

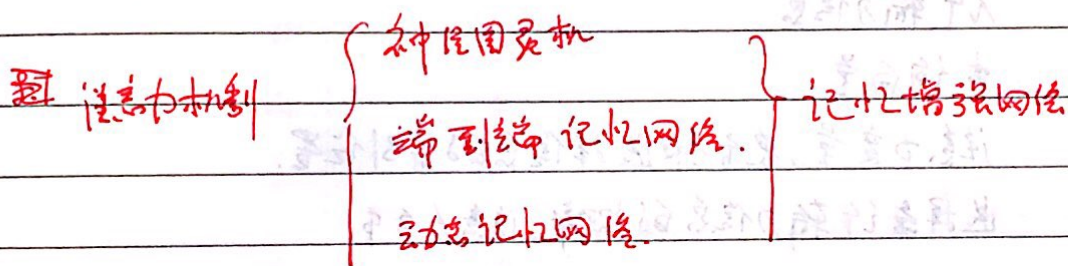
第8章 chap-8 注意力机制与外部记忆

模型复杂度 — 表达能力

网络容量

注意力: 通过自上而下的信息选择机制来过滤掉大量无关信息

外部记忆: 优化神经网络的记忆结构来提高存储容量.



自注意力: 可以作为神经网络中的一层来使用, 有效地解决长距离依赖问题.

记忆单元

外部记忆 不直接参与运算 通过接口进行操作.

记忆周期 神经网络

短期 状态 (神经元活性)

中期 外部记忆

长期 可学习参数

存储方式

内容寻址为主

8.1 注意力

特点：自下而上信息，主动 (主动)

显著性：自上而下信息，被动

8.1.1 注意力机制

自下而上显著性 $\left\{ \begin{array}{l} \text{max pooling} \\ \text{门控} \end{array} \right.$

\bar{x}_{in} 一个输入信息

\vec{q} 查询向量

z 注意力变量，表示被选择信息的索引位置。

α_i 选择每个输入信息的概率，注意力分布

$$\alpha_i = p(z=i | \bar{x}_{in}, \vec{q})$$

$$= \text{softmax}(s(\bar{x}_i, \vec{q}))$$

$$= \frac{\exp(s(\bar{x}_i, \vec{q}))}{\sum_{j=1}^N \exp(s(\bar{x}_j, \vec{q}))}$$

$s(\bar{x}_i, \vec{q})$ 注意力打分函数

加性模型

$$s(\bar{x}_i, \vec{q}) = v^T \tanh(w \bar{x}_i + u \vec{q})$$

点积模型

$$s(\bar{x}_i, \vec{q}) = \bar{x}_i \vec{q}$$

缩放点积模型

$$s(\bar{x}_i, \vec{q}) = (\bar{x}_i \vec{q}) / \sqrt{d} \quad (\text{推荐})$$

双线性模型

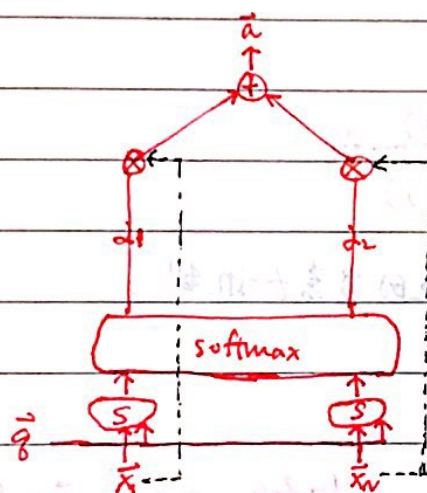
$$s(\bar{x}_i, \vec{q}) = \bar{x}_i^T W \vec{q}$$

注意力分布 α_i 可解释为上下文查询 q 时, 第 i 个

软注意力机制

$$\text{att}(X, \vec{q}) = \sum \alpha_i \vec{x}_i$$

$$= \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} | X, \vec{q})} [\vec{x}]$$



x_i 乘 $q \Rightarrow s(x_i, \vec{q})$ 即分数

$s(x_i, \vec{q})$ 经过 softmax $\Rightarrow \alpha_i$

α_i 乘 \vec{x}_i 信息 $\Rightarrow \text{att}(X, \vec{q})$

8.1.3 注意力机制的变体

\Rightarrow 硬性注意力

区别:

只关注到某一位置上的信息, 而不是注意力分布下的期望

1. 选取最高概率率的输入信息

$$\text{att}(X, \vec{q}) = \vec{x}_j, \quad j = \arg \max_{i=1}^N \alpha_i$$

2. 注意力分布 α 的随机采样

No.

Date.

⇒ 键值对注意力

键: 计算注意力分布 d_i

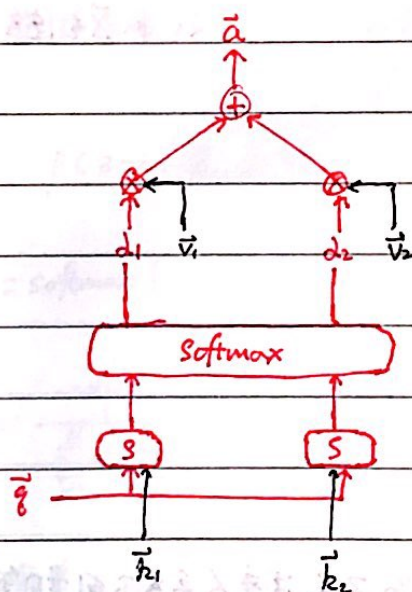
值: 计算聚合信息

$$(k, v) = [(\vec{k}_1, \vec{v}_1), \dots, (\vec{k}_N, \vec{v}_N)] \quad N \text{ 个输入信息}$$

$$\text{att}((k, v), \vec{q}) = \sum d_i \vec{v}_i$$

$$= \sum \frac{\exp(s(\vec{k}_i, \vec{q}))}{\sum \exp(s(\vec{k}_j, \vec{q}))} \vec{v}_i$$

当 $k=v$ 时, 键值对注意力等价于普通的注意力机制



k_i 和 $q \Rightarrow s(\vec{k}_i, \vec{q})$ 分数

$s(\vec{k}_i, \vec{q})$ 经过 softmax $\Rightarrow d_i$

d_i 和 \vec{v}_i 信息

⇒ 多头注意力

利用多个查询 $Q = [\vec{q}_1, \dots, \vec{q}_m]$ 来平行地计算从输入信息中选取多个信息, 每个注意力关注不同部分。

$$\text{att}((k, v), Q) = \text{att}((k, v), \vec{q}_1) \oplus \dots \oplus \text{att}((k, v), \vec{q}_m)$$

No.

Date.

⇒ 结构化注意

层次化解注意力