Application Note

通过减少使用的内存和 DDR 带宽来优化 TI 深度学习性能

Adam Hua, Reese Grimsley

摘要

TIDL 是 TI 的 AI 推理框架，在 TDA4x 和 AM6xA 系列处理器上运行，利用内置的 C7xMMA AI 加速器实现高效的 AI 模型推理。C7xMMA 作为专用的 AI 推理加速器，具有复杂的架构。虽然 TIDL 推理框架已广泛优化资源分配，从而更大限度地提高利用率，但在模型推理期间仍可能会出现较高的内存带宽消耗。为了进一步利用推理资源并减少内存用量，TIDL 上运行的模型需要进行额外的优化。本文档详细介绍了旨在减少 DDR 带宽消耗的模型优化方法。

内容

商标

1 简介

目前能够进行 AI 模型推理的 SoC 通常采用以下两种架构之一：集成通用 GPU 的架构和包含专用 AI 推理加速器（通常称为 NPU）的架构。TI TDA4x 和 AM6xA 产品系列中的 AI 加速型 SoC 采用后一种方法，其 NPU 通常称为 C7xMMA。该名称来源于 NPU 的两个组成部分：C7000 系列浮点数字信号处理器和矩阵乘法加速器 (MMA)。C7x 系列 DSP 内核会运行 RTOS，在模型内负责数据调度和非线性处理。MMA 与 C7x 深度耦合，负责线性代数运算（如矩阵乘法和 2D 卷积），这些运算占大多数神经网络计算要求的 99% 以上。

TI 提供了 TI 深度学习 (TIDL) 推理框架，TIDL 架提供了统一的接口，便于高级操作系统（*例如* Linux、QNX）轻松调用。具体调用方法不在本文讨论范围内，我们默认读者已熟悉相关接口，我们将重点关注模型优化技术。用户可以利用 TIDL 工具为特定处理器编译模型。然后，TIDL 将量化编译模型部署到 NPU 上，允许用户使用 TIDL 运行时 (TIDL-RT)、tivxTIDLNode 或开源运行时框架 (OSRT)（如 ONNX Runtime、Tensorflow-Lite）来调用推理。

TIDL 内存读取/写入带宽是指 DDR 接口上的负载。例如，如果单个推理帧需要从 DDR 读取 100MB（包括模型权重、输入和中间层特征映射）并向 DDR 写入 50MB（包括模型输出和中间特征映射），则实现 30fps 帧率需要 4.5GB/s 的总 DDR 读取/写入带宽。由于单通道 DDR4 可能提供大约 8GB/s 的实际带宽，TIDL 模型推理会消耗大量带宽。

AM67A 和 TDA4VH 等器件包括多个 C7x 实例。尽管 TDA4VH 等处理器将包括多个 DDR 接口，但在并行加速器上运行的并行推理任务将进一步影响 DDR 的利用率和争用情况。查看系统级性能时，应考虑系统级 DDR 争用情况，但本文中的优化仍然有利于首先降低 DDR 利用率，以提高模型性能。

为了优化 DDR 带宽，需要对 C7xMMA 高速缓存结构有一定的了解，从而有效地利用 TIDL 工具，并可能需要对模型进行修改。

2 C7xMMA 高速缓存结构

DDR 带宽优化要从了解 TI C7xMMA 的内存层次结构开始，如下图简化版结构所示。

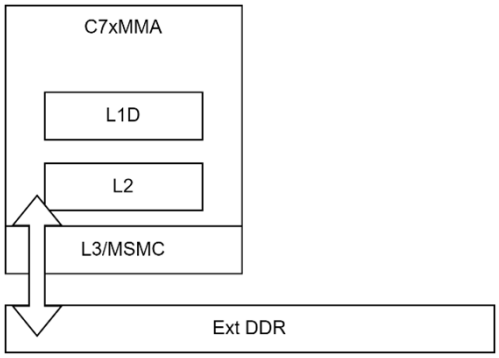

图 2-1. C7xMMA 三级高速缓存结构

C7xMMA 采用三级高速缓存结构。除外部 DDR 之外，它还整合了内部 L1D、L2 和 L3/MSMC 高速缓存。L1D 最小，最接近计算内核（典型大小为 16KB）。L2 相对远一点（典型大小为 224KB、448KB），但与 MMA 的数据移动机制紧密耦合。TDA4x 上的 L3 是多核共享内存控制器 (MSMC)，而在其他 SoC 上，它是由每个 C7xMMA 单独管理的 SRAM。注意：此处的 L1D、L2、L3 术语对应于 TIDL 框架中的说明；芯片数据表中可能称为 L1P、L1D、L2；在某些 SoC 中，也可能指 L3（即 TDA4VM 上的 MSMC）。L2 和 L3 区域的大小可在 tidl 工具包含的 device_config.cfg 文件中找到

下图显示了典型层的推理过程中（涉及四个操作）缓存的使用情况。操作 1 是 DMA 将数据直接从 DDR 传输到 L2。操作 2 将数据从 L3 移动到 L2。操作 3 将数据从 L2 传输到 L3。操作 4 将数据从 L3 移动到 DDR。操作 2 和 3 的效率比操作 1 和 4 高十倍以上。利用前一层的特征映射可能导致三种情况：只有操作 1（如果输入层和前一层输出完全位于 DDR 中）；只有操作 2（如果前一个特征映射完全适合 L3/MSMC）；或操作 1 和 2（如果前一个输出对于 L3 太大，部分存储在 DDR 中）。计算出当前层的特征映射后，操作 3 会优先将数据移动到 L3。如果超过 L3 容量，操作 4 会将多余的数据存储在 DDR 中。权重值始终存储在 DDR 中，并在需要时直接提取到 L2。

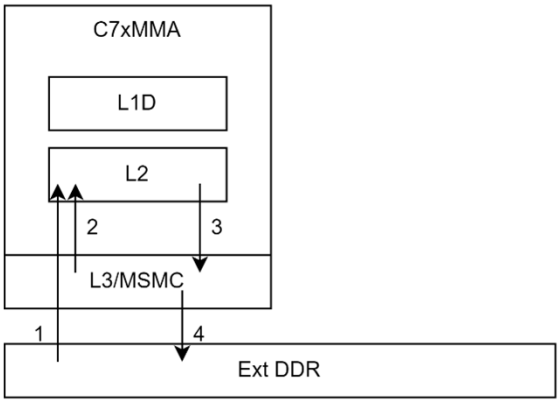

图 2-2. C7xMMA 高速缓存操作

这种三级高速缓存架构避免了计算周期内缓慢的 DDR 读取/写入操作并节省了 DDR 带宽，从而显著提高了推理效率。提高效率和节省带宽的关键在于最大限度地提高 L3 利用率，从而防止在 DDR 中存储特征映射。下一节将介绍如何分析模型的内存使用情况。

3 为编译的 TIDL 模型进行 DDR 读取/写入分析建模

TI 模型编译工具提供了相应的接口，有助于快速进行 DDR 读取/写入分析。编译模型时，将在输出文件夹中生成 bufinfolog_0.csv 文件。确切地说，它将位于一个命名为 artifacts/tempDir/$SUBGRAPH_NAME_tidl_net 文件夹中，其中 SUBGRAPH_NAME 取决于模型。在不同的 SDK 版本中，文件夹结构的确切名称可能会略有变化。

需要注意的是，使用 RTOS（即 tidl_model_import.out）工具时，必须确保提供 perfSimConfig 并且该工具正常运行。在模型编译失败的任何情况下，必须在分析带宽消耗之前解决错误。编译过程中的任何错误日志都可能导致不准确的分析结果。

下表说明了 bufinfolog CSV 文件中的关键字段。

表 3-1. *bufinfolog* CSV 文件中的字段说明

字段	说明
Ni	输入特征映射通道尺寸。
No	输出特征映射通道尺寸。
InW	输入特征映射宽度尺寸。
InH	输入特征映射高度尺寸。
OutW	输出特征映射宽度尺寸。
OutH	输出特征映射高度尺寸。
In-Write-size	实际输入特征映射大小，以字节为单位计算。可能涉及少量用于填充操作或 DMA 总线大小调整的额外开销。
In-Write-memSpace	L2 或 DDR；除数据层（输入/输出层）外，所有层都使用 L2
Out-Read-memSpace	当前层的计算结果的临时存储位置。值通常为空、L2、MSMC 或 DDR。空表示计算出的结果直接存储到 Out-Write-memSpace 中。当存在某个值时，表示需要首先将当前层输出特征映射的一部分存储在 Out-Read-memSpace 中，然后从该空间读取并写入到 Out-Write-memSpace。
Out-Write-memSpace	当前层的特征映射的最终存储位置。值为 DDR 或 MSMC。DDR 表示当前层的输出特征映射无法完全放入 MSMC 中，因此部分或全部存储在 DDR 中。MSMC 表示输出完全存储在 MSMC 中，供后续层使用。
Out-Write-size	当前层存储在 Out-Write-memSpace 中的数据量。当 Out-Write-memSpace 为 DDR 时，该值表示当前层对 DDR 写入带宽消耗的影响。可能涉及少量用于填充操作或 DMA 总线大小调整的额外开销。
Wt-Write-memSpace	当前层的权重的加载位置。始终为 L2。权重值完全存储在 DDR 中，并在使用时加载到 L2 中。
Wt-Write-size	当前层的权重数据的大小。该值会影响 DDR 读取带宽的消耗。可能涉及用于填充操作或 DMA 总线大小调整的额外开销。

在上述字段中，DDR 写入带宽主要受 **Out-Write-memSpace** 值的影响。通过减小一层的输出特征映射大小，可以显著降低 DDR 写入带宽消耗。

DDR 读取带宽主要受 **Wt-Write-size** 和 **In-Write-size** 的影响。值得注意的是，仅当前一层的输出部分或完全存储在 DDR 中时，**In-Write-size** 才会产生影响。

因此，减少 DDR 读取/写入带宽的关键在于优化模型，以尽可能减少权重导致的读取带宽以及大型特征映射产生的读取/写入带宽。

4 模型优化

根据上述情况，降低 DDR 带宽依赖于在 L3 中保留中间特征映射，因此需要特意设计模型。

4.1 简单结构模型

简单模型具有线性、非分支结构。下面所示的 EfficientNet 的初始部分是完全按顺序排列的。在这里，确保每个层的输出适合 L3 就足够了。TIDL 可以自动配置层以使用 L3 运行，从而避免 DDR 交互。当中间特征映射超过 L3 的大小时，仍需要 DDR 交互。

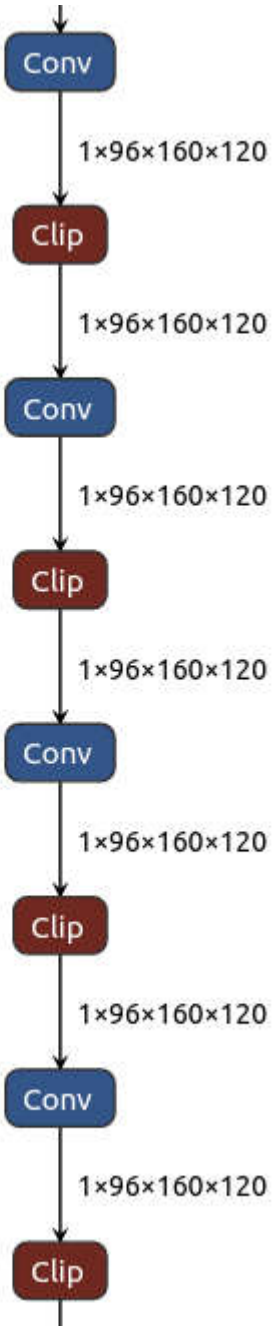

图 4-1. 具有顺序层排序的简单模型

4.2 复杂结构

在实践中，许多模型的结构和图模式比单纯的顺序层更为复杂。接下来的几节将介绍一些更复杂的结构示例，并说明其对 DDR 和高速缓存使用情况的具体影响。

4.2.1 残留结构

许多骨干架构使用如图所示的残留结构，创建称为“残差”的本地化并行路径，在训练过程中非常有用。残差可避免梯度消失问题。

在编译期间，TIDL 会针对不同的计算顺序（左分支优先、右分支优先、交错）模拟 DDR 带宽，并选择其中最高效的。它还会决定第一个 Conv 层的输出是保持在 L3/MSMC 中直到出现添加操作，还是立即写入 DDR。存储在 DDR 中会导致直接带宽成本，而保留在 L3 中可能会在左分支计算期间占用内存，从而可能强制左分支的部分使用 DDR。

TIDL 将选择一种能够更大限度地增加 L3 占用的策略，但大型中间特征映射可能需要使用 DDR。在这种情况下，建议优先优化较长路径（图中的左侧）的大小，以避免多个特征映射进入 DDR，而跳跃连接仅涉及单个特征映射。

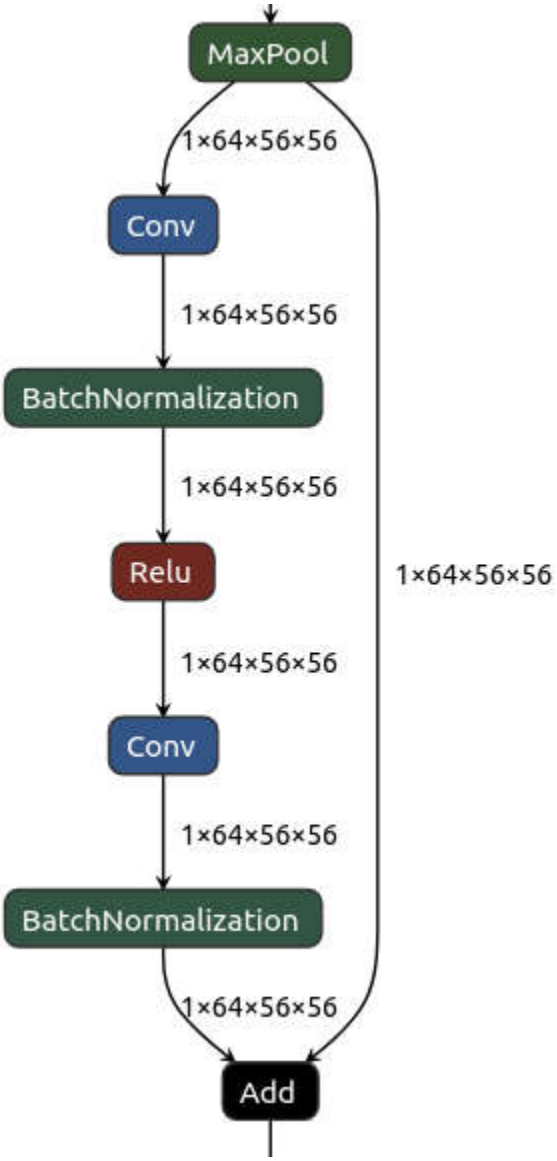

图 4-2. 神经网络中的残留结构。右侧路径的“跳跃”连接具必须存储特征映射，直到左侧路径完成

4.2.2 并行分支合并

应用通常涉及将多个深度并行分支合并到一个分支中，或将一个分支拆分为多个深度并行路径。这对于多输入神经网络尤其常见。图中展示了经典四输入 BEV 网络在 gridsample 算子合并路径后的部分结构。

由于路径在合并之前比较深，因此必须将特征映射放置在 DDR 中的合并点，而这不可避免地需要 DDR 带宽消耗。此类架构应仅在需要时使用，以避免超出 DDR 带宽并由此产生瓶颈。但是，可以降低由权重引起的 DDR 读取带宽。可以修改模型架构以合并多个输入头，并将某些模型层的批次维度设置为大于 1 的值，这样权重只需加载一次。

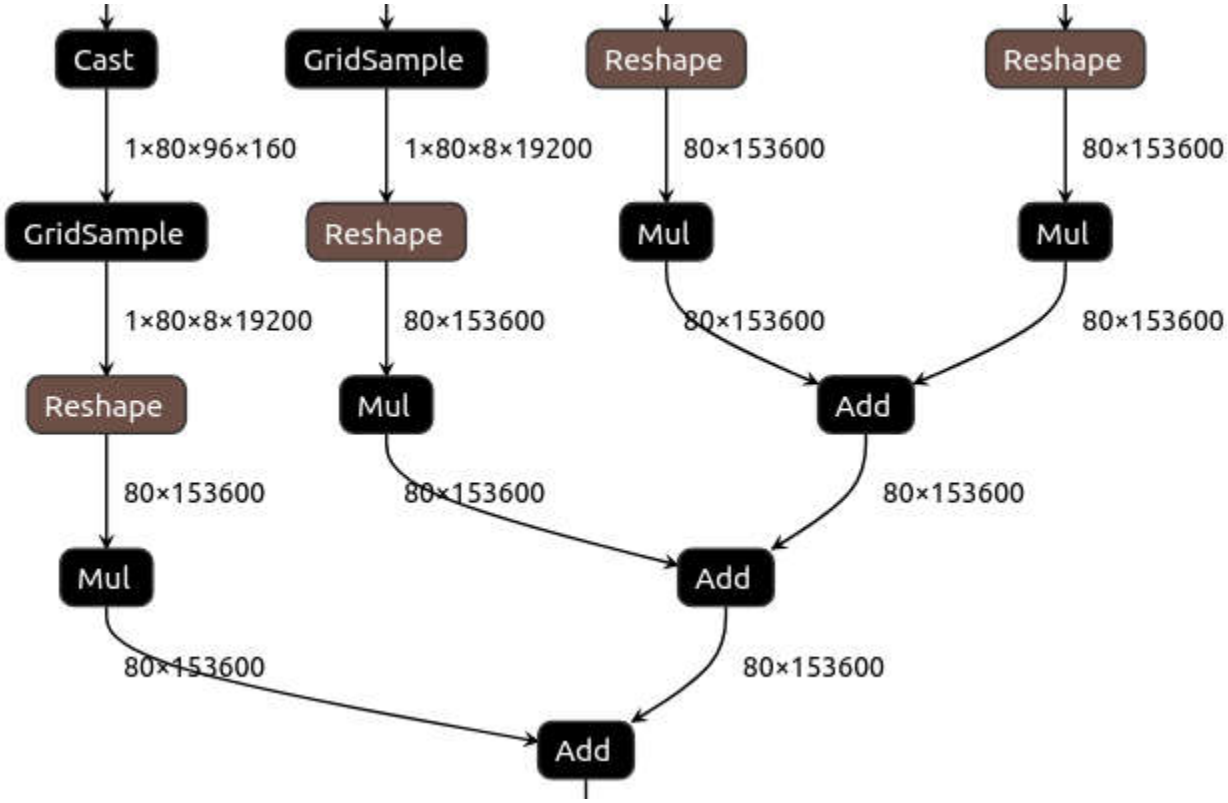

图 4-3. 合并多个并行分支的复杂结构

例如，在上图中，相同的骨干架构在 GridSample 层之前，此骨干架构每一层的特征映射相对较小。四个分支可以合并为两个甚至一个分支，并相应调整批次大小。然后是相应的层或数据整形层，以再次分离批次，从而可以通过所示的添加层重新组合这些批次。这种方法可以减少甚至防止相同权重的重复加载，从而降低 DDR 读取带宽开销。这种方法需要注意合并的骨干架构中的特征映射的大小。

5 总结

DDR 带宽的模型优化主要涉及减小每层特征映射大小和增加深度。对于复杂的结构，DDR 带宽消耗可能无法避免。TI 的 Model Zoo 提供了许多经过优化和验证的模型和骨干架构。考虑到通用架构的成熟度，可以考虑使用 TI 优化的版本来替换模型的骨干架构，以实现快速改进。

本文档详细介绍了分析模型 DDR 带宽消耗并优化模型以降低带宽消耗的方法。这些内容与 TDA4x、AM6xA 系列 SoC 和 TIDL 推理框架的用户密切相关。应用这些方法通常可获得经过优化的模型，这些模型仅在输入和输出时占用带宽，从而为整个系统释放大量资源。

6 参考资料

1. https://www.ti.com.cn/product/cn/AM62A7
2. https://www.ti.com.cn/tool/cn/PROCESSOR-SDK-AM62A
3. https://www.ti.com.cn/product/cn/AM67A
4. https://www.ti.com.cn/product/cn/TDA4VM
5. https://www.ti.com.cn/product/cn/TDA4VE-Q1
6. https://www.ti.com.cn/product/cn/TDA4VM-Q1
7. https://www.ti.com.cn/product/cn/TDA4AL-Q1
8. https://www.ti.com.cn/product/cn/TDA4VL-Q1
9. https://www.ti.com.cn/product/cn/TDA4VP-Q1
10. https://www.ti.com.cn/product/cn/TDA4VH-Q1
11. https://www.ti.com.cn/product/cn/TDA4VEN-Q1

重要通知和免责声明

TI“按原样”提供技术和可靠性数据（包括数据表）、设计资源（包括参考设计）、应用或其他设计建议、网络工具、安全信息和其他资源，不保证没有瑕疵且不做出任何明示或暗示的担保，包括但不限于对适销性、与某特定用途的适用性或不侵犯任何第三方知识产权的暗示担保。

这些资源可供使用 TI 产品进行设计的熟练开发人员使用。您将自行承担以下全部责任：(1) 针对您的应用选择合适的 TI 产品，(2) 设计、验证并测试您的应用，(3) 确保您的应用满足相应标准以及任何其他安全、安保法规或其他要求。

这些资源如有变更，恕不另行通知。TI 授权您仅可将这些资源用于研发本资源所述的 TI 产品的相关应用。严禁以其他方式对这些资源进行复制或展示。您无权使用任何其他 TI 知识产权或任何第三方知识产权。对于因您对这些资源的使用而对 TI 及其代表造成的任何索赔、损害、成本、损失和债务，您将全额赔偿，TI 对此概不负责。

TI 提供的产品受 TI 销售条款)、TI 通用质量指南 或 ti.com 上其他适用条款或 TI 产品随附的其他适用条款的约束。TI 提供这些资源并不会扩展或以其他方式更改 TI 针对 TI 产品发布的适用的担保或担保免责声明。 除非德州仪器 (TI) 明确将某产品指定为定制产品或客户特定产品，否则其产品均为按确定价格收入目录的标准通用器件。

TI 反对并拒绝您可能提出的任何其他或不同的条款。

最后更新日期：2025 年 10 月