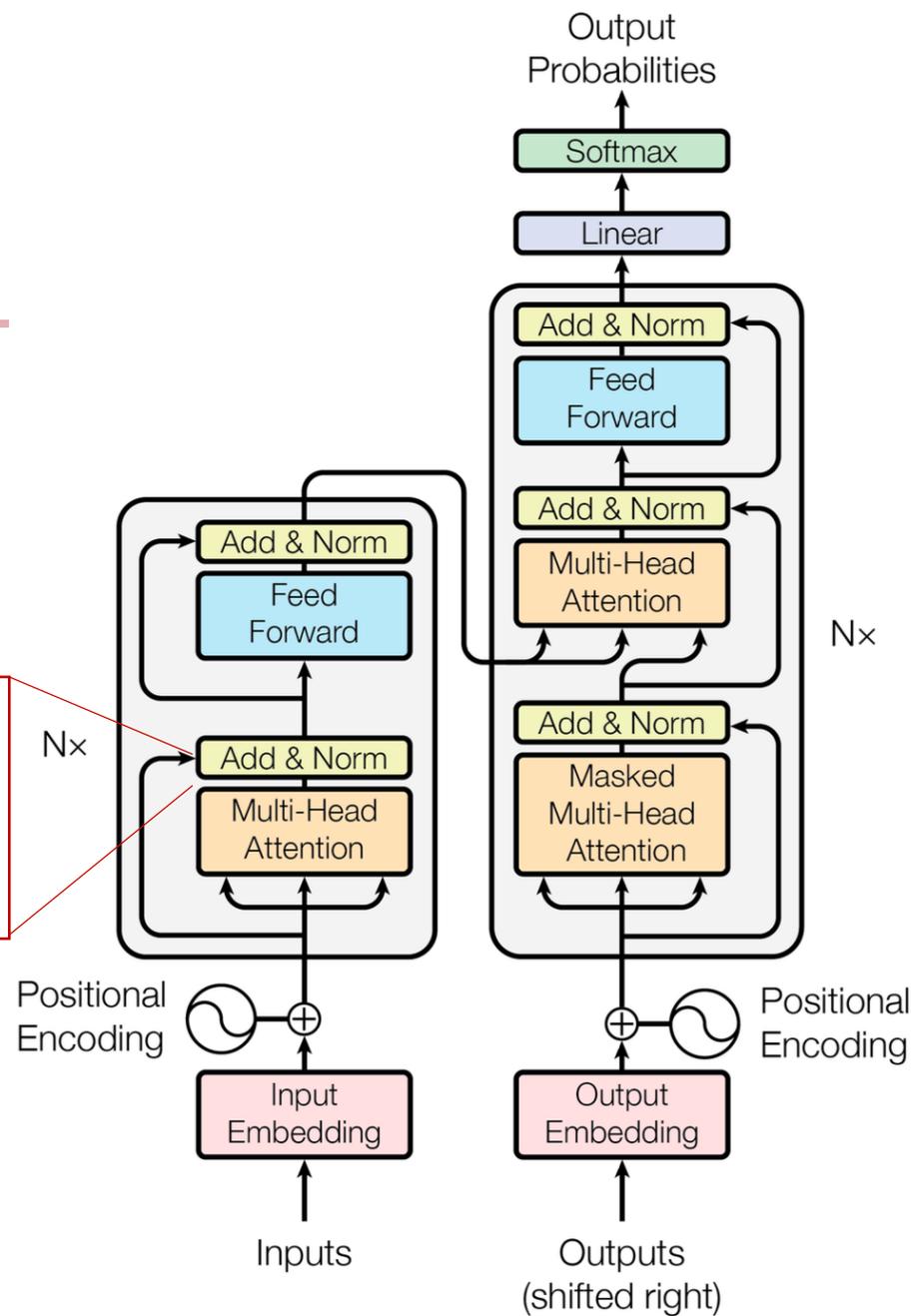
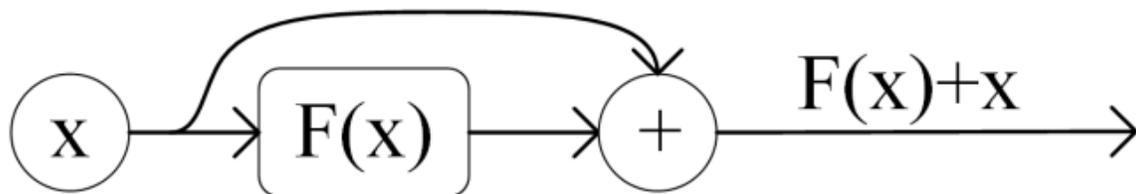
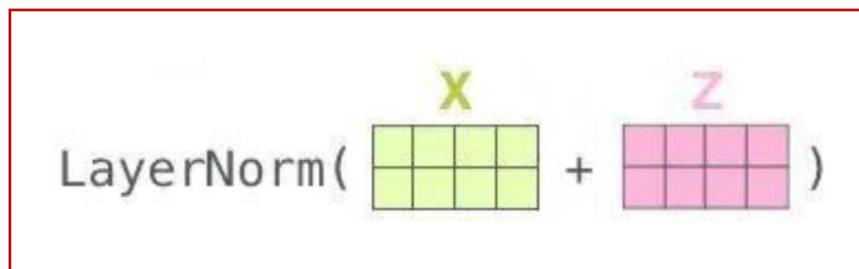
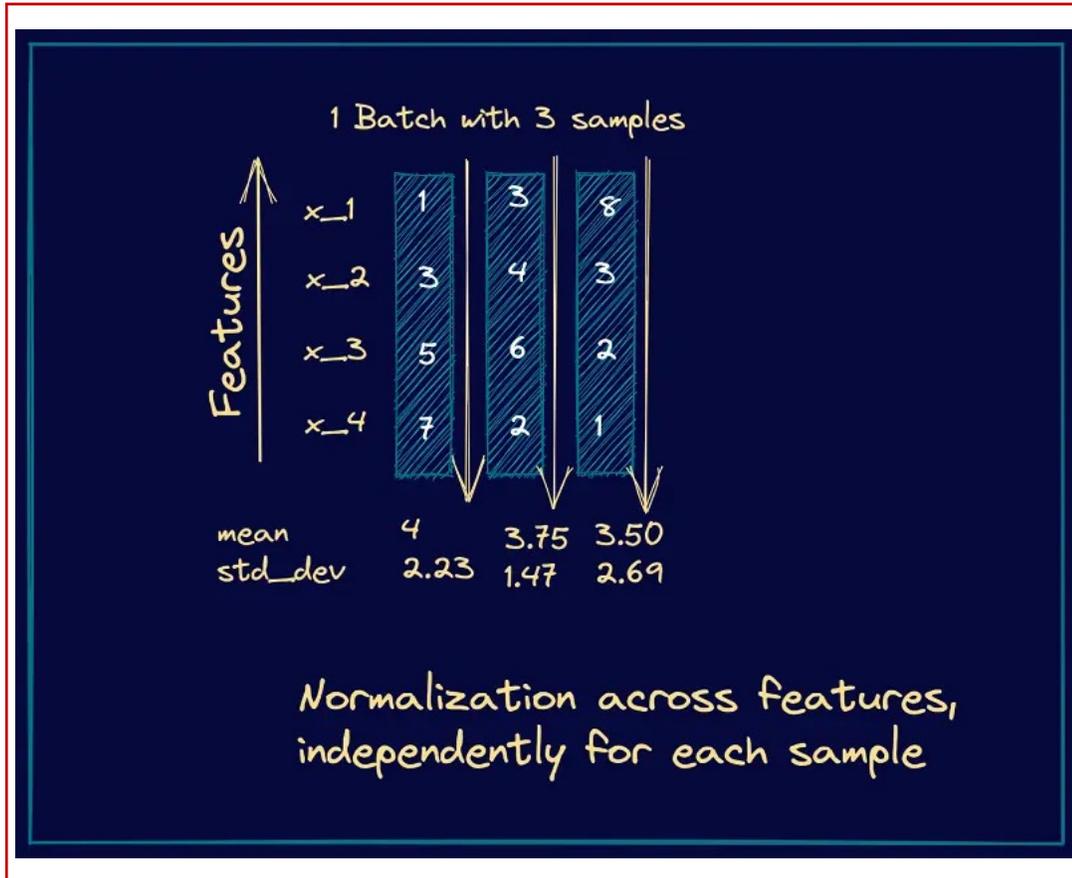


Figure 1: The Transformer - model architecture.

Layer Normalization



<https://arxiv.org/abs/1607.06450>

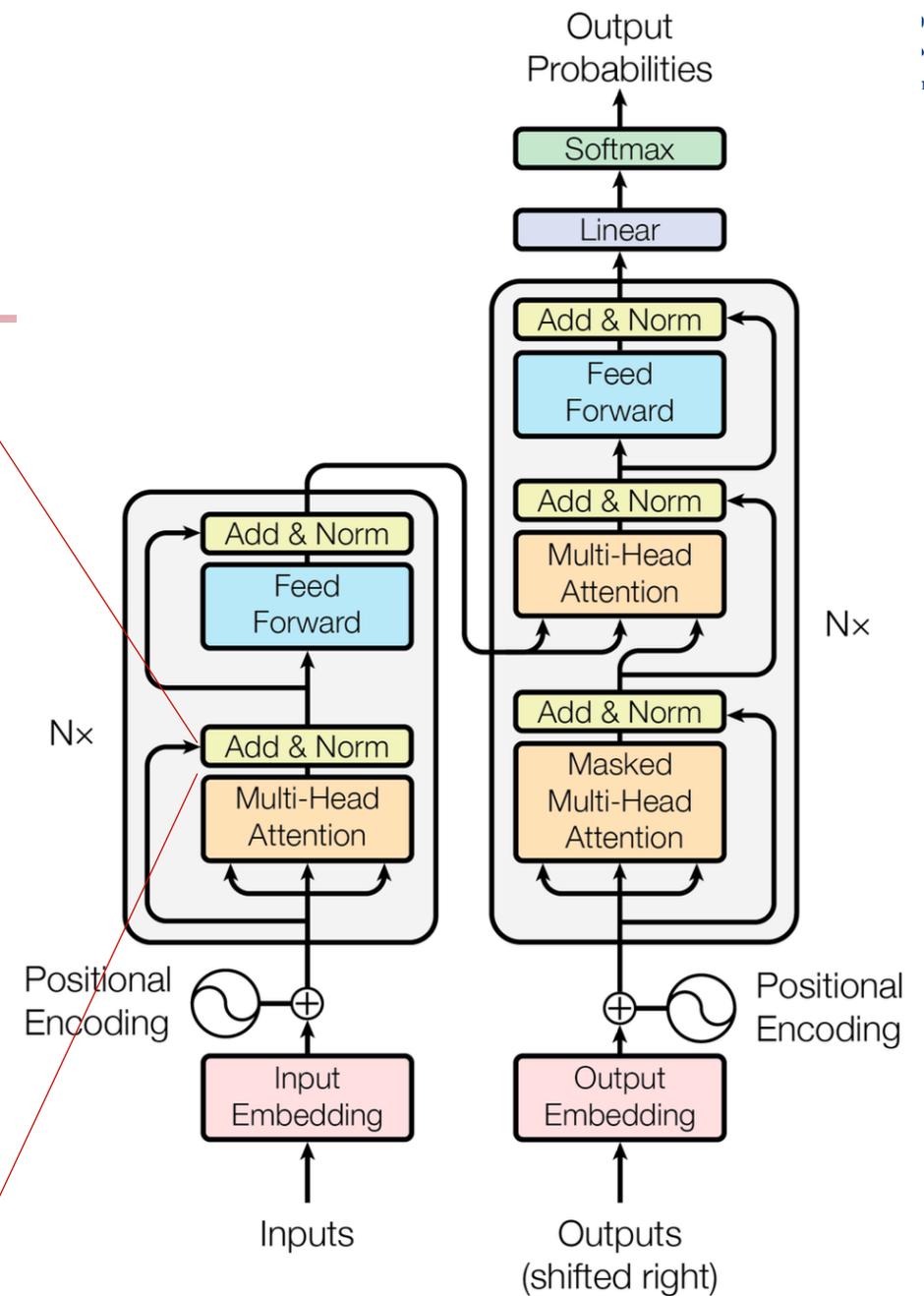
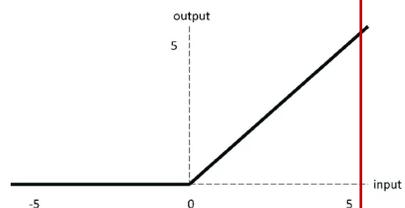


Figure 1: The Transformer - model architecture.

Feed Forward层

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2$$



ReLU激活函数

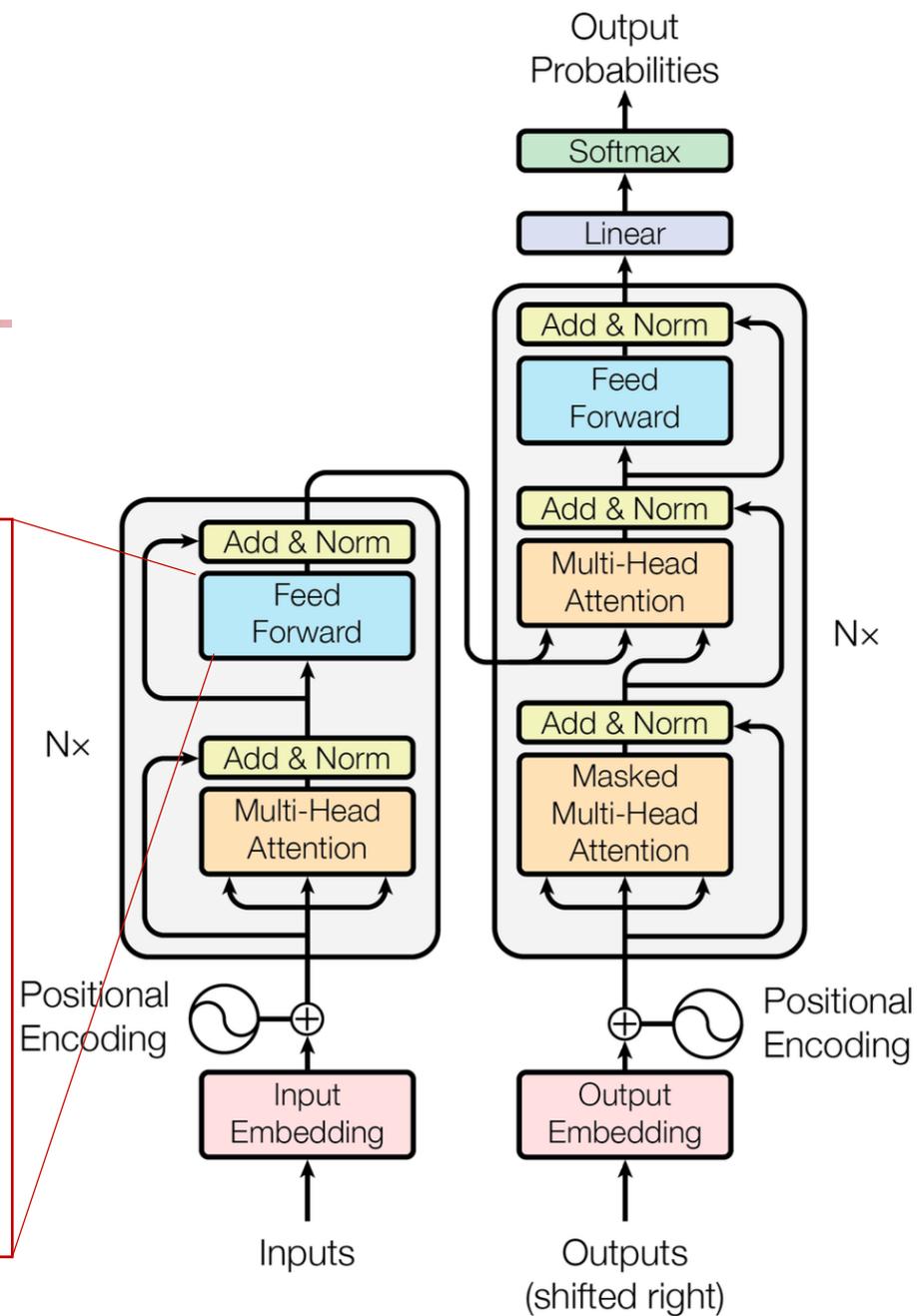
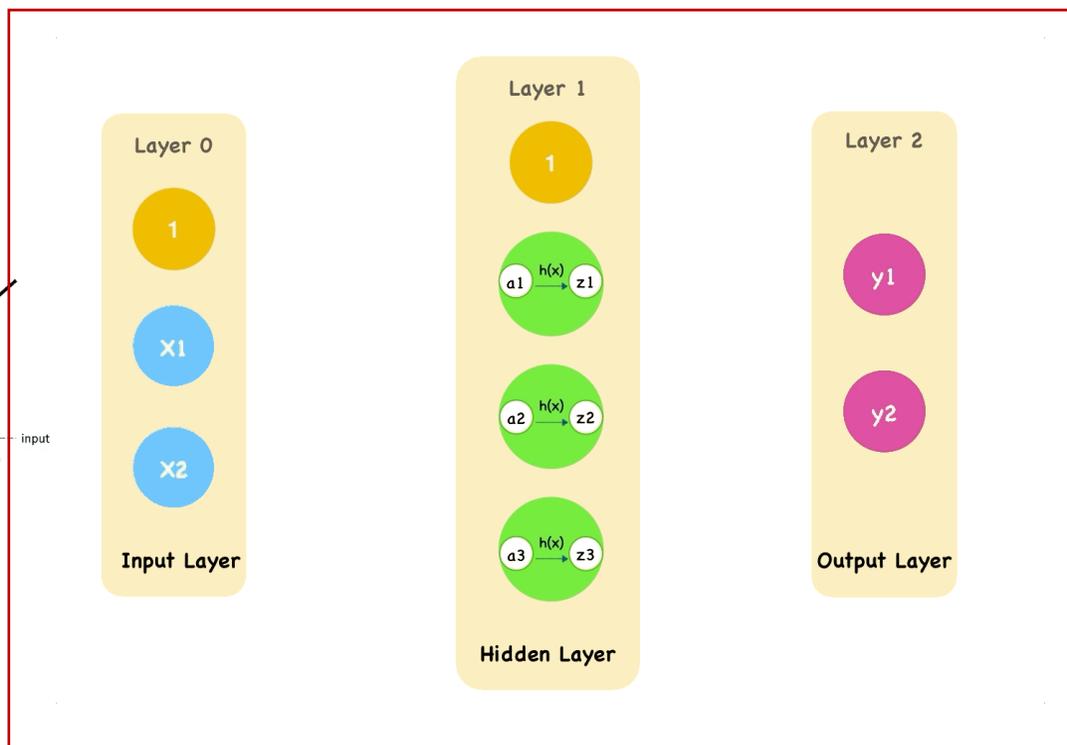


Figure 1: The Transformer - model architecture.

Add&Norm

将token原始编码与经过FFN层后相加

$$\text{LayerNormal}(X + \text{FeedForward}(X))$$

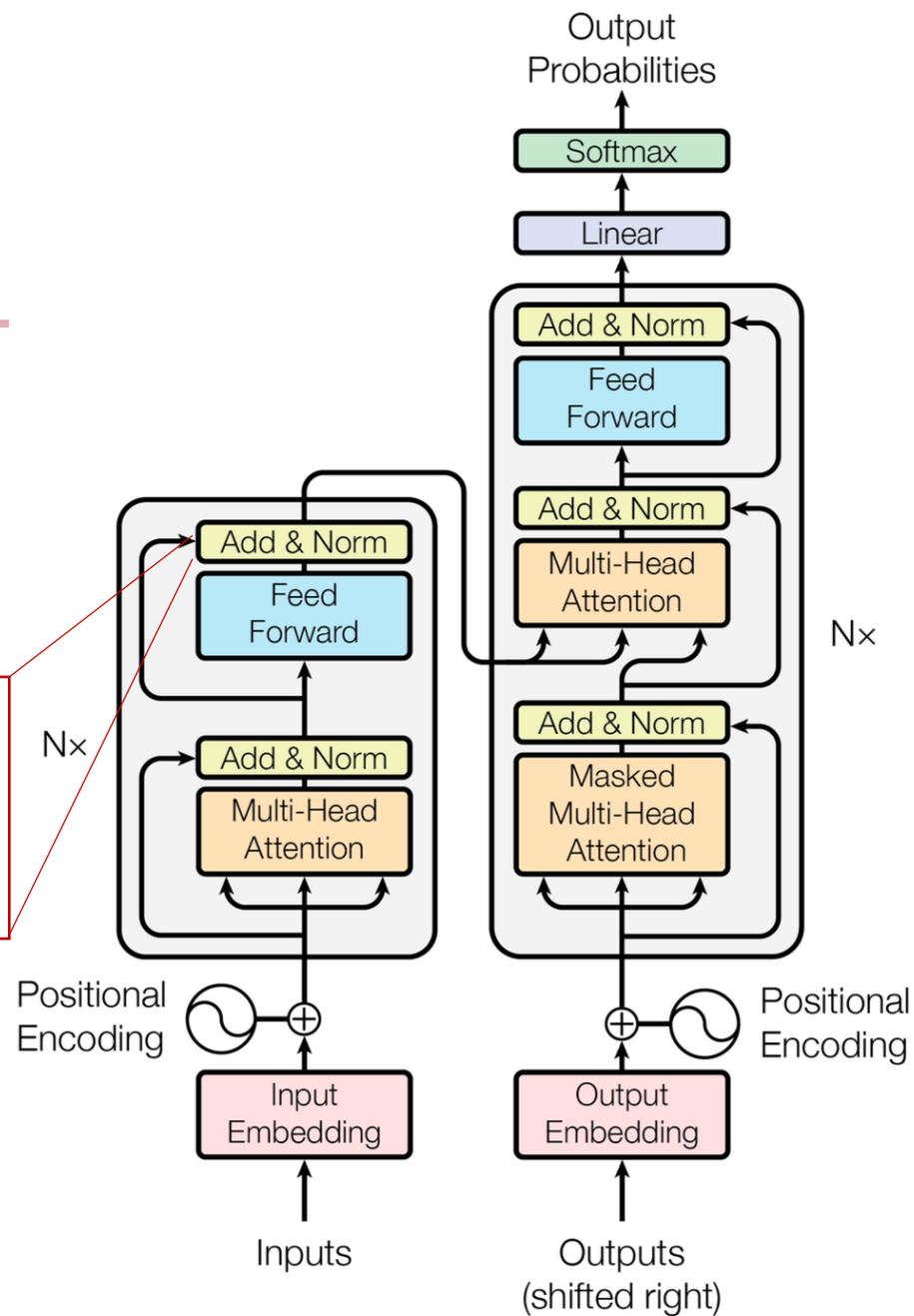
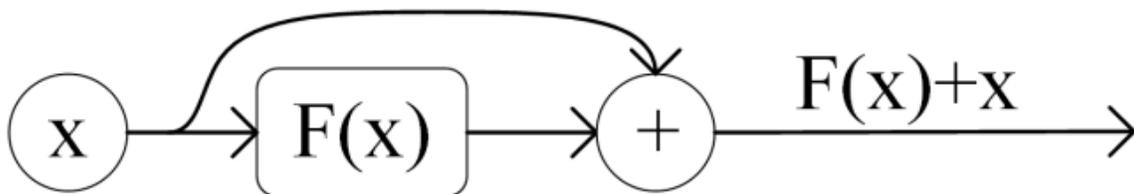


Figure 1: The Transformer - model architecture.

Masked MHA掩码多头注意力

将所有指向未来位置的注意力权重，强行加上一个负无穷大 ($-\infty$) 的值，人为地设置了只能看左侧信息

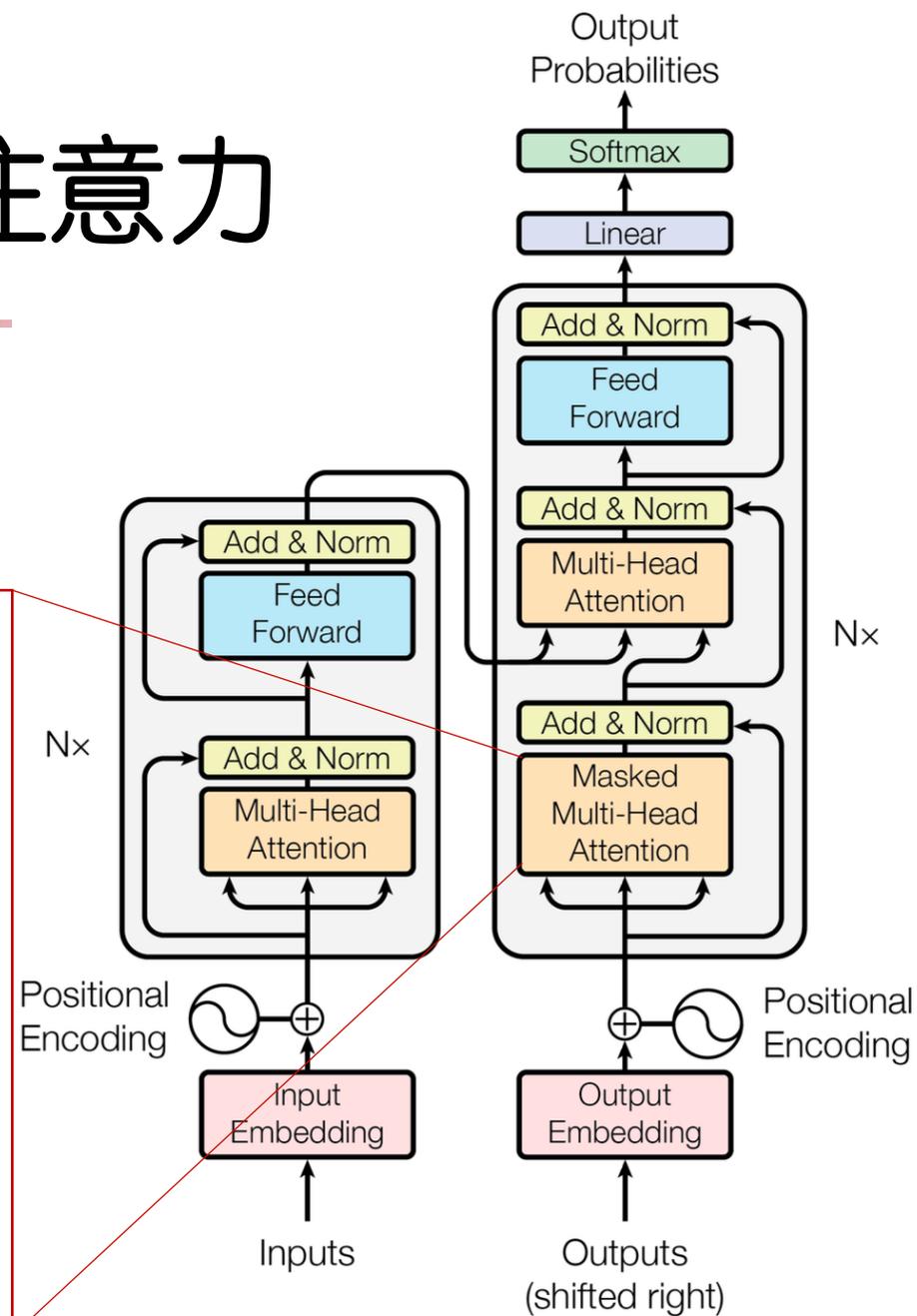
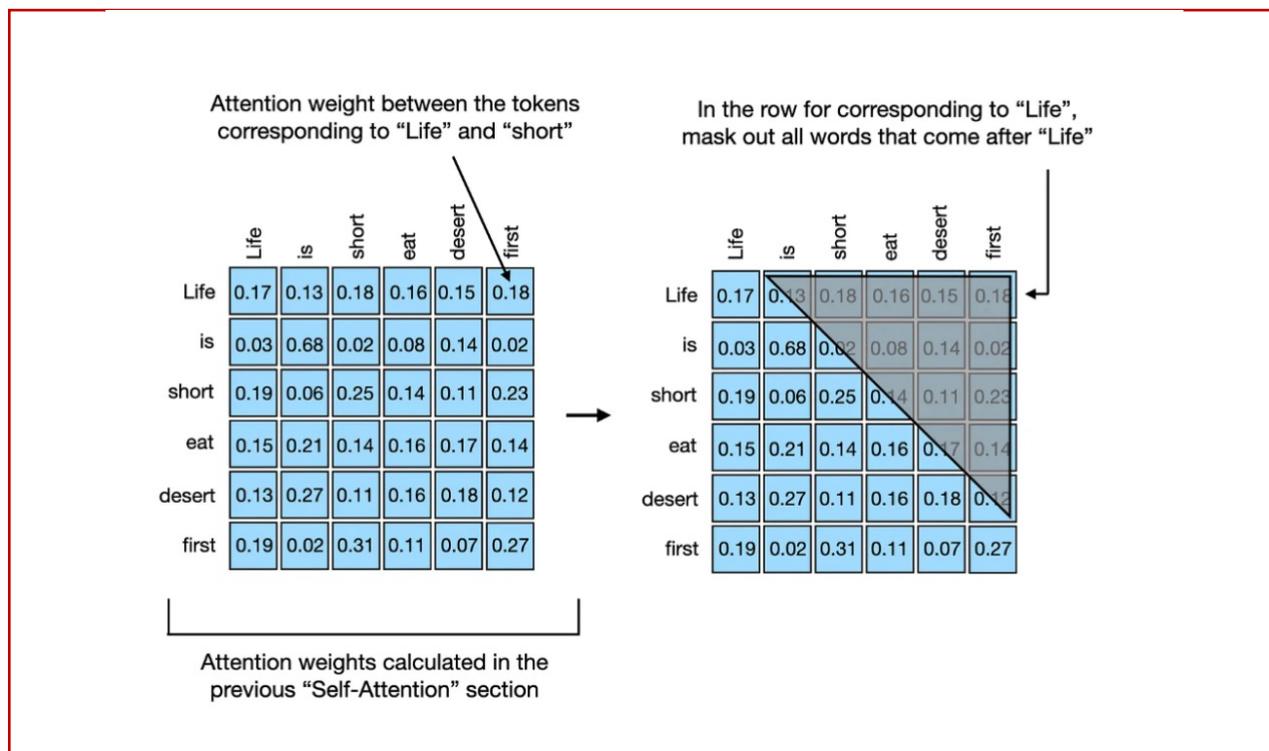


Figure 1: The Transformer - model architecture.

交叉注意力

Q来自解码器, KV来自编码器

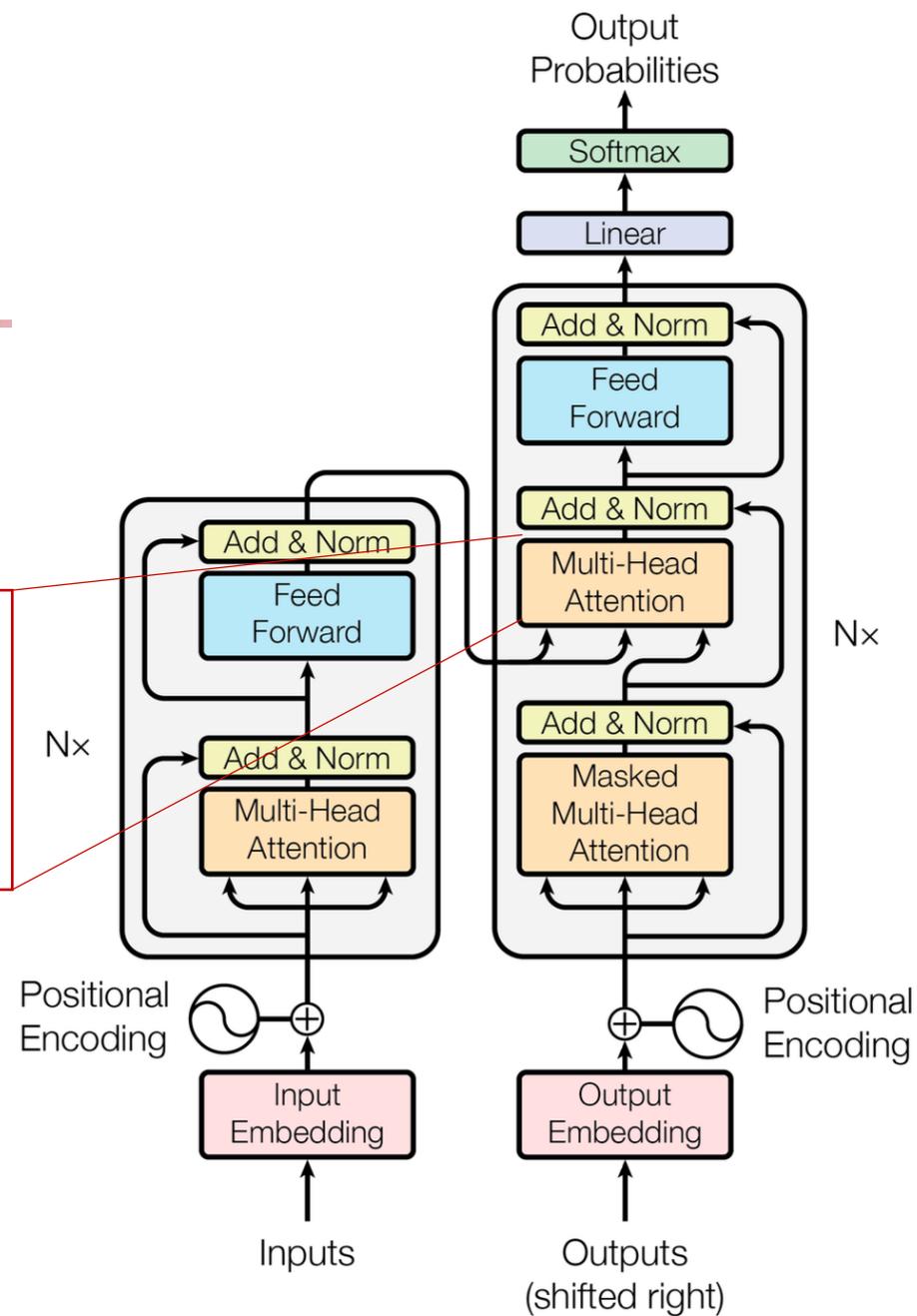


Figure 1: The Transformer - model architecture.

线形层

将注意力值扁平化

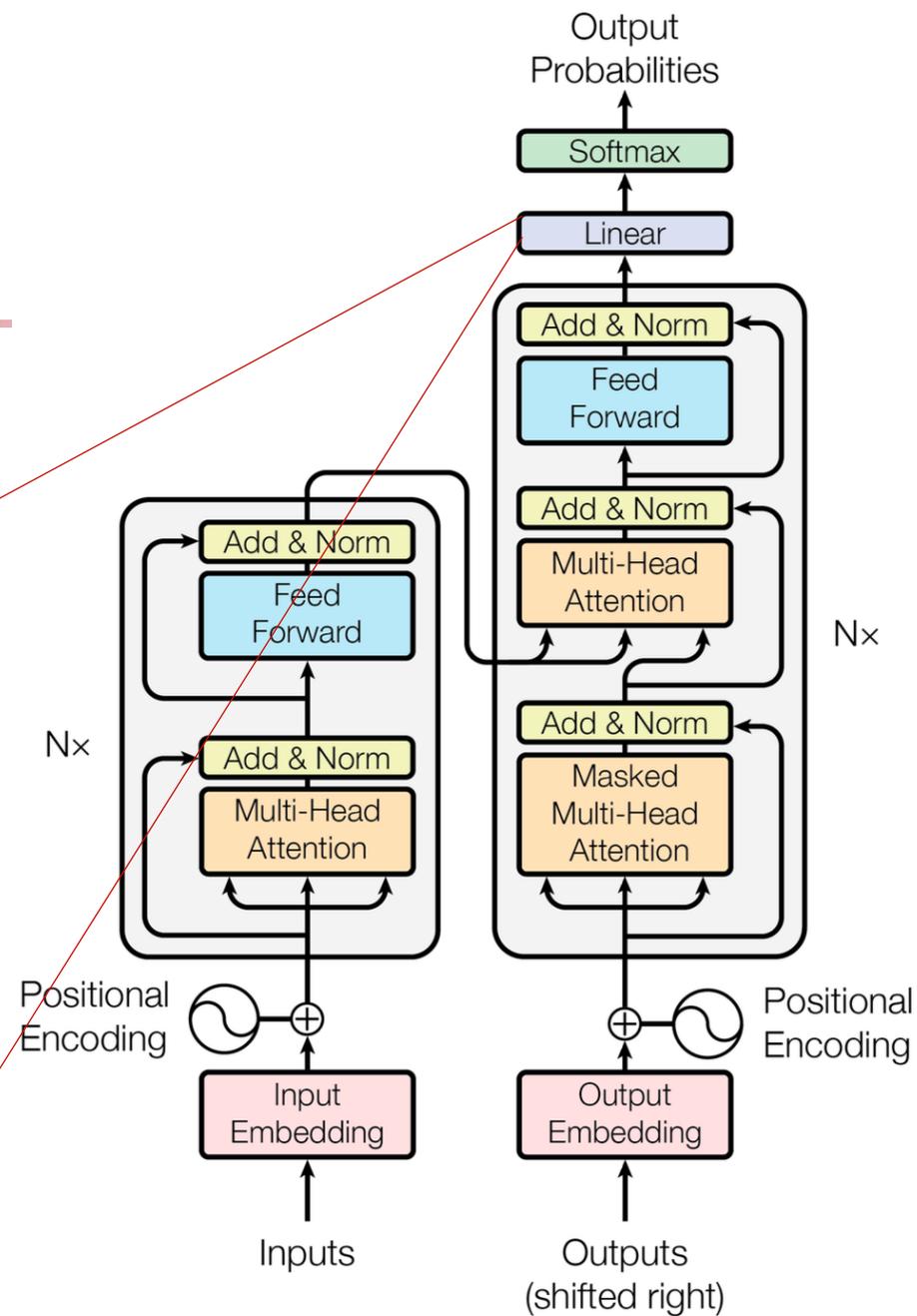
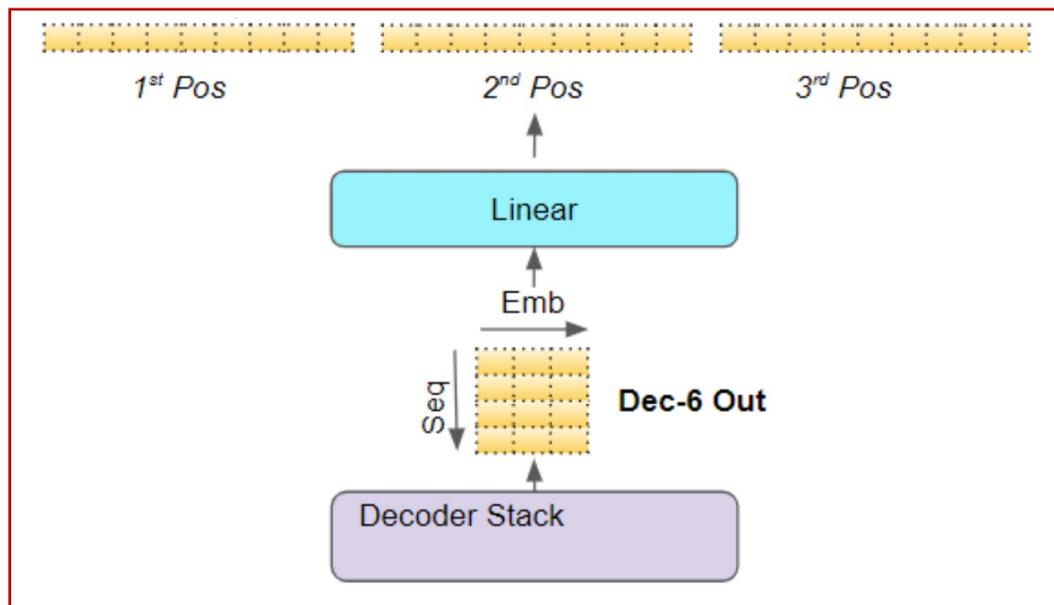


Figure 1: The Transformer - model architecture.

Softmax

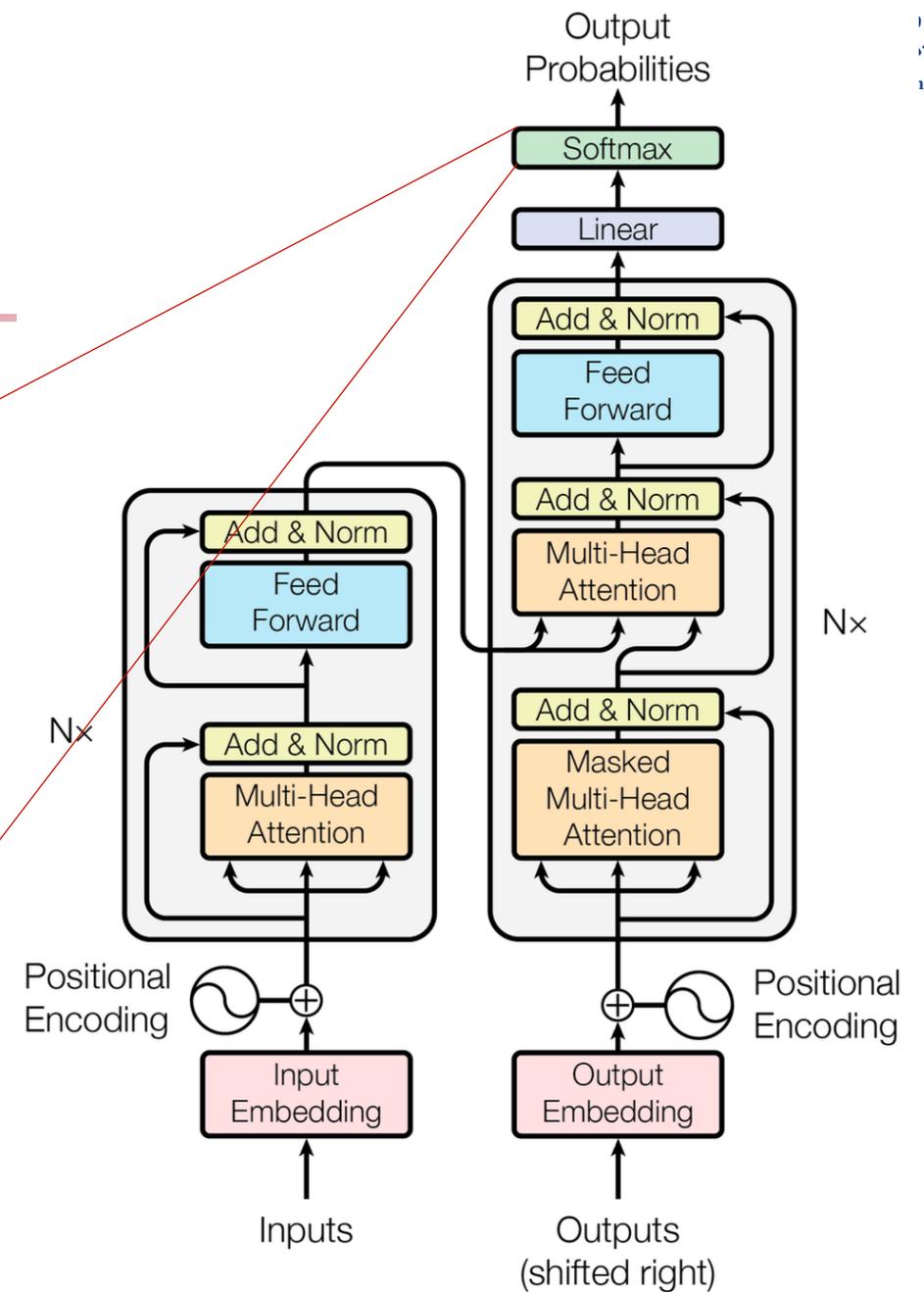
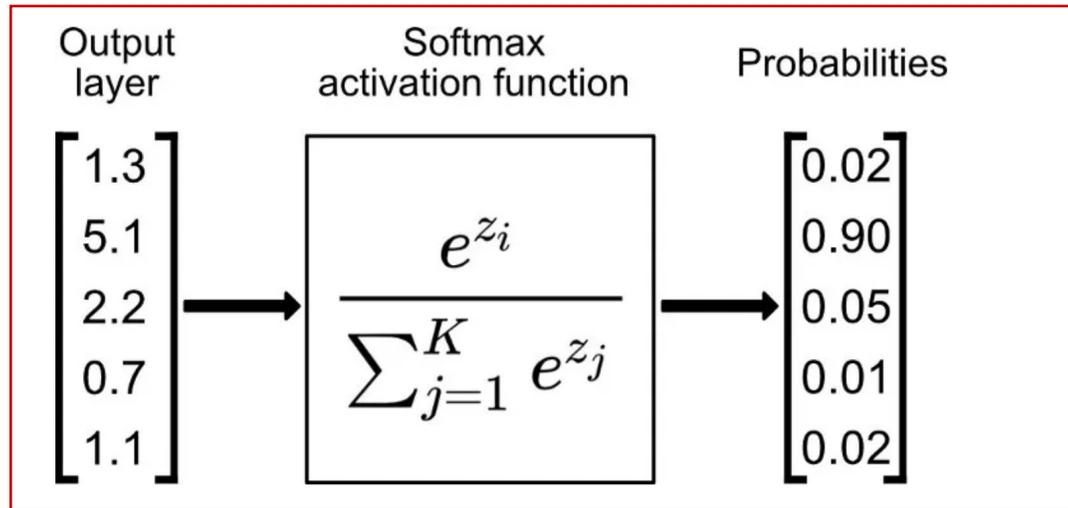
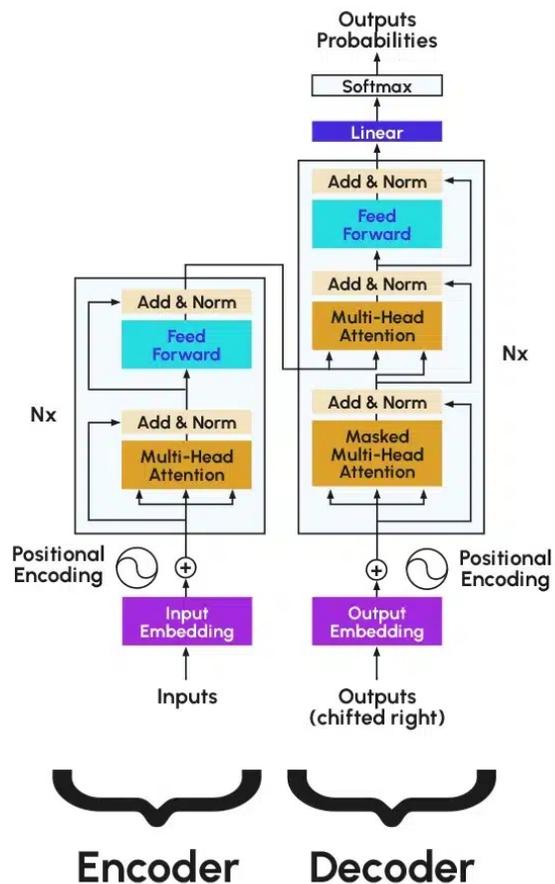


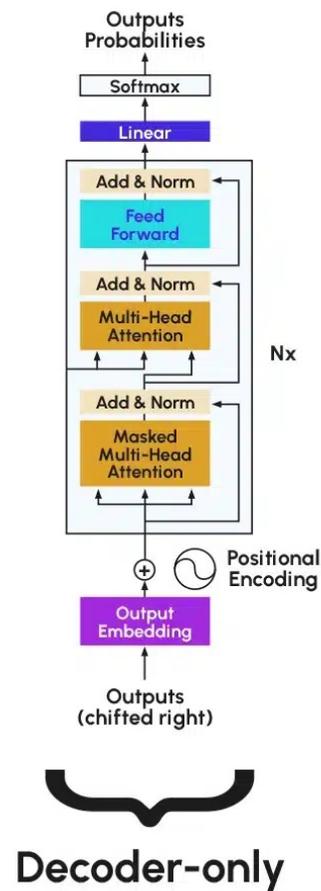
Figure 1: The Transformer - model architecture.

Transformers的变体

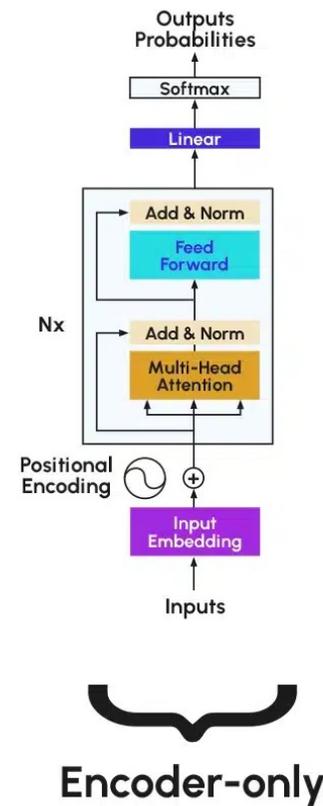
Transformer



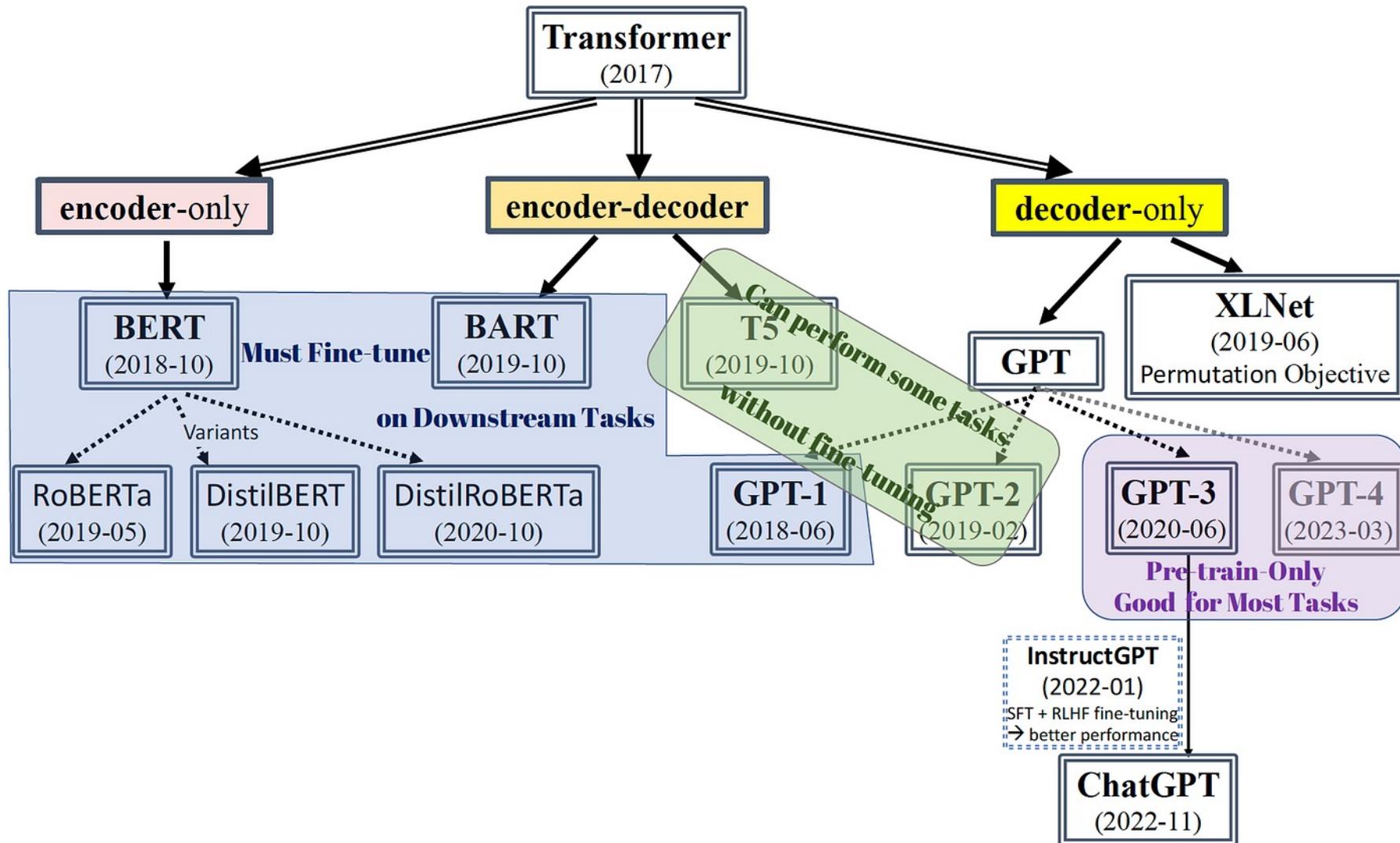
GPT*



BERT*



使用Transformers的模型



本节复习

□ Encoder-Decoder中的Attention

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

□ Position Embedding

□ Transformers Architecture

测测你的注意力



**Count how many times
the players wearing
white pass the ball**

参考文献

- ❑ Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder–decoder for statistical machine translation." *EMNLP*. 2014.
- ❑ Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
- ❑ Su, Jianlin, et al. "Roformer: Enhanced transformer with rotary position embedding." *Neurocomputing* 568 (2024): 127063.

作业

- 阅读“Attention is all you need” 论文
 - 写论文+代码的阅读笔记，并在线发布/提交作业
 - 加分题：代码复现
 - <https://github.com/tensorflow/tensor2tensor>
 - 完成时间：第5周课（4月3号）前

致谢

- 胡玥、曹亚男、方芳：国科大《自然语言处理基础》
- 曹亚男、任昱冰：国科大《深度学习与自然语言处理概述》





THANKS