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.

With the revolution of AI technologies a greater need 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 power alongside new and constatnly evolving AI applications, inference at the edge can provide better privacy and latency than the remote datacenter alternatives. This dissertation will focus on methodologys for improving inference performance with preexisting compression techniques.

These models can have a huge number of parameters so inference can sometimes be impractical. [1] - "see Table 1"

Issues with limited resource computation [2]

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