Neural Network Compression

David Turner

January 15, 2020

1 Abstract

2 Introduction

Motivations and goals. There should also be the main hypothesis of the project. Why is this an interesting hypothesis to investigate. Use illistrations

3 Literature Review

15-20 pages

3.1 Processor Architectures for deep learning

3.2 High Performance Devices

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

3.2.1 GPUs

3.2.2 TPUs

3.2.3 CPUs

3.3 Low Power Edge Devices

Numbers of memory and performance metrics for each of these

3.3.1 FPGAs

- General Structure
- What makes them a good choice?

3.3.2 USB Accelerators

- Intel Neural Compute Stick
  - VPU Structure
  - vpu figures

- Google Coral USB Acellerator
    - TPU At Edge

3.3.3 Embedded GPUs

Embedded within phones for example arm stuff and apple

3.3.4 Smart Home

Google home now has neural processing units

3.3.5 Edge Custom Solutions

- 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
- custom asic

4 Compression Techniques

4.1 Methods/Algorithms

4.1.1 Pruning

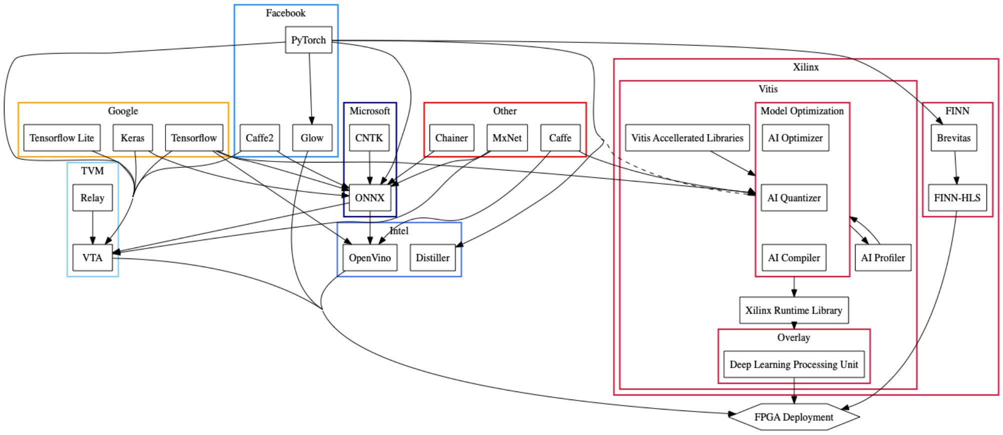
4.1.2 Quantization

4.1.3 Knowledge Distillation

4.1.4 Regularization

4.1.5 Conditional Computation

4.2 Frameworks



4.2.1 Intel Distiller

4.2.2 FINN

4.2.3 Intel OpenVino

4.2.4 Xilinx Vitis

5 Requirements Analysis

3 pages atleast [1]

5.1 Research Questions

5.2 hypothesis

5.3 Aim

5.4 Objectives

6 Methodology

Datasets Preliminary ideas fo model or system Experimantal setup and evaluation

7 Project Plan

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

7.1 Gantt Chart

7.2 Risk Analysis

References

[1] Wenlin Chen, James T Wilson, Stephen Tyree, Kilian Q Weinberger, and Yixin Chen. Compressing convolutional neural networks. *arXiv preprint arXiv:1506.04449*, 2015.