[ASPLOS 2023] FLAT: An Optimized Dataflow forMitigating Attention Bottlenecks

王天宸

2024 年 1 月 23 日

1 Introduction and Background

Transformer 中的自注意力机制是其核心部分，其中一个重要属性是序列长度 N。直观上来说增加序列长度可以使基于注意力的模型更好地捕捉输入句子的上下文或图像片段之间的关系。

- 但是最先进的应用于注意力操作中的神经网络数据流会在序列长度达到 512 时就达到性能顶峰
- 注意力模型中内存需求为 O(N2) 其中 N 为序列长度
- 本文通过算子融合和控制片上活动内存的大小，使得序列长度 N 能达到最大 64k 元素，且内存需求为 O(N)

本文中的

2 Transformer 中的对应部分与挑战

对于注意力层中的输入张量 $[B, N, D]$ 中，B 为 Batch size，N 为序列长度 N，即序列中有 N 个 token；D 是一个 token **向量**的维度 (1*D)。

- 根据 $[B, N, D]$ 和权重张量相乘得到 K、Q、V 张量。K、Q、V 三个张量生成多个头得 H(head 个数) 使得 $[B, N, D]$ 变为 $[B, H, N, d]$ 其中 $d = N/H$。

- 计算 L(logits score)，将 Q **向量**和 K **向量**点积看，得到每个令牌与序列中其他令牌之间的相关性分数（logits)。输出张量为 $[B, H, N, N]$
- 用 softmax 归一化 L 的输出
- V 与上一步输出进行 matmul 得到是注意力输出张量 $[B, H, N, d]$
- 多头注意力层输出

3 问题

3.1 L/A 计算密度低

对于激活-权重算子 (Q/K/V/O) 来说操作密度是 $O(\frac{BND^2}{BND+D^2+BND})$ 分子中的 $O(BND^2)$ 是操作次数，分子中的三项分别是输入、权重、输出对应的访存次数。增加 B 的大小可以增加操作密度。

对于激活-激活算子 (L/A) 来说，操作次数为 $O(BN^2D)$，两个算子的输入分别为 $O(BND)$，输出为 $O(BN^2)$ 总的操作密度是 $O(\frac{BN^2D}{2BND+BN^2})$。对于现在采用的多头注意力来说是是 $O(\frac{BN^2D}{2BND+BHN^2})$。可见不能简单通过增加 B 的大小增加操作密度。

对三种常见的基于注意力的模型和广泛使用的 CNN 网络 ResNet50 的操作符的屋顶线分析，可见 CONV 一般有好的计算利用率，FC 可以通过增加 batchsize 来提高计算利用率，但 L/A 则严重被访存限制。L/A 操作符的低操作强度使得它们基本上是内存受限的，任何单个操作符级别的数据流/映射探索都不能进一步提高性能。

3.2 融合困难

对于计算密度低的算子来说算子融合是个不错的策略。由于 softmax 在元素层面上来说是一个多对多算子融合不会简单地自动融合。本文设计的数据流成功将这两个算子融合。

3.3 L/A 中间张量大小随 N 二次增长

L 的输出是和 N 二次相关的，将其中间结果保留在片上是不现实的，超过 16k 的序列长度时存储超过 8GB。故整个中间张量保存在片上不可扩展。

本文的 tiling 策略可以根据片上内存约束控制活动内存占用。

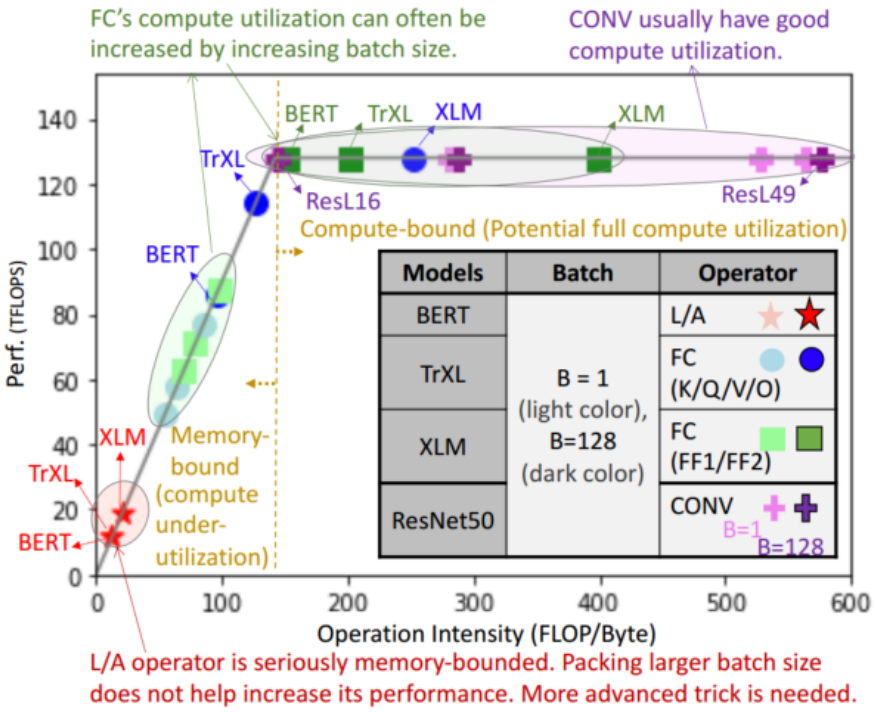

图 1: 对 Bert 等模型中算子的 roofline 测试

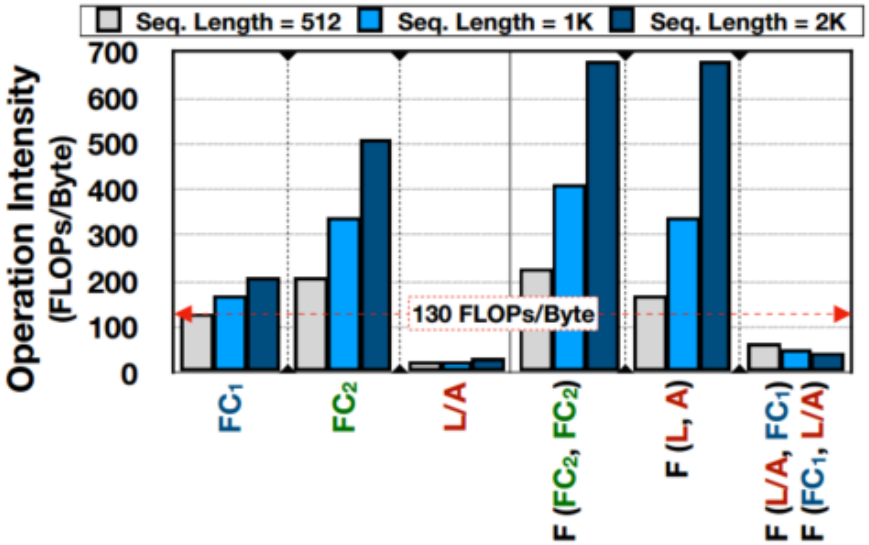

图 2: Fig 4

4 选择策略

根据 fig4 可见只有 L/A 算子没有达到构成计算瓶颈的操作密度，而融合后操作密度获得了提升。由于别的算子已经达到受计算约束的标准了，并且可以通过增加 B 的大小提高操作密度，且融合会导致潜在的权值复用减少的情况，故不融合

5 FLAT 的实现

要融合任意的两个张量算子，可以将算子内部的循环分为“外循环”和“内循环”。用 L 和 A 举例，其中外循环是 L 和 A 共享的循环，内循环由各个算子决定。在融合后将共用外循环而内循环依次执行。

对于 L-softmax-A 来说，最小粒度由 softmax 带来的数据依赖关系决定。由于 Softmax 的规约是沿着一个主维度进行的，故在这个维度上的向量构成了数据依赖关系的最小粒度。

外层循环的粒度受到片上存储的约束，其随着序列长度 N 的增加二次增长。对于其他问题上，减小外层循环粒度可能会导致权重复用的减少；但

Table 2: Buffer requirement for tiling granularity. M: batched Multi-head, B: Batch, H: Head, R: Row.

Granularity	M-Gran	B-Gran	H-Gran	R-Gran
Buffer Req.	$\mathcal{O}(8BDN+BHN^2)$	$\mathcal{O}(8DN+HN^2)$	$\mathcal{O}(8Nd+N^2)$	$\mathcal{O}(4Rd+4Nd+RN)$

图 3: 表 2

对于 L/A 融合问题来说其中所有算子都是激活-激活算子，每一次都要重新获得权值故不存在这个问题。

内层循环，也就是 softmax 带来的最小粒度上粒度大小的选择来说，减小粒度会减少矩阵乘中的数据复用，故也需要考虑其大小。

表 2 中列出了不同粒度下所需片上存储的计算结果。