# Summary

Quantisation is a commonly applied technique to improve the efficiency of a neural network on edge devices. Apart from a quantised model, most of the quantisation-aware training methods maintain an extra full-precision model to prevent large accuracy losses. But keeping full-precision model needs sufficient memory and energy. To address this problem, this paper proposed the Adaptive Precision Training method (APT), which only keeps a quantised model to save memory. To solve the issue of quantisation underflow, they employ a matric called G to quantify the learning ability of each layer and dynamically adjusts per-layer bitwidth.

Experiments on image classification and text classification tasks suggest that APT saves 60-72.5% memory space for model parameters with limited accuracy loss, compared with the 8-bit traditional QAT method.

Several questions are considered as follows:

1. In introduction of page 2, you mentioned that the training of a quantised neural network requires more memory space and communication than full-precision does. But why