# 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 which edge devices don’t have. 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 is that? Quantisation is meant to reduce the storage space of the model at the cost of accuracy, so it is supposed to require less memory space. Why it requires more communication should be explained as well.
2. In figure 1 (page 4), the meaning of M in (c) is not explained, and the relationship between the three graphs should be clarified at the beginning when introducing them, showing what they represent respectively, like they represent the forward, backward and update.
3. In page 5, you mentioned that the communication computation ratio of the update operations is higher than that in the forward and backward pass. What exactly is the definition of the communication computation ratio? What do these two parts refer to? Please add the definition.
4. In formula 3(page 6), the maximum and minimum of weight tensor is needed to calculate the minimum resolution of the quantisation. But how do we get the maximum and minimum? Are they related to k? Why the resolution is given in this form? This formula needs further explanation.
5. In section 3.2(page 8), you used an interval of 100 iterations to valuates per-layer Gavg. How is this chosen? Does different number of iterations influence the training result? Have you done any experiment on it?
6. In the experiment, the interval of evaluation was set to 100 iterations, while the interval of adjustment was set to 1000 iterations. Why should the adjustment be performed every 10 evaluations instead of every time or other times of evaluations?
7. In table 5(page 16), the comparison with other training method should involve accuracy on a certain experiment, which is also meaningful despite the former experiment on different datasets.
8. In figure 6(b) (page 20), the legends of the table are not explained, and the difference between bars of different legend should also be analysed.