Implémentation FPGA performante pour réseaux de neurones quantifiés

Lieu et contacts

Laboratoire TIMA, 46 avenue Félix Viallet, Grenoble, France.
Contact : Adrien Prost-Boucle, adrien.prost-boucle@univ-grenoble-alpes.fr
Contact : Frédéric Pétrot, frederic.petrot@univ-grenoble-alpes.fr

(Version française)

Mots-clés : *Accélération Matérielle, FPGA, Réseaux de Neurones*

Contexte : Nous avons un environnement de modélisation et de génération de circuits accélérateurs de réseaux de neurones sur FPGA. Afin de réduire l'énergie consommée, les communications avec l'extérieur du FPGA sont réduite au maximum. L'architecture des circuits accélérateurs considérés est adaptée à cet objectif. Ces circuits sont de type pipeline, avec de nombraux types de composants de calcul connectés en série. Ces composants implémentent les opérations de base nécessaires aux réseaux de neurones numériques, notamment de unités de mémoire avec lecture par fenêtre à décalage, des unités de multiplication matrice-vecteur, opérations ReLU, etc.

Objectifs : Des développements supplémentaires sont nécessaires afin d'augmenter les performances, réduire les ressources matérielles utilités et améliorer la qualité des résultats. Cela concerne à la fois l'amélioration de composants existants, et la création de nouveaux composants.

Les réseaux ciblés sont fortement quantifiés, ce qui signifie que les valeurs manipulées sont représentées sur un petit nombre de bits (binaire, ternaire, 8 bits, etc). On s'intéresse à l'amélioration des réseaux de neurones binaires, à la gestion de la quantification hétérogène, et à la rédction des ressources consommées.

Les développement principaux à effectuer sont les suivants :

— implémenter un composant de *binarisation* (pour activations binaires),
— implémenter un composant de répétition de trames de données à base de *buffer ping-pong*,
— implémenter des composants de sélection et insertion de valeurs dans des trames de données,
— réduire le temps d'exécution d'un composant mémoire avec lecture par fenêtre glissante,
— réduire les ressources utilisées par les composants de neurones parallélisés au maximum,
— tester une implémentation optimisée additionneur multi-entrée signé,
— valider les nouveau développements par simulation RTL,
— effectuer des tests de synthèse et d'exécution sur carte FPGA.

Prérequis :

— Solides connaissances en architectures de circuits numériques, notamment FPGA
— Maîtrise du langage VHDL
— Bases de la programmation en C/C++
— Une première expérience avec les outils Xilinx Vivado/Vitis serait un plus
— Intérêt pour les technologies FPGA et l'accélération matérielle

Références :

— High-Efficiency Convolutional Ternary Neural Networks with Custom Adder Trees and Weight Compression (2019)
`https://cnrs.hal.science/hal-01686718v2`

(English version)

Title : Efficient FPGA implementation of quantized neural networks

Keywords : *Hardware acceleration, FPGA, Neural Networks*

Context : We have a framework to model and generate hardware accelerators of neural networks on FPGA. To reduce energy consumption, off-chip communications are reduced to the strict minimum. The architecture of considered accelerator circuits is designed to fit this objective. These circuits are mainly pipelines with numerous computing components of various types connected in series. These components implement base operations necessary to digital neural networks, in particular memory units with sliding window based read operation, matrix-vector multiplication units, ReLU operation, etc.

Objectives : Further development is necessary to increase performance, reduce hardware resource usage and improve quality of results. This applies both to existing components and to creation of new components.

Target networks are strongly quantized, which means the values involved in the computations are represented on a a low number of bits (binary, ternary, 8 bits, etc). The interest is on improvements of binary neural networks, on handling of heterogeneous quantization, and on reduction of hardware resource usage.

The main tasks of this internship are the following :
— to implement a *binarization* component, for binary activations,
— to implement a data repeater component based on a ping-pong buffer,
— to implement data selection and insertion components from streams of data transactions,
— to reduce the execution time of an existing sliding window memory component,
— to reduce the resources used by a neuron component under maximum parallelism,
— to test an optimized implementation of multi-input signed adder tree,
— to validate new developments with RTL simulation testbenches,
— and to perform synthesis tests and actual runs on an FPGA board.

Prerequisites :
— Solid knowledge in architecture of digital circuits, especially on FPGA
— Solid knowledge of VHDL language
— Good programming skills C/C++
— Experience with Xilinx tools Vivado/Vitis would be a plus
— Interest in FPGA technologies and hardware acceleration

References :
— High-Efficiency Convolutional Ternary Neural Networks with Custom Adder Trees and Weight Compression (2019)
`https://cnrs.hal.science/hal-01686718v2`