

*Summarising the context of the dissertation project, stating the aim and objectives of the project, identifying the problems to be solved to achieve the objectives, and sketching the organisation of the dissertation. Good intro: hypothesis (assertion) one aim, set of objectives for meeting aim, try and quantify objectives (25% better perf)*

## 0.1 Motivation

With the continued revolution of AI technologies a desire to perform inference at the edge is becoming ever more prevalent. The argument for localising inference is only becoming stronger with the ever increasing availability 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 and knock-on effects of compression. This alone can make compression a daunting task, even for experienced machine learning practitioners, however these compression algorithms often feature complex parameters that make compression even less accessible.

## 0.2 Hypothesis

*Using a systematic model selection process combined with a bayesian optimisation algorithm we can partially automate compression parameter selection and improve inference latency above an accuracy threshold in a typical edge computing environment.*

### 0.3 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 compression methodologies and apply a bayesian optimisation strategy on the compression parameters.

**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 framework to optimise compression parameters according to a single metric.

#### 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 computer vision model to use for testing.
- **O2:** Select 2 suitable datasets for testing with a significant distinction between the cardinality of classes.
- **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 framework to parameterise select compression methods.
- **O8:** Evaluate a model using a bayesian optimisation approach on compression parameters.

### 0.4 Document structure

outline the document: We start with ..., then we cover x, y, and z ...