一种基于 FPGA 的卷积神经网络加速器设计与实现

仇越 1，马文涛 2，柴志雷 1，2
（1 江南大学 物联网工程学院，江苏 无锡 214122；
2 数学工程与先进计算国家重点实验室，江苏 无锡 214125）

摘要： 针对卷积神经网络模型 ZynqNet 现有 FPGA 实现版本中卷积运算单元并行度低，存储结构过度依赖片外存储等问题，提出一种针对 ZynqNet 的 FPGA 优化设计.设计了双缓冲结构将中间运算结果放到片内以减少片外存储访问；将数据位宽从 32 位降为 16 位；设计了具有 64 个卷积运算单元的并行结构.实验结果表明，在 ImageNet 测试准确度相同的情况下，本文所提出的设计工作频率可达 200 MHz，运算速率峰值达到 1.85 GMAC/s，是原 ZynqNet 实现的 10 倍，相比 i5-5200U CPU 可实现 20 倍加速.同时，其计算能效达到了 NVIDIA GTX 970GPU 的 5.4 倍.

关键词： 卷积神经网络；现场可编程门阵列（FPGA）；ZynqNet；并行计算；加速
中图分类号： TP39
文献标识码： A
文章编号： 1000-7180(2018)08-0068-05

Design and Implementation of a Convolutional Neural Network Accelerator Based on FPGA

QIU Yue1, MA Wen-tao2, CHAI Zhi-lei1,2
(1 School of Internet of Things, Jiangnan University, Wuxi 214122, China;
2 State Key Laboratory of Mathematical Engineering and Advanced Computing, Wuxi 214125, China)

Abstract： In the hardware design of ZynqNet implemented on FPGA, the parallelism of convolution unit is low and the storage structure is almost dependent on off-chip memory. A FPGA accelerator optimization is proposed based on ZynqNet and it is easy to apply in other CNN models. The double buffering stores intermediate result of the network into the chip to reduce off-chip access; The data precision is changed from 32 bits to 16 bits, thus a parallel structure of 64 convolution operation units is designed to improve computing parallelism. The ImageNet results show that the optimized accelerator based on FPGA can achieve peak performance of 1.85 GMAC/s under 200 MHz, it is 10 times speedup compared to the original ZynqNet and 20 times speedup compared to i5-5200U CPU. In terms of performance power ratio, the FPGA accelerator is 5.4 times of NVIDIA GTX 970GPU version.

Key words： convolutional Neutral Network (CNN) ; field-programmable gate Array(FPGA); ZynqNet; parallelism computing; acceleration

作者简介：
仇越女，(1992-)，硕士研究生.研究方向为可重构计算.E-mail：942979548@qq.com.
马文涛男，(1981-)，工程师.研究方向为分布式计算.
柴志雷男，(1975-)，博士，副教授.研究方向为嵌入式系统设计技术、FPGA 的可重构计算等.