Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation

Doctoral Program in Electrical, Electronics and Communications Engineering

(35^{th}cycle)

Techniques and Optimization Strategies for Efficient Hardware Acceleration of Neural Networks

Tap-Wisely-Quantized Winograd Algorithm and Capsule Networks

By

Beatrice Bussolino

******

Supervisor(s):

Prof. Maurizio Martina, Supervisor

Doctoral Examination Committee:

Prof. Francesca Palumbo. , Referee, Università degli Studi di Sassari

Prof. Stefania Perri, Referee, Università della Calabria

Dr. Renzo Andri, Computing Systems Lab, Huawei Zurich Research Center

Prof. Guido Masera, Politecnico di Torino

Dr. Riccardo Peloso, ST Microelectronics

Politecnico di Torino

2023

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Beatrice Bussolino
2023

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Techniques and Optimization Strategies for Efficient Hardware Acceleration of Neural Networks

Beatrice Bussolino

The growing popularity of deep neural networks (DNNs) has intensified the demand for efficient hardware accelerators to handle the complex computations required by these models. This trend has led to increased research and development of domain-specific hardware accelerators (DSAs) to achieve the high performance and energy efficiency needed to support the deployment of DNNs in a wide range of applications. The effective execution of a DNN on a hardware accelerator depends on the workload presented by the model, the peak performance offered by the accelerator, and the efficiency with which the accelerator's resources are used.

With these three pillars in mind, in the first part of this thesis, we present a technique to enable convolutional neural networks (CNNs) acceleration by combining the Winograd algorithm for fast convolution and integer-only inference. A novel tap-wise quantization method overcomes the numerical issues arising when combining int-8 quantization and Winograd algorithm with larger tiles. Furthermore, custom hardware units and a carefully-tailored dataflow allow the processing of the Winograd transformations in a power- and area-efficient way. An extensive experimental evaluation on a large set of state-of-the-art computer vision benchmarks is conducted. This reveals that applying the tap-wisely-quantized Winograd algorithm with 4×4 tiles leads to a negligible accuracy loss compared to FP32 baselines. A domain-specific accelerator (DSA), enhanced with the Winograd custom hardware units, achieves up to $1.85\times$ gain in energy efficiency and up to $1.83\times$ end-to-end speed-up for state-of-the-art segmentation and detection networks.

In the second part of this thesis, we present our efforts to enhance the hardware-friendliness of capsule networks, a novel DNN model, by utilizing quantization methods and optimizing the model architecture in an HW-oriented fashion. This work aims to facilitate the acceleration of capsule networks, improving their computational efficiency and, in turn, enabling their deployment in resource-limited environments. First, we present a study on the quantization possibilities for capsule networks and provide a framework for a fast generation of per-layer quantization parameters. When tested on a deep capsule network model for the CIFAR10 dataset,

the proposed approach reduces the memory footprint by 6.2× with only a 0.15% accuracy loss. Secondly, we present NASCaps, an automated framework for the hardware-aware neural architecture search (NAS) of different types of DNNs, covering both traditional convolutional CNNs and capsule networks. The aim is to optimize the network accuracy and hardware efficiency, expressed in terms of energy, memory, and latency of a given hardware accelerator executing the DNN inference. The framework is tested on different datasets, generating various network configurations, and demonstrating the tradeoffs between the different output metrics.

Overall, this thesis presents novel techniques to overcome the challenges of CNNs and capsule networks and achieve efficient hardware acceleration. The results demonstrate that the proposed techniques improve the throughput and energy efficiency of the neural networks, which can have a significant impact on the development of efficient and accurate AI systems.