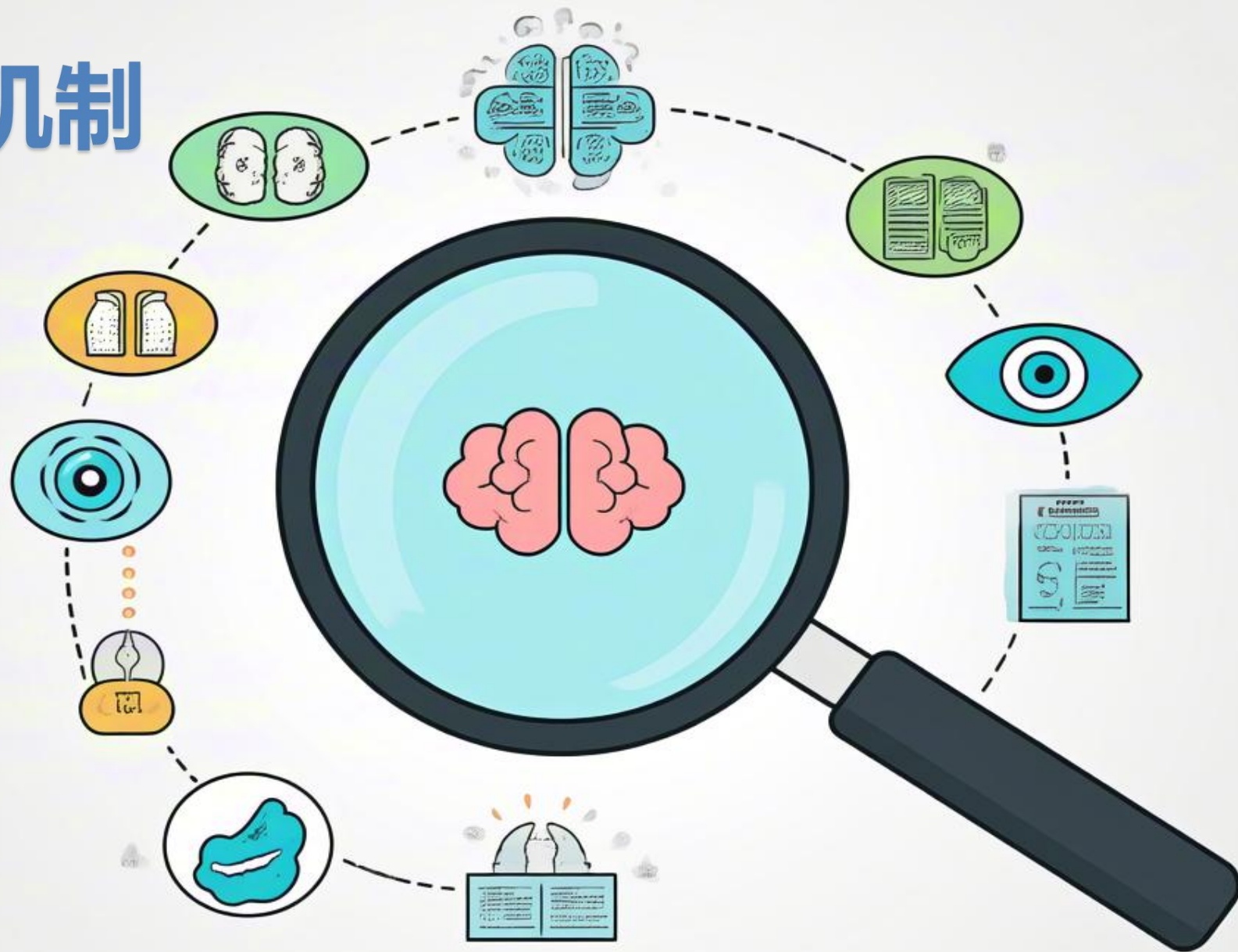


第09章 注意力机制

欧 新 宇

注意力机制





- 什么是注意力机制
- 注意力分数
- 基于注意力的Seq2Seq





什么是注意力机制

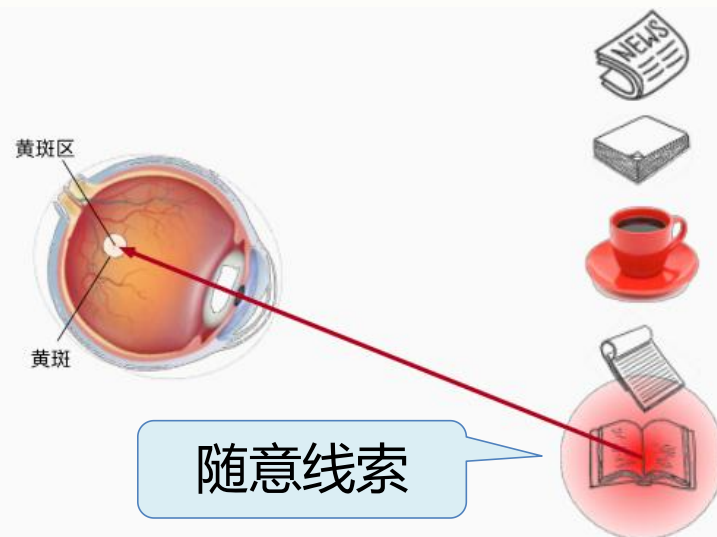
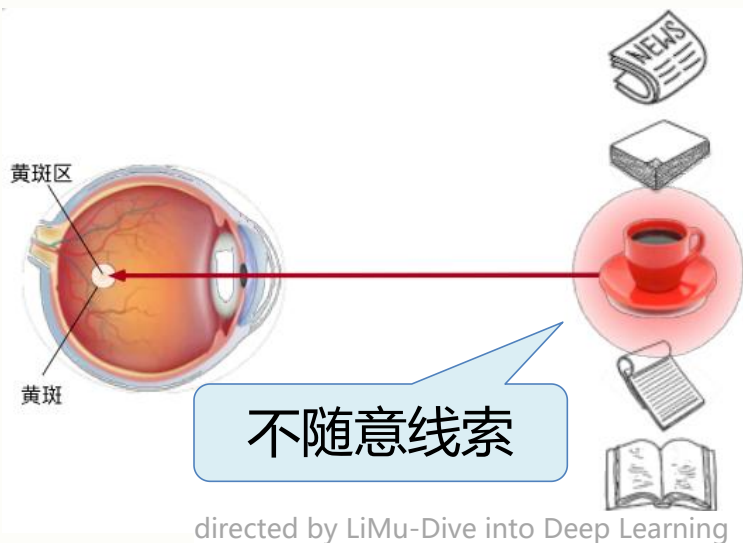


注意力机制

注意力机制的起源

19世纪下半叶，心理学家威廉·詹姆斯在《心理学原理》中提出，注意力是大脑以一种清晰生动的方式在多个同时存在的对象中选择其中一个的过程，它是**意识**对**特定对象**的**集中与聚焦**。

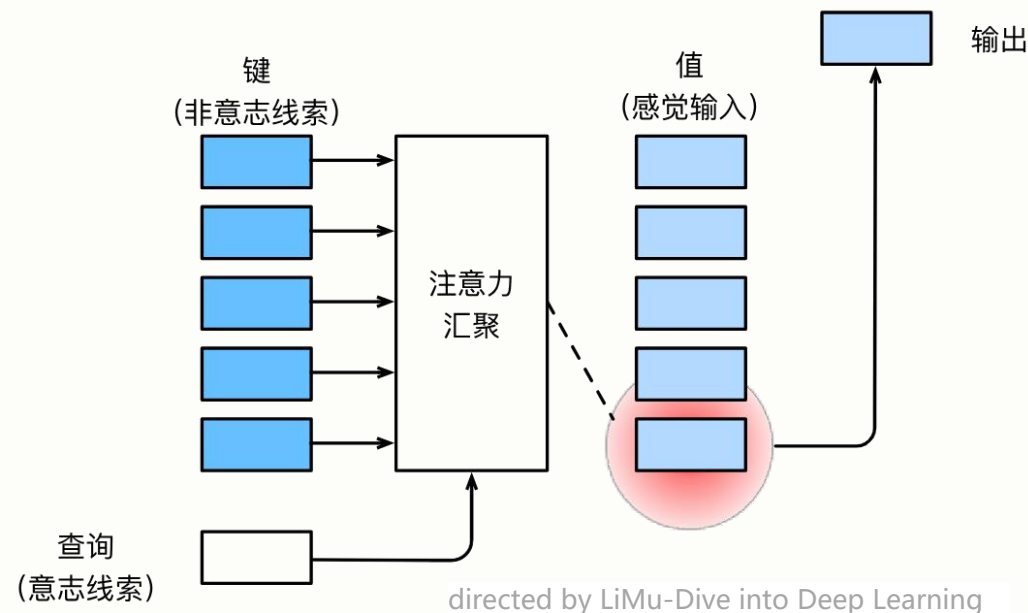
- ✓ **注意力** 是人类在长期进化中形成的一种**生存机制**。
- ✓ **心理学框架**：生物总是会根据**自主的随意线索**和**非自主的不随意线索**选择关注的焦点。



注意力机制

什么是注意力机制

- 卷积、全连接、池化层都只考虑不随意线索
- 注意力机制则显示地考虑随意线索
 - ✓ 随意线索又称为查询 (query)
 - ✓ 输入是不随意线索 (key) 和值 (value) 组成对
 - ✓ 注意力机制通过注意力汇聚层来有偏向性地选择某些输入



非参注意力池化

■ 给定数据对: $(x_i, y_i), i = 1, \dots, n$

✓ 均值池化: $y = f(x) = \frac{1}{n} \sum_i x_i$

✓ 非参注意力池化 (Nadaraya-Watson 核回归):

The diagram shows the formula for Nadaraya-Watson kernel regression:
$$f(x) = \sum_{i=1}^n \frac{K(x - x_i)}{\sum_{j=1}^n K(x - x_j)} y_i$$
 Annotations include:

- A green box labeled "query" pointing to the variable x in the numerator.
- A green box labeled "key" pointing to the variable x_i in the numerator.
- A blue box labeled "value: 加权求和" (value: weighted sum) pointing to the y_i term in the numerator.

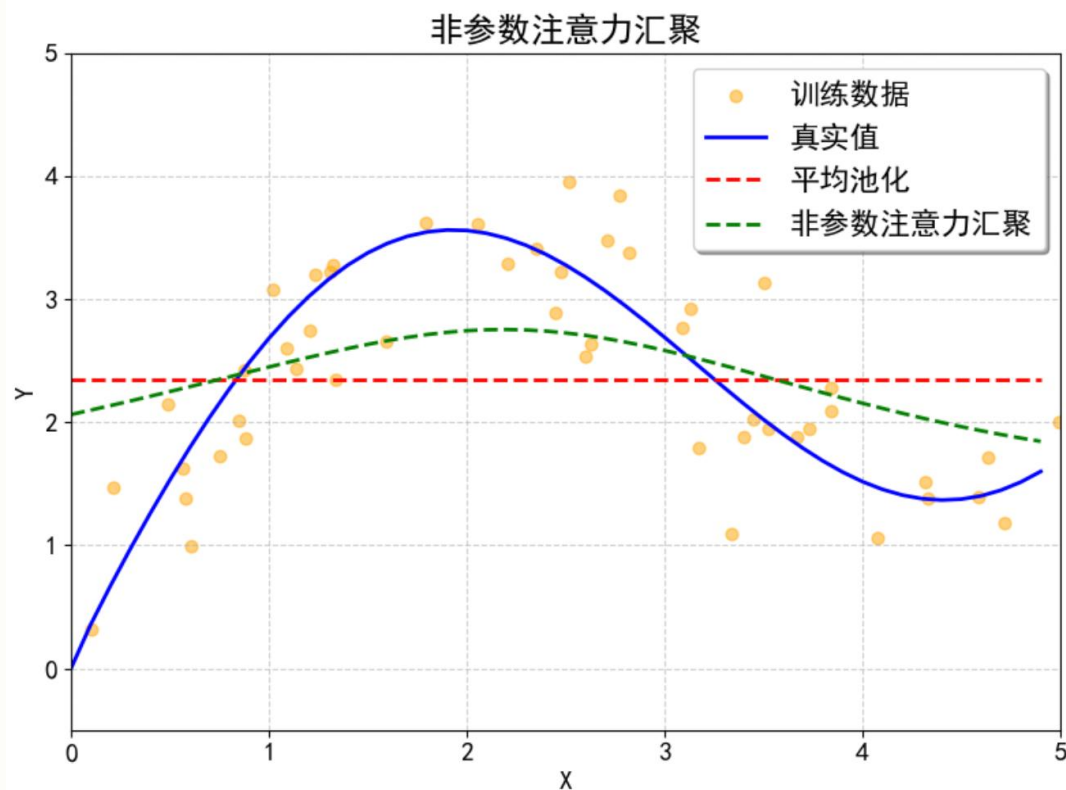
✓ 注意力权重的一般形式: $f(x) = \sum_{i=1}^n \alpha(x, x_i) y_i$

Nadaraya-Watson 核回归

- 高斯核: $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2})$

- 基于高斯核的注意力汇聚:

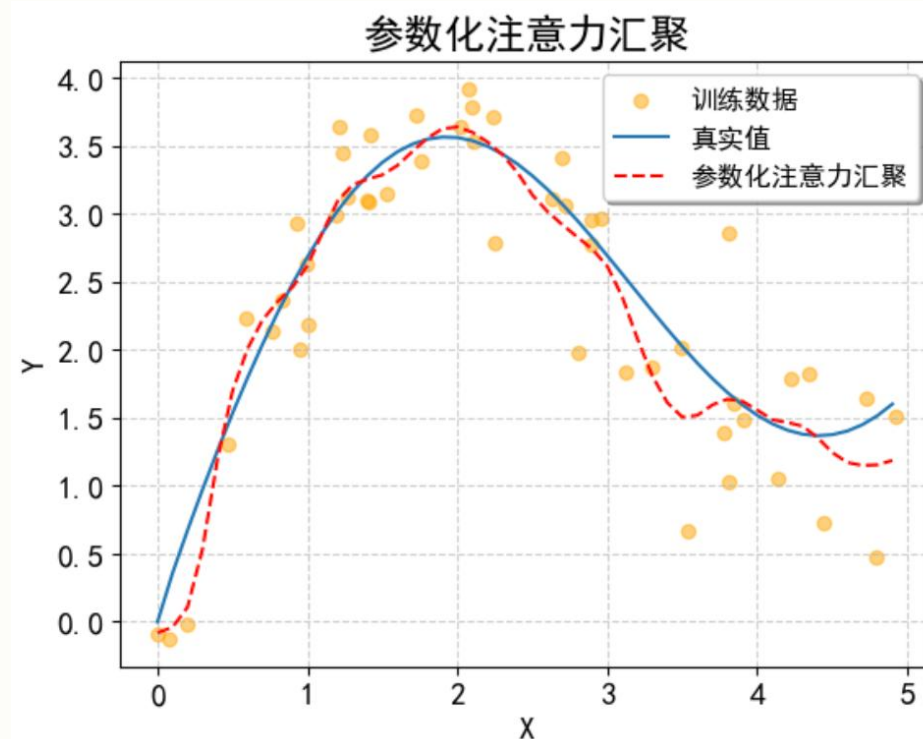
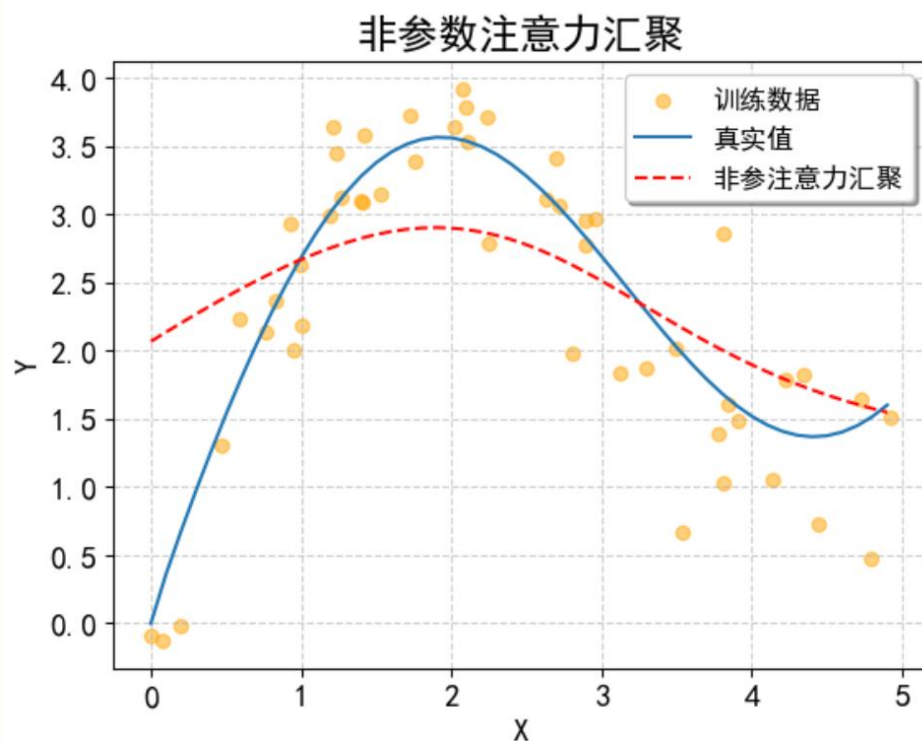
$$\begin{aligned}
 f(x) &= \sum_{i=1}^n \alpha(x, x_i) y_i \\
 &= \sum_{i=1}^n \frac{\exp(-\frac{1}{2}(x - x_i)^2)}{\sum_{j=1}^n \exp(-\frac{1}{2}(x - x_j)^2)} y_i \\
 &= \sum_{j=1}^n \text{softmax}(-\frac{1}{2}(x - x_j)^2) y_i
 \end{aligned}$$



注意力机制

参数化的注意力机制

$$f(x) = \sum_{i=1}^n \alpha(x, x_i) w y_i = \sum_{j=1}^n \text{softmax}\left(-\frac{1}{2}(x - x_i)w^2\right) y_i$$



注意力机制

小 结

- 人类的注意力是有限的、有价值 and 稀缺的资源。
- 心理学：自主性选择的**随意线索**与非自主性感知的**不随意线索**
- 注意力机制与全连接层或者汇聚层的主要区别在于**增加了自主提示**，它通过注意力使选择偏向于更有意义的值（value，感官输入），它来源于查询（query，自主性提示）和键（key，非自主性提示）的交互。
- 注意力机制的一般形式 $f(x) = \sum_{i=1}^n a(x, x_i) y_i$ ，其中， $a(x, x_i)$ 是注意力权重
- 注意力机制实现了“**既凭直觉抓取模式，又按目标筛选信息**”的平衡





注意力分数

注意力分数

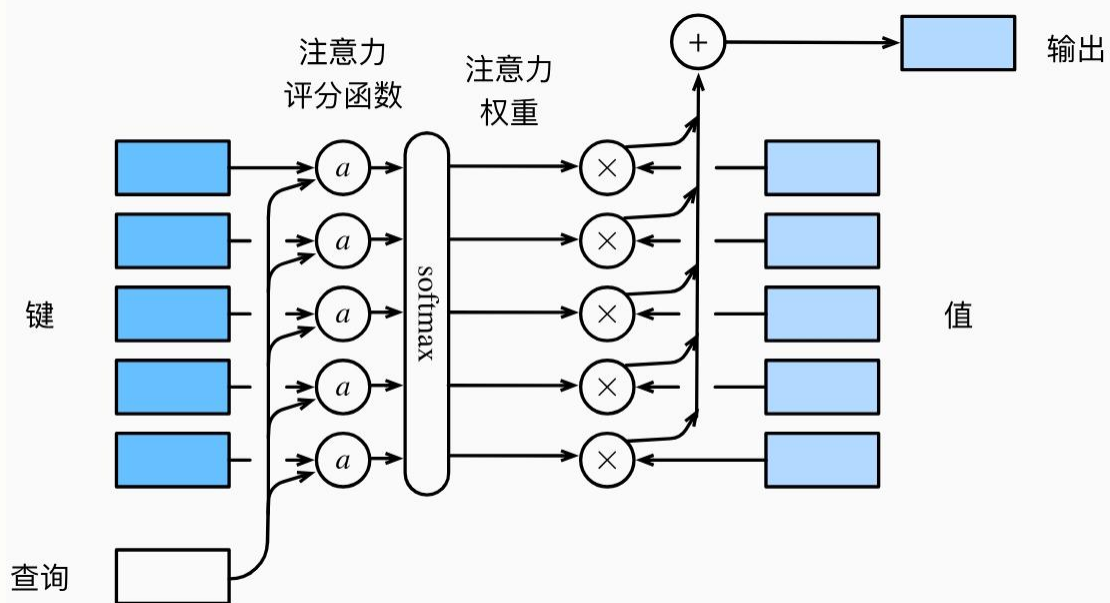
回顾:

注意力权重

注意力分数

注意力分数

$$f(x) = \sum_{i=1}^n \alpha(x, x_i) y_i = \sum_{j=1}^n \text{softmax}(-\frac{1}{2}(x - x_i)^2) y_i$$



注意力分数

高维度注意力

假设 query $q \in \mathbb{R}^q$, m 对 key-value $(k_1, v_1), \dots$, 这里 $k_i \in \mathbb{R}^k$, $v_i \in \mathbb{R}^v$

则注意力池化层变为:

$$f(q, (k_1, v_1), \dots, k_m, v_m) = \sum_{i=1}^n \alpha(q, k_i) v_i \in \mathbb{R}^v$$

$$\alpha(q, k_i) = \text{softmax}(\alpha(q, k_i)) = \frac{\exp(\alpha(q, k_i))}{\sum_{j=1}^n \exp(\alpha(q, k_i))} \in \mathbb{R}$$

注意力分数

注意力分数

加性注意力 (Additive Attention)

- 给定查询向量 q , 键向量 k
- 可学习参数: $W_k \in \mathbb{R}^{h \times k}, W_q \in \mathbb{R}^{h \times q}, v \in \mathbb{R}^h$
- 相似度得分:

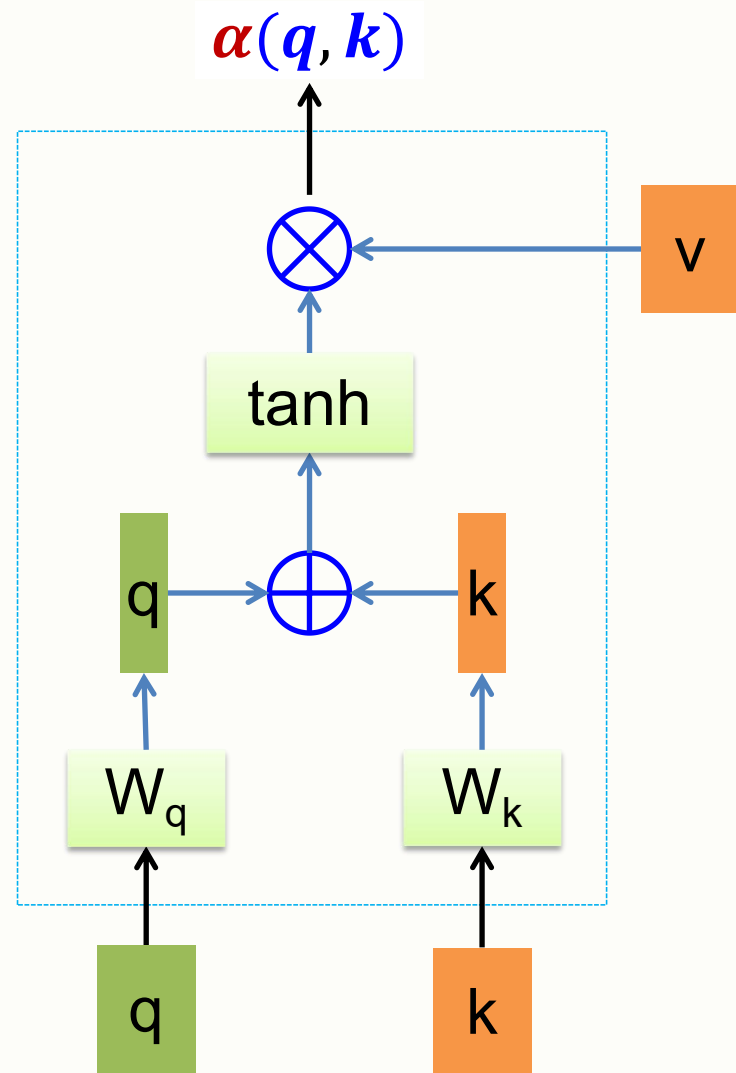
$$\alpha(q, k) = v^T \tanh(W_k k + W_q q)$$

加性融合

非线性激活

加权求和

- 等价于一个大小为 h , 输出为 1 的单隐层 MLP
- 支持不同维度且需要复杂交互的场景
- 表达能力强, 但计算开销大



注意力分数

缩放点积注意力 (Scaled Dot-Product Attention)

当 query 和 key 的长度相等时, 且 $q, k_i \in \mathbb{R}^q$, 那么注意力分数可以表示为:

$$\alpha(q, k_i) = \langle q, k_i \rangle / \sqrt{d}$$

向量化版本:

✓ $Q \in \mathbb{R}^{n \times d}, K_q \in \mathbb{R}^{m \times d}, V \in \mathbb{R}^{m \times d}$

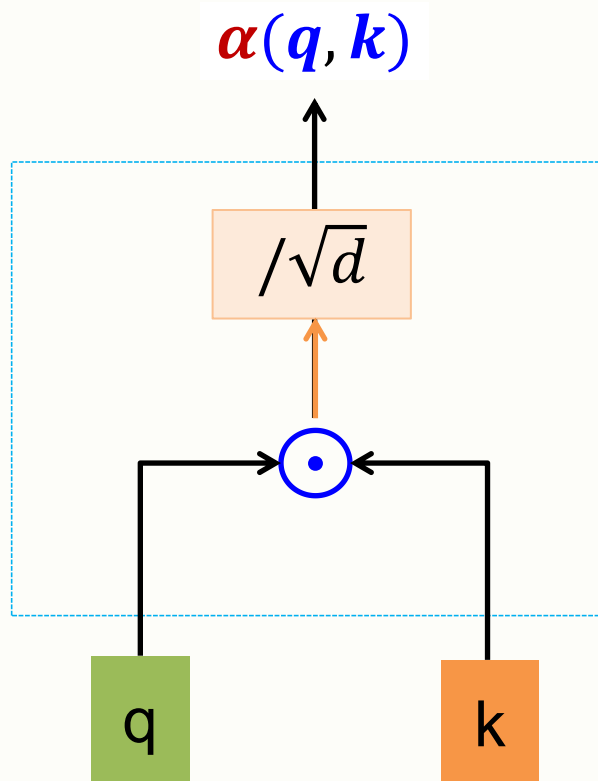
✓ 注意力分数: $\alpha(Q, K) = QK^T / \sqrt{d} \in \mathbb{R}^{n \times m}$

✓ 注意力权重:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \in \mathbb{R}^{n \times v}$$

加权求和

缩放因子

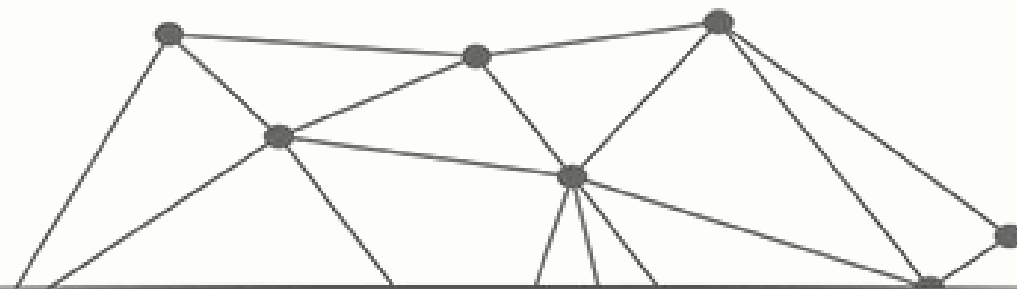


小结

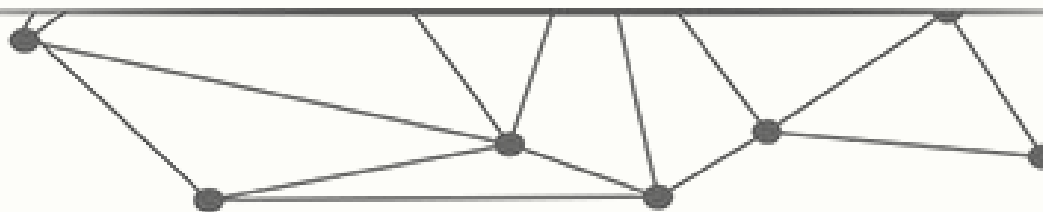
- 注意力分数衡量了查询向量 (query) 与键向量 (key) 之间的**相似度**，它反映的是模型对不同键值对的**关注程度**。
- Softmax 实现将注意力分数转化为更加**稳定的注意力权重**，确保了模型的有效训练。
- 加性注意力 (Additive Attention) 和缩放点积注意力 (Scaled Dot-Product Attention) 是两种常见的注意力分数计算方法。
 - ✓ **加性注意力**：将 query 和 key 合并起来进入一个单输出单隐藏层的MLP，适用于任务复杂、且需要强拟合的场景
 - ✓ **缩放点积注意力**：直接将 query 和 key 做内积，复杂度低，但更高效。



基于注意力的Seq2Seq



课堂互动 9.1



读万卷书 行万里路 只为最好的修炼



QQ: 14777591 (宇宙骑士)

Email: ouxinyu@alumni.hust.edu.cn

Website: <http://ouxinyu.cn>

Tel: 18687840023

地址: 安宁校区 诚远楼201

南院 智能应用研究院A306-2