Final Report

Project #43: Reconfigurable Neural Processing Unit (NPU) for Energy-Efficient AI at the Edge

Kristelle Sampang

Project Report

Project Partner: Pratham Chhabra

Supervisor: Dr Morteza Biglari-Abhari

Co-Supervisor: Dr Maryam Hemmati

15th October 2025

Waipapa Taumata Rau
University of Auckland

COMPSYS043-2025

RECONFIGURABLE NEURAL PROCESSING UNIT (NPU) FOR ENERGY-EFFICIENT AI AT THE EDGE

Kristelle Sampang

ABSTRACT

Abstract goes here.

DECLARATION

Student I hereby declare that:

1. This report is the result of the final year project work carried out by my project partner (see cover page) and I under the guidance of our supervisor (see cover page) in the 2025 academic year at the Department of Electrical, Computer and Software Engineering, Faculty of Engineering, University of Auckland.
2. This report is not the outcome of work done previously.
3. This report is not the outcome of work done in collaboration, except that with a potential project sponsor (if any) as stated in the text.
4. This report is not the same as any report, thesis, conference article or journal paper, or any other publication or unpublished work in any format.

In the case of a continuing project, please state clearly what has been developed during the project and what was available from previous year(s):

Signature:

Date: 19/10/2025

Table of Contents

Acknowledgements

Thank important people here.

Glossary of Terms

Term	Definition

Abbreviations

AOA	**Angle of attack**

1. Introduction

1.1 Motivation

- As the demand for artificial intelligence grows to become prominent in today's society, its energy consumption has become a key issue.
- The processing required to run machine learning models is large, with millions of computations needed to be executed.
- This means that low-power processing devices at edge are limited to its use.
- The processing power to run machine learning models is large, limiting its use for low-power processing devices at the edge.
- With high computational power required, energy-efficiency is a key concern, especially nowadays when energy is an essential resource to not waste.
- In the past decade, common types of processors for running AI are CPU, GPU, and FPGA.
- However, as technology continues to improve, Neural Processing Units (NPU) are developed. These are processors that specialise in processing AI computations.
- By integrating NPU alongside other processors, the performance and energy-efficiency are improved compared to a standalone processor.

1.2 Problem Statement

1.3 Report Structure

The remainder of this report goes as follows: Section X covers Y....

2. Background

The rapid growth of Artificial Intelligence (AI) in the past decade has driven significant advancements across numerous field. Machine Learning (ML) models, particularly Deep Neural Networks (DNNs), are popular for its applications in image classification to autonomous driving [1]. There has been significant shift from AI inferencing occur at a cloud-level to resource-constrained edge devices. This move motivates the "AI at the Edge" in the research, where lower latency, enhanced data privacy, and real-time processing capabilities must are requirements that must be met without relying on constant network connection [2].

However, this shift presents an alarming issue where DNNs are computationally intensive and power-hungry, whilst edge devices operate under strict power and resource limitations. To bridge this gap, hardware accelerators, such as Neural Processing Units (NPUs) are introduced to execute AI algorithms at faster rates than general-purpose CPUs [3]. This section provides necessary background on the core technologies that support this project, starting with the most popular model for image-based tasks, the Convolutional Neural Network.

2.1 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) is a prominent type of DNN model, mainly utilised image processing tasks, such as object recognition, classification, and detection. The network is compromised of four main layers: convolutional, activation layer, pooling layers, and fully connected layers, where this research focuses on the convolutional and activation layers. The convolutional layer is where majority of the computations occur [4], as it extracts features from an image and converts it into numerical values. In a convolutional layer, there are several filters that slide through the image, searching for a specific pattern. These filters are typically in the size of 3x3 or 5x5, and is applied to the image by multiplying the filter by the 2D pixel representation of the image. Mathematically, this operation can be represented as

$$O[h][w][c] = \sum_{i=1}^{f_h} \sum_{j=1}^{f_w} \sum_{k=1}^{i_c} I[h + i \cdot s_h][w + j \cdot s_w][k] \times F[i][j][k][c]$$

where I, O, F are input activation, output activation, and filter weights respectively [4]. This can be represented as an enormous number of Multiply-Accumulate (MAC) operations, making CNNs computationally intensive.

An activation layer determines whether a neuron should be activated based on its input. Its primary role is to introduce non-linearity into the network. Without this, a neural network can only learn simple, linear patterns. Non-linearity allows the network to execute complex tasks, such as learning complicated patterns. A commonly used function is the Rectified Linear Unit (ReLU). The main functionality is to allow for positive inputs to remain unchanged whist setting any negative input to zero. A critical consequence of this is that it introduces significant sparsity as approximately half of the elements are zero [5]. This sparsity is a key property that this project exploits to improve energy-efficiency, and will be discussed further in Section X.

References

[1] A. Parashar et al., *SCNN: An Accelerator for Compressed-sparse Convolutional Neural Networks*, en, arXiv:1708.04485 [cs], May 2017. DOI: 10.48550/arXiv.1708.04485. Accessed: Sep. 18, 2025. [Online]. Available: http://arxiv.org/abs/1708.04485.

[2] B. Kim, S. Lee, A. R. Trivedi, and W. J. Song, "Energy-Efficient Acceleration of Deep Neural Networks on Realtime-Constrained Embedded Edge Devices," en, *IEEE Access*, vol. 8, pp. 216 259–216 270, 2020, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3038908. Accessed: Mar. 31, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/9262933/.

[3] E. Manor and S. Greenberg, "Custom Hardware Inference Accelerator for TensorFlow Lite for Microcontrollers," en, *IEEE Access*, vol. 10, pp. 73 484–73 493, 2022, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3189776. Accessed: Mar. 31, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/9825651/.

[4] J. Choi et al., "Enabling Fine-Grained Spatial Multitasking on Systolic-Array NPUs Using Dataflow Mirroring," en, *IEEE Transactions on Computers*, vol. 72, no. 12, pp. 3383–3398, Dec. 2023, ISSN: 0018-9340, 1557-9956, 2326-3814. DOI: 10.1109/TC.2023.3299030. Accessed: Apr. 2, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10198513/.

[5] W. Sun, D. Liu, Z. Zou, W. Sun, S. Chen, and Y. Kang, "Sense: Model-Hardware Codesign for Accelerating Sparse CNNs on Systolic Arrays," en, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 31, no. 4, pp. 470–483, Apr. 2023, Publisher: Institute of Electrical and Electronics Engineers (IEEE), ISSN: 1063-8210, 1557-9999. DOI: 10.1109/tvlsi.2023.3241933. Accessed: Jul. 22, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10043636/.