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 avaliablilty of computation resources alongside new and constatnly 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 the case.

0.2 Hypothesis

An appropriate combination of compression techniques applied in a layer-context-aware manner can improve inference latency without reducing accuracy significantly whilst constrained within a typical edge computing environment.

0.3 Research Aims

This dissertation will research methodologies for reducing inference latency with 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 internal structure of the neural network.

1. This research project will explore a pool of compression techniques and apply a varied composition of them in a manner that is sensitive to the context of the individual layer structures within a network.

2. We will seek to optimise inference latency within a typical edge environment.

3. The reasearch will attempt to provide evidence of effective compression techniques for a given layer type within a conventional neural network.

Issues with limited resource computation [1]

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