Heriot-Watt University

Masters Thesis

Neural Network Compression

Author:
David Turner

Supervisor:
Dr. Robert Stewart

A thesis submitted in fulfilment of the requirements for the degree of MSc. Artificial Intelligence with Speech and Multimodal Interaction

in the

School of Mathematical and Computer Sciences

February 2020

Declaration of Authorship

I, David Turner, declare that this thesis titled, 'Neural Network Compression' and the work presented in it is my own. I confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed: _____

Date: _____

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Introduction to the report

Chapter 2

Literature Review

15-20 pages

2.1 Processor Architectures for deep learning

2.1.1 High Performance Devices

Include numbers here relating to memory and performance metrics from papers including speed, accuracy, model size

2.1.1.1 GPUs

Hardware structure, Benefits drawbacks and current performance on inference

2.1.1.2 TPUs

Structure, benefits, drawbacks, and current performance

2.1.1.3 CPUs

Hardware structure, drawbacks, current performance

2.1.2 Low Power Edge Devices

Numbers of memory and performance metrics for each of these

2.1.2.1 FPGAs

- General Structure
- What makes them a good choice?

2.1.2.2 USB Accelerators

For each item in the list describe processor architecture and the current available performance figures

- Intel Neural Compute Stick
  - VPU Structure
  - VPU Stats and figures
- Google Coral USB Accelerator
  - TPU At Edge

2.1.2.3 Embedded GPUs

Qualcomm Arduino line, Apple Bionic Chips.

Embedded within phones for example arm and apple

2.1.2.4 Smart Home

Google home now has neural processing units

2.1.2.5 Edge Custom Solutions

Current companies offering solutions focused on accelerating machine learning and neural network inference

Nvidia Jetson Line NVIDIA EGX Graphcore Qualcomm adapteva viatech mediatek - Supplimenting cloud ai chip in device NeuroPilot Kalray AWS Inferentia Arm Intel® Nervana™ Neural Network processors. Inside Xeon CPUs custom asic

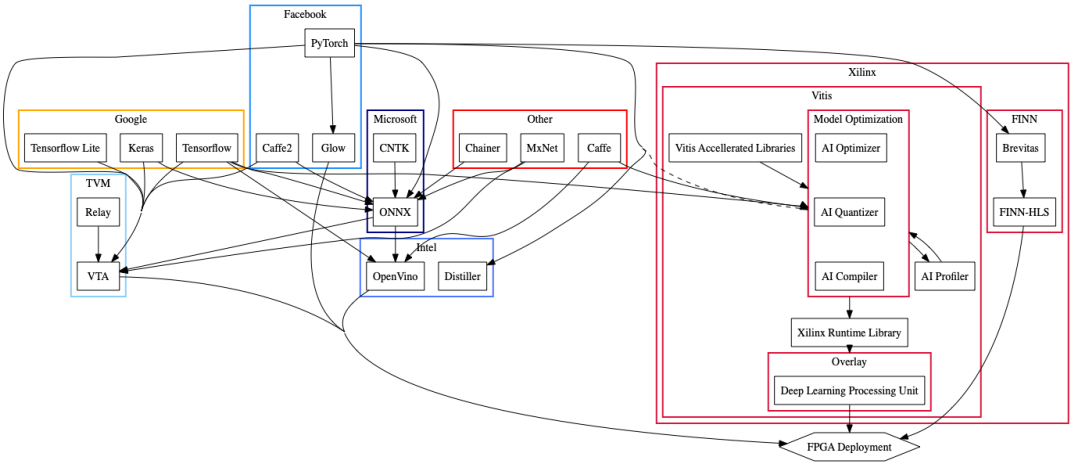

FIGURE 2.1: The state of Neural Network Frameworks for Edge Deployment

2.2 Model Compression Techniques

List of techniques and current results

2.2.1 Methods/Algorithms

2.2.1.1 Pruning

list of most interesting algorithms how they work Current available results

2.2.1.2 Quantization

Bit widths or weights and activation functions

2.2.1.3 Knowledge Distillation

2.2.1.4 Regularization

2.2.1.5 Conditional Computation

2.2.2 Frameworks

2.2.2.1 Intel Distiller

2.2.2.2 FINN

2.2.2.3 Intel OpenVino

2.2.2.4 Xilinx Vitis

2.3 Inference on edge devices

2.3.1 offloading edge inference

Offloaded execution of deeplearning inference at Edge: Challenges and Insights

2.4 benchmarking Neural networks

2.4.1 Image classification

2.4.1.1 Datasets

2.4.2 Object Detection

2.4.2.1 Datasets

Chapter 3

Requirements Analysis

3 pages atleast

3.1 Research Questions

How does model size impact time to inference? Does combining Pruning and Quantization help Model size and inference? How does offloading completely to cloud for inference compare to device inference? How can the DNN inference pipeline be transformed into a stream processing pipeline

3.2 hypothesis

3.3 Aim

3.4 Objectives

Below structure is from RM Lecture

3.5 Project Goals

Who are stakeholders

what are their needs

- Requirements analysis

Prioritise Needs

Distinguish functional and non functional requirements

Must, Should Could and Wont for requirements

Use Case studies

3.5.1 Functional Requirements

for example the network should be built using cnn architecture

3.5.2 Non Functional Requirements

for example which framework will be used

3.6 Deliverables and tasks

Chapter 4

Methodology

Datasets Preliminary ideas to model or system Experimantal setup and evaluation What is Acceptable Accuracy? Critical Accuracy? Take into account specificity and sensitivity in accuracy.

There is a need to measure these to evaluate accuracy and whether the acheived accuracy is good enough.

Chapter 5

Project Plan

How will each objective achieve the aim to allow for the hypothesis to be proved or disproved

5.1 Gantt Chart

5.2 Risk Analysis

Bibliography