SoCaffe: 基于 Zynq SoC 平台的高性能深度学习框架

信息科学技术学院
赵睿哲
1200012778

1 摘要

近年来，深度学习与神经网络领域飞速发展，并逐渐应用于无人机、自动驾驶汽车等平台。这类平台使用嵌入式设备进行计算，对功耗、算法的实时性、计算资源都有一定的限制，与传统的深度学习应用所依赖的 GPU 或者集群在性能上相差甚远。因此直接移植深度学习应用到嵌入式设备有很大的难度。

本研究实现了 SoCaffe—一个基于 Caffe 与和 Zynq SoC 的深度学习框架。Caffe 作为最流行的 CPU/GPU 深度学习框架之一，效率高、配置简单，并为大部分研究人员所熟悉。Zynq SoC 是 Xilinx 公司推出的“全可编程”嵌入式开发平台，搭载高效的 ARM Cortex-A9 双核 CPU 与 Xilinx FPGA，能同时满足计算性能与功耗比的要求。

本研究综合考虑了 Zynq SoC 平台的特点与 Caffe 的计算特性，对 Caffe 的功能进行软硬件逻辑划分，挑选 Caffe 中密集使用的 GEMM 计算作为 FPGA 加速的目标。针对 GEMM 的硬件逻辑设计，本研究对其性能进行了细致的分析和数学建模，从硬件资源占用与延迟两个角度进行了充分的优化，达到了相对于软件版本 5.4x 加速比和最高 12.31GFLOPS 的性能。同时，本研究生成的 SoCaffe 框架的功能基本与 Caffe 完全一致，基于其他硬件平台的 Caffe 应用可以直接移植到 Zynq SoC 上实现。最后，SoCaffe 的整体计算性能也有最高 2.34x 的加速比。此外，本研究完全基于 Xilinx 新推出的 SDSoC 开发平台进行软硬件协同设计，大大提升了开发效率。

综上，SoCaffe 同时拥有 Zynq SoC 平台的高性能与 Caffe 框架的通用性和易用性，是针对嵌入式平台深度学习应用开发的实际解决方案，具有一定的实用价值。

2 主要工作

本研究的主要工作如下：

1. 使用 Vivado HLS 工具实现了 GEMM 计算在 FPGA 上的设计与实现，针对 Zynq-7000 的系统资源特点给出了相应的优化方案与数学分析模型，找到了能实现的矩

阵块大小的上界；同时使用了半精度浮点数优化了对计算资源的使用。最终获得最高 12.31GFLOPS 和 5.42 倍加速比。

2. 系统地学习和掌握了 Xilinx SDSoC 开发环境的使用方式，通过文档与探索自主掌握了该工具的高级使用方法，并针对 SoCaffe 的特点定制了一套开发流程。

3. 使用 ARM GNU 工具链使用交叉编译的方式编译了全部 Caffe 需要使用的第三方库，使用 SDSoC 构建了 SoCaffe 的系统镜像文件包和动态链接库。

4. 测试了 SoCaffe 的计算性能，可用性，精度等等特性，找到了 SoCaffe 的最佳使用条件：即针对大量使用卷积神经网络，网络规模较大的时候，SoCaffe 可以取得不错的加速比。

3 未来工作

SoCaffe 存在一些缺点和亟待提高的地方：1) 首先，因为 Zynq SoC 平台的内存有限，SoCaffe 无法支持特别大的神经网络结构；2) 其次，SoCaffe 的 GEMM 固定块算法依然有一定的比重是冗余计算，如果输入矩阵形状比较小，则会造成较大比例的性能损失；3) 最后，SoCaffe 目前只支持 GEMM 的优化，其他不用到 GEMM 操作的网络层不能得到相应的性能提升。这些工作都是本研究下一步要进行优化的工作。

4 测试数据

		Baseline	Latency	ResAlloc	IrrShape	HalfFloat
矩阵形状	M	32	32	56	64	96
	N	32	32	56	64	96
	K	32	32	56	56	96
FPGA 资源占用(%)	BRAM18K	35.71	35.71	63.57	71.79	92.5
	DSP48E	72.73	72.73	96.36	96.36	43.64
	FF	27.43	27.34	40.27	40.89	38.51
	LUTs	34.62	34.57	56.59	59.3	55.73
延迟 (cycles)		4404	2352	6736	8673	19030
时钟频率 (MHz)		143	143	143	143	143
GFLOPS预测		2.227735	4.1713197	7.65610451	7.76648449	13.50436994
GFLOPS块测试		2.558022	3.31	6.77	7.213274	12.31
GFLOPS完整测试		1.272	1.45	2.61	3.039285	4.44

Figure 1: GEMM 的多种优化策略对比

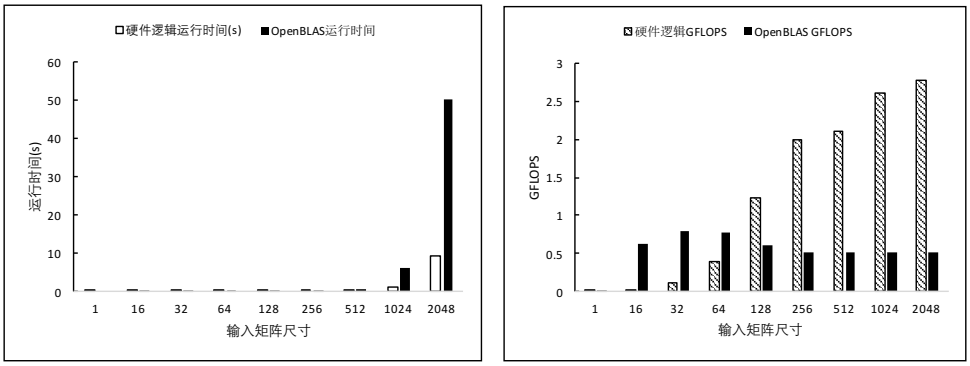

Figure 2: GEMM 硬件加速与 OpenBLAS 版本的对比

	OpenBLAS	Latency	ResAlloc	IrrShape	HalfFloat
forward(ms)	1713.13	1264.9	830.649	999.284	787.999
backward(ms)	3088.98	2489.66	1676.27	1636.56	1222.61
total(ms)	4802.7	3754.9	2507.2	2636.1	2010.8
加速比	1	1.28	1.92	1.82	2.39

Figure 3: SoCaffe 卷积层速度测试

		IrrShape	Half	Latency	ResAlloc	OpenBLAS
MNIST		Time (ms)	Time (ms)	Time (ms)	Time (ms)	Time (ms)
avg fw		567.105	588.082	456.306	463.133	577.453
avg bw		893.551	925.128	700.1	721.521	892.443
avg fw-bw		1460.9	1513.46	1156.66	1184.92	1470.14
mnist	forward:	0.6379	0.74858	0.65728	0.66402	0.73992
mnist	backward:	0.00382	0.00368	0.0041	0.00426	0.00396
conv1	forward:	82.1427	83.5672	70.0164	71.4867	84.2427
conv1	backward:	71.8496	71.3425	60.4915	60.5835	71.5036
pool1	forward:	39.2125	39.1425	39.3205	39.2049	39.1716
pool1	backward:	19.4461	19.1329	18.7064	18.8129	19.1824
conv2	forward:	328.47	352.57	249.294	254.82	344.042
conv2	backward:	622.903	649.87	483.504	501.371	627.029
pool2	forward:	16.4577	16.4013	16.4905	16.5062	16.5277
pool2	backward:	9.93022	9.83388	9.58154	9.55594	9.88252
ip1	forward:	97.5931	93.0654	78.1103	78.0357	90.2199
ip1	backward:	165.487	170.657	124.497	127.782	160.885
relu1	forward:	0.87482	0.87578	0.85868	0.85458	0.85834
relu1	backward:	1.4372	1.43284	1.35666	1.41778	1.42318
ip2	forward:	1.1313	1.14832	0.9504	0.95464	1.05204
ip2	backward:	2.41664	2.78364	1.86154	1.89474	2.45312
loss	forward:	0.50968	0.5038	0.5171	0.51792	0.52312
loss	backward:	0.02338	0.02402	0.02384	0.024	0.02416
Loss		0.0418068	0.0418068	0.0418068	0.041807	0.0418068
accuracy		0.9862	0.9862	0.9862	0.9862	0.9862

Figure 4: SoCaffe MNIST 速度测试