Towards Agile DNN Accelerator Design Using Incremental Synthesis on FPGAs

Source: Architecture and CAD, FPGA '22

Last Modified: 29/12/2022

0. Abstract

增量编译框架名为 Acoda，首先利用一种"图编辑距离算法"检测软件上的修改，然后通过"多层次的复用"将软件上的修改映射到硬件上的修改。

1. Introduction

神经网络的算法硬件协同设计可以提高应用时的表现，通过将各层利用定制化的模块实现。通常的设计流程中包括算法和硬件的流式迭代，但是目前硬件的综合部署需要在每次算法修改后都整体重新完成，消耗了大量的时间。

在协同设计流程中，对网络结构的调整一般是局部和小的且每次只改变一个部分。因此可以进行增量综合，即只重新综合变化的部分。

Acoda 通过计算图编辑距离来检测网络的修改，并通过三层级的复用来尽可能复用硬件设计和增量编译。三层级包括子图复用，节点复用和部分复用。

2. Background and motivation

2.1 DNNs on FPGAs

神经网络通常由多个常用的模式或结构组成。许多新型的神经网络实际是之前神经网络的变种。在协同设计过程中，可以根据目标器件的算力来选择适合的变种以定制神经网络。

2.2 DNN accelerator Design

神经网络加速器通常包括多种硬件模块来实现不同的运算，这些模块可以使用各种 IP 来进行实现。但目前的设计中，框架在生成硬件时不存在任何复用，即任何小的修改都需要整体重新综合部署来生成新的硬件。而通过指令集的可软件编程的加速器设计又通常在具有更高灵活性的同时损失了性能和能效。

3. Acoda

用户输入为神经网络算法，输出为针对该算法的最终加速器设计。Acoda 需要一个参考设计来完成当前设计。首先，将神经网络变成一个计算图，其中节点表示层，边表示会数据依赖。基于这张图，每个节点都被映射为单独的硬件模块。当硬件资源不足以实现复杂神经网络时，会将整张图划分为子图，并对每个子图生成硬件架构。在每个架构中，不同的模块被放在器件的固定位置上。Acoda 会存储所有的中间逻辑网表和物理实现以应对之后的使用。

基于之前的参考运行，Acoda 增量地实现当前部署。Acoda 检测出当前和参考运行之间的修改并通过层级式的复用来实现硬件上的修改。

（1） 当子图没有发生改变时，Acoda 不进行修改直接复用；

（2） 当子图中的单一节点发生变化时，Acoda 复用没有发生改变的节点对应的硬件模块，后端通过参数上的探索实现被修改节点的模块设计；

（3） 如果多个节点发生改变以至于需要整个架构重新部署，Acoda 部分地复用已有的模块设计。

4. DNN Revisions

4.1 Revision patterns

通过分析已有的神经网络优化方法，将网络上的修改识别为七个模式：

（1） 插入：将一个新节点插入图中；

（2） 去除：将一个已有节点删除；

（3） 替换：将一个节点替换为不同种的另一个节点；

（4） 压缩：将一个节点的参数进行改变而变成另一个节点；

（5） 重排：将两个节点交换顺序；

（6） 解压：将一个节点变成一系列节点：

（7） 复制：插入一个子图的复制。

4.2 Revision detection algorithm

Acoda 通过计算近似图编辑距离来找到修改序列，其中距离对应着图编辑操作的消耗。消耗由所需综合时间决定，而综合时间与修改节点的资源占用成正比，即与节点计算复杂度成正比。计算复杂度定义如下：

（1） 插入：加权后的对应节点计算复杂度；

（2） 去除：加权后的对应节点计算复杂度；

（3） 解压：加权后的对应节点前后计算复杂度的较大值；

（4） 替换：加权后的对应节点前后计算复杂度的和；

（5） 压缩：特征向量的欧式距离；

（6） 重排：某个常数；

（7） 复制：由后处理决定，任何相同的子图均被视为复制。

通过近似编辑距离算法来寻找最小值具有 O(N3)或 O(N2)复杂度。

5. Reuse hierarchy

5.1 Underlying Architecture

Acoda 针对的场景为对单一神经网络设计专用加速器且每个网络节点均使用专用硬件模块实现。这样的专用化设计既有利于取得较好的性能指标，也利于硬件资源的复用，即减少需要重新综合的模块数量。

5.2 Methodology

由于不同种修改设计的节点数量不同，这些修改对于硬件的影响也不同。三不同层级的复用策略如下：

（1） 子图复用：复用一个子图的整个物理部署；

（2） 节点复用：复用节点的物理实现不进行重新物理部署；

（3） 部分复用：复用网表，但需要重新进行物理部署。

三种复用针对的修改类型如下：

（1） 子图复用：复制；

（2） 节点复用：重排，替换和压缩；

（3） 部分复用：插入，删除和解压。

下面具体介绍不同层级复用：

（1） 子图复用：完全复制原有物理设计并删除多余的部分；

（2） 节点复用：将原先用于某一节点的硬件资源给新节点使用；

（3） 部分复用：当重分配硬件资源不能满足新节点需求时，通过降低原先节点的并行度等方式整体重新综合已完成新网络的部署。