Attention 升级面

来自： AiGC面试宝典

宁静致远

2023年09月29日 13:26

1 传统 Attention 存在哪些问题?

1. 传统 Attention 存在 上下文长度 约束问题；
2. 传统 Attention 速度慢，内存占用大；

2 Attention 优化方向

1. 提升上下文长度
2. 加速、减少内存占用

3 Attention 变体有哪些?

- 稀疏 attention。将稀疏偏差引入 attention 机制可以降低了复杂性；
- 线性化 attention。解开 attention 矩阵与内核特征图，然后以相反的顺序计算 attention 以实现线性复杂度；
- 原型和内存压缩。这类方法减少了查询或键值记忆对的数量，以减少注意力矩阵的大小；
- 低阶 self-Attention。这一系列工作捕获了 self-Attention 的低阶属性；
- Attention 与先验。该研究探索了用先验 attention 分布来补充或替代标准 attention；
- 改进多头机制。该系列研究探索了不同的替代多头机制。

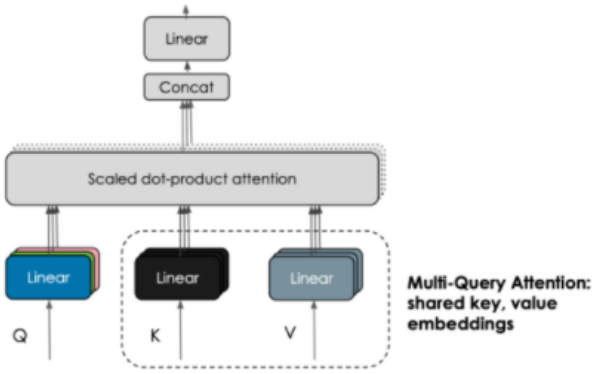
4 Multi-Query Attention 篇

4.1 Multi-head Attention 存在什么问题?

- 训练过程：不会显著影响训练过程，训练速度不变，会引起非常细微的模型效果损失；
- 推理过程：反复加载 巨大 的 KV cache , 导致 内存开销大，性能是内存受限；

4.2 介绍一下 Multi-Query Attention?

Multi-Query Attention 在所有注意力头上 共享 key 和 value.



4.3 对比一下 Multi-head Attention 和 Multi-Query Attention?

- Multi-head Attention：每个注意力头都有各自的query、key和value。
- Multi-query Attention: 在所有的注意力头上共享key和value。

模型	n_heads	head_dim	FFN中间维度	维度h
LLaMA	32	128	11008	4096
baichuan	32	128	11008	4096
ChatGLM-6B	32	128	4h, 16384	4096
ChatGLM2-6B	32	128	13696	4096
Bloom	32	128	4h, 16384	4096
Falcon	71	64	4h, 18176	4544

Falcon、PaLM、ChatGLM2-6B都使用了Multi-query Attention，但有细微差别。

- 为了保持参数量一致，
    - Falcon: 把隐藏维度从4096增大到了4544。多余的参数量分给了Attention块和FFN块
    - ChatGLM2: 把FFN中间维度从11008增大到了13696。多余的参数分给了FFN块

4.4 Multi-Query Attention 这样做的好处是什么?

减少 KV cache 的大小，减少显存占用，提升推理速度。

4.5 有 哪些模型 是 使用 Multi-Query Attention?

- 代表模型：PaLM、ChatGLM2、Falcon等

5 Grouped-query Attention

5.1 什么是 Grouped-query Attention?

Grouped query attention: 介于multi head和multi query之间，多个key和value。

5.2 有哪些大模型使用 Grouped-query Attention?

ChatGLM2，LLaMA2-34B/70B使用了Grouped query attention。

6 FlashAttention

- 核心：用分块softmax等价替代传统softmax
- 优点：节约HBM，高效利用SRAM，省显存，提速度
- 代表模型：Meta推出的开源大模型LLaMA，阿联酋推出的开源大模型Falcon都使用了Flash Attention来加速计算和节省显存
- 关键词：HBM、SRAM、分块Softmax、重计算、Kernel融合。

7 并行 transformer block

用并行公式替换了串行，提升了15%的训练速度。

在8B参数量规模，会有轻微的模型效果损失;在62B参数量规模，就不会损失模型效果。

Falcon、PaLM都使用了该技术来加速训练