Getting familiar with optimization methods:

Theoretical side:

1. What is the difference between eager mode and graph mode(script mode) in deep learning frameworks?

2. What are the approaches for converting a model from eager mode to script mode in Pytorch?

3. According to the previous question, what are the advantages and disadvantages of these methods?

4. How does TensorRT help in optimizing a deep learning network? What are the strategies that it uses for optimization?

Practical Side:

1. Install PyTorch Docker Container from this link which contains all the libraries you need for compiling with TensorRT. (you need to first install Nvidia drivers, Docker Engine, and Nvidia Container toolkit on your Linux system)

2. Get familiar with docker images and discover how you can run your evaluation project on a docker image.

3. Use Torch-TensorRT for compiling your project and evaluate your increase in performance and memory usage. (this link is helpful for this part and the next one)

4. Use PTQ Option for quantizing your model.