Energy-Efficient Near-Memory Computing Architecture for CNN at Cache Level

Masoud Nouripayam, Arturo Prieto, Vignajeth Kuttuva Kishorelal, and Joachim Rodrigues
Dept. of Electrical and Information Technology, Lund University, Sweden
firstname.lastname@{eit.lth.se}

Abstract

A non-von Neumann Near-Memory Computing architecture, optimized for CNN inference in edge computing, is integrated in the cache memory sub-system of a microcontroller unit with a RISC-V core. The NMC co-processor is evaluated using an 8-bit fixed-point quantized CNN model, and achieves an accuracy of 98% on the MNIST dataset. A full inference of the CNN model executed on the NMC processor, demonstrates an improvement of more than 34× in performance, and 28× in energy-efficiency, compared to the baseline scenario of a conventional RISC-V processor. The design achieves a performance of 1.39 GOPS (at 200 MHz) and an energy-efficiency of 49 GOPS/W, with negligible area overhead of less than 1%. [1]

Index Terms

Near-Memory Computing (NMC), SRAM, Cache, Embedded Systems, Microcontroller Unit (MCU), IoT, Convolutional Neural Networks (CNN), Low-Power Computing.

REFERENCES

[1] M. Nouripayam, A. Prieto, V. K. Kishorelal and J. Rodrigues, "An Energy-Efficient Near-Memory Computing Architecture for CNN Inference at Cache Level," *28th IEEE International Conference on Electronics, Circuits, and Systems (ICECS)*, pp. 1-4, 2021.