**Why is this an important area of study? [0.5 points]**

Quantization allows us to use smaller data types for calculation, which makes training or inference faster and results in less memory-consumption.

**Describe two different techniques/approaches discussed [0.5 points]**

1. Post-training Quantization (PTQ): perform quantization without re-training the model by mapping the model weights and input into a smaller data type using a scale factor and bias.
2. Quantization-aware Training (QAT): use an additional quantization error as auxiliary loss function during training. This minimizes the precision loss. Gradients for the round operation are estimated.

**Discuss relative strengths [0.5 points] and weaknesses [0.5 points] of the two techniques described above. [1 point total]**

* PTQ can be easily applied to a pre-trained model. But it requires calibration, but the precision loss can be big depending on the choice of parameters.
* QAT requires training the model, but the precision loss due to quantization can be minimized.