0.1 Motivation

With the continued revolution of AI technologies a desire to **perform inference at the edge (rephrase)** is becoming ever more prevalent. The argument for localising inference is only becoming stronger with the ever increasing avaliablilty of computation resources alongside new and constantly evolving AI applications, inference at the edge can provide better privacy and latency than the remote datacenter alternatives.

Neural network compression is one avenue for bringing inference to the edge, intuitively we might think that a network with a smaller memory footprint would naturally have lower inference latency but this is often not always the case. Utilising neural network compression effectively requires expert level knowledge of not only the network structure but the consequences of compression because compression techniques such as pruning can have cascading effects througout a neural network. This alone can make compression a daunting task, even for experienced machine learning practicioners, it gets worse however, these compression algorithms often feature complex parameters with implications that may not be revealed until a substantial amount of time has already been invested in retraining a compressed model.

0.2 Terminology

0.3 Hypothesis

Using a systematic compression method selection process combined with a bayesian optimisation algorithm we can partially automate compression parameter selection and imporve inference latency based on an accuracy threshold in a typical edge computing environment.

0.4 Research Aims

Aim 1 - This dissertation will research methodologies for reducing inference latency using a collection of off-the-shelf compression techniques, we will investigate which compression techniques have a positive effect on inference latency, and consider the context of this improvement with respect to the layer structure of the neural network.

Aim 2 - We will use this contextual information to select approriate compression methodologies and reduce the search space down to a single pruning algorithm per sub domain.

Aim 3 - Maintain a valid testing environment by using an edge based ai accelerator to perform inference, while training and compression will be performed on a GPU.

Aim 4 - Develop a interface to optimise compression parameters according to a metric representing the union of accuracy and latency.

Objectives

- **O0:** Develop a methodology to verify that the compression methods are actually being applied to the model being represented.
- **O1:** Select at least 1 neural network model to use for testing.
- **O2:** Select 2 suitable datasets for testing with a significant distinction between the cardinality of categories.
- **O3:** Evaluate a pool of compression algorithms with respect to end-to-end latency.
- **O4:** Measure latency for individual layers during inference.
- **O5:** Investigate the effect of composing select algorithms from different compression categories.
- **O6:** Select compression parameters to optimise.
- **O7:** Develop a interface to parameterise select compression methods.
- **O8:** Evaluate a model using a bayesian optimisation approach on compression parameters.