Summer 7-1-2020

Optimizing Neural Network Operations using Hexagon Vector eXtensions

Abhilash Panda

Optimizing Neural Network Operations using Hexagon Vector eXtensions

A project report submitted

to

MANIPAL ACADEMY OF HIGHER EDUCATION

For Partial Fulfillment of the Requirement for the

Award of the Degree

of

Bachelor of Technology

in

Information Technology

by

Abhilash Panda

Reg. No. 160911 078

Under the guidance of

Ipsita Upasana
Assistant Professor
Department of I&CT

Manipal Institute of Technology
Manipal, India

Pradeep N S
Senior Chief Engineer
On-Device AI

Samsung R&D Institute India- Bangalore
Bangalore, India

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

JULY 2020

ABSTRACT

AI-based applications have in the recent past been cloud-based. This involves information gathered and sent from a device on to a cloud that has the resources to apply computationally intensive machine learning models. Cloud-powered AI has some disadvantages including latency, reliability, security, and privacy. On-device AI solves this issue by localizing computation to the device. The trend toward Localized AI has been driven by two factors. The first factor is the increase in computing power available on end devices. The second is the efforts directed at making AI algorithms more efficient through neural network acceleration methods. The focus of this project is based on this objective of accelerating neural networks by leveraging Hexagon Vector Extension (HVX) – extensions to the Qualcomm Hexagon Digital Signal Processor. These extensions are designed to handle computer vision and image processing workloads.

The project required working with nnlib which is a library for the Hexagon NN Offload Framework. The offload framework is used by the Hexagon SDK for accomplishing computational offload to a DSP runtime.

As part of a broader objective of working on the optimization of neural networks, the project involved working on new approaches to performing convolution. The approach undertaken and covered in this report is a GEMM+im2col algorithm.

[Computing Methodologies]: Machine Learning- Machine Learning Approaches
[Computer Systems Organization]: Architectures- Parallel Architectures