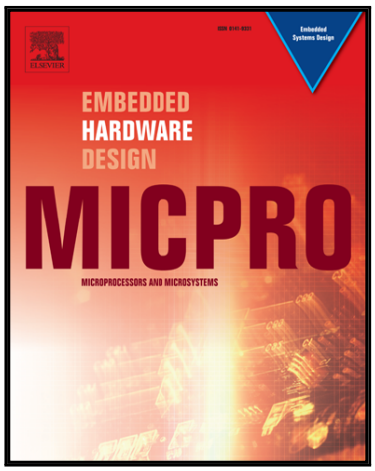
Accepted Manuscript



Throughput Optimizations for FPGA-based Deep Neural Network Inference

Thorbjörn Posewsky, Daniel Ziener

PII: S0141-9331(17)30296-X
DOI: 10.1016/j.micpro.2018.04.004
Reference: MICPRO 2675

To appear in: *Microprocessors and Microsystems*

Received date: 2 June 2017
Accepted date: 12 April 2018

Please cite this article as: Thorbjörn Posewsky, Daniel Ziener, Throughput Optimizations for FPGA-based Deep Neural Network Inference, *Microprocessors and Microsystems* (2018), doi: 10.1016/j.micpro.2018.04.004

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Throughput Optimizations for FPGA-based Deep Neural Network Inference

Thorbjörn Posewsky

Institute of Embedded Systems
Hamburg University of Technology (TUHH)
21073 Hamburg, Germany

Daniel Ziener

Computer Architectures for Embedded Systems
University of Twente
7500 AE Enschede, The Netherlands
Email: d.m.ziener@utwente.nl

Abstract

Deep neural networks are an extremely successful and widely used technique for various pattern recognition and machine learning tasks. Due to power and resource constraints, these computationally intensive networks are difficult to implement in embedded systems. Yet, the number of applications that can benefit from the mentioned possibilities is rapidly rising. In this paper, we propose novel architectures for the inference of previously learned and arbitrary deep neural networks on FPGA-based SoCs that are able to overcome these limitations. Our key contributions include the reuse of previously transferred weight matrices across multiple input samples, which we refer to as batch processing, and the usage of compressed weight matrices, also known as pruning. An extensive evaluation of these optimizations is presented. Both techniques allow a significant mitigation of data transfers and speed-up the network inference by one order of magnitude. At the same time, we surpass the data throughput of fully-featured x86-based systems while only using a fraction of their energy consumption.

Keywords: Deep Neural Networks, Batch processing, Pruning, Compression, FPGA, Inference, Throughput Optimizations, fully-connected

ACCEPTED MANUSCRIPT

Download English Version:

https://daneshyari.com/en/article/6885860

Download Persian Version:

https://daneshyari.com/article/6885860

Daneshyari.com