INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	Palash Das
Roll Number	:	156101001
Programme of Study	:	Ph.D.

Thesis Title: Near-Memory acceleration of Convolutional Neural Networks by exploiting Parallelism, Sparsity, and Redundancy.

Name of Thesis Supervisor(s)	:	Prof. Hemangee K. Kapoor.
Thesis Submitted to the Department/ Center	:	CSE
Date of completion of Thesis Viva-Voce Exam	:	07.04.2022
Key words for description of Thesis Work	:	Near-memory Processing, Convolutional Neural Networks, Accelerated Architectures, CNN accelerators.

SHORT ABSTRACT

The gap between the processing speed of the CPU and the access speed of the memory is becoming a bottleneck for many emerging applications. This gap can be reduced if the computation can be taken closer to the memory through near-memory processing (NMP). Among the logic options, application-specific integrated circuits (ASICs) are highly efficient in terms of power and area overhead for NMP logic integration. In this thesis, we aim to accelerate Convolutional Neural Networks (CNNs) by integrating custom hardware near the memory. As CNNs are widely used in several emerging applications, the designed hardware can be extensively used in all such cases. To design an NMP-based system with high performance and energy efficiency, we explore various techniques such as leveraging parallelism, exploiting data sparsity, and utilizing computation redundancy to reduce the number of operations. All such techniques result in hardware designs that implement the appropriate dataflow and data-parallel algorithm. The designs have positively impacted the system's performance and energy efficiency. To examine the deployability of the NMP approach, we perform experiments on various memory technologies like 3D memory, hybrid memory, and the commodity DRAM. Additionally, we also measure the efficacy of NMP for other applications like database operations. The proposed systems have performed substantially well while comparing them with various baselines and state-of-the-art works.