Quantization Practice

1. Uniform Quantization

- **Objective:**
  - Run uniform quantization and dequantization on a 1D array of random values.
  - Observe the effects of these processes.

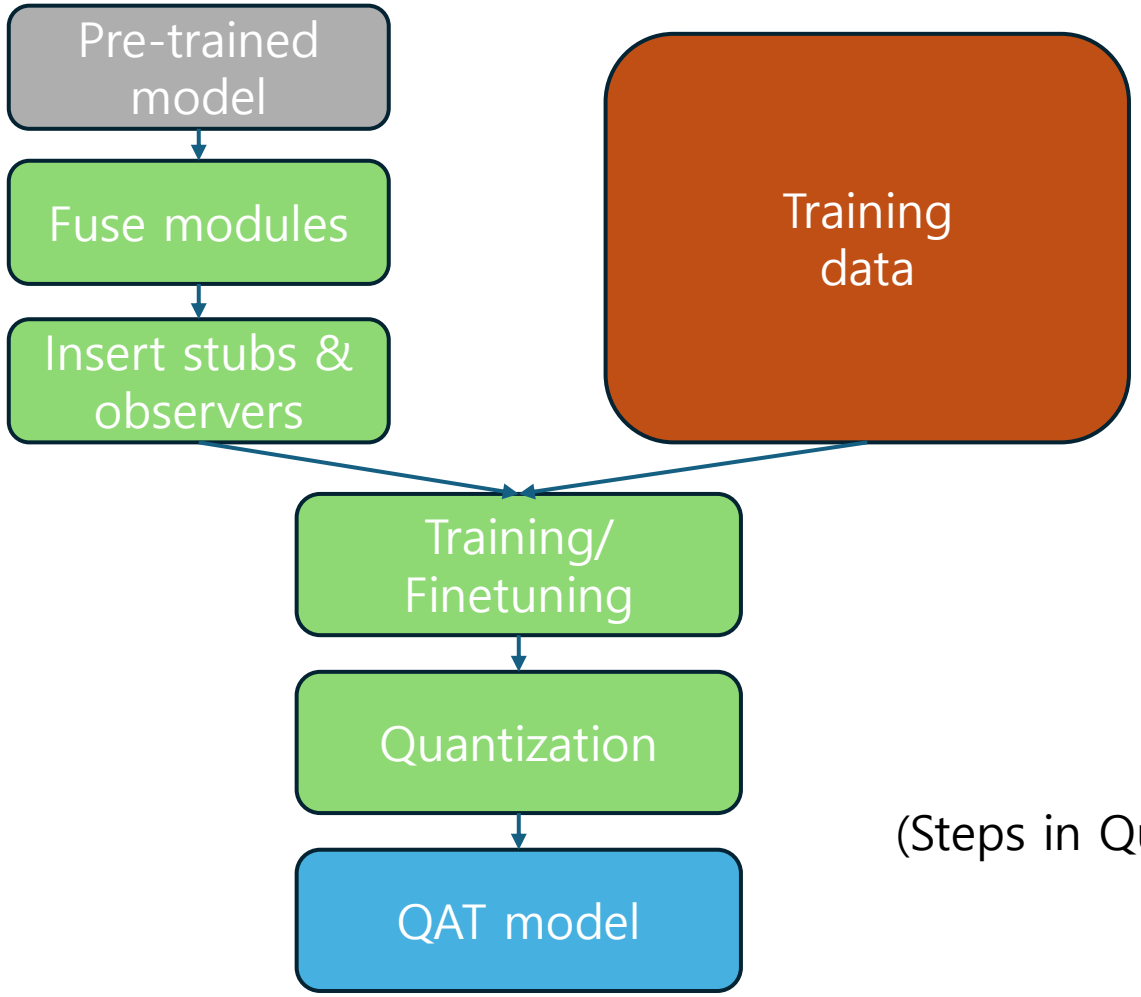
- **Practice:**
  - Experiment with the number of quantization bits and observe the (corresponding) quantization error

2. PyTorch Quantization

- **Objective:**
  - Learn how to use PyTorch Quantization Framework.
  - Compare the difference of the model statistics before and after quantization.
  - Measure and analyze the latency difference before and after quantization.

- **Practice:**
  - Modify custom quantization observer functions to understand their impact on model performance.

2. PyTorch Quantization



(Steps in Quantization aware training)

3. TensorRT INT8 calibration

- **Objective:**
  - Learn how to make TensorRT engine (executable model).
  - Apply TensorRT calibration.

- **Practice:**
  - Modify the parameters ‘calib_input’, ‘calib_num_images’, and ‘calib_batch_size’ to observe their effects on calibration accuracy.
  - [Optional] Experiment with ₩the image augmentation in the preprocess_image() function in image_batcher.py.