Task 1: Run the code on the github for this paper https://arxiv.org/abs/2201.11113

I was not able to produce any results because the code base requires a large amount of storage (300 GB). I was looking into renting an Amazon S3 container to run it.

I struggled with running the code because of the sheer size of the dataset. The ImageNET dataset is the one used in the paper and the code. This dataset is quite large (300 GB). I do not have the disk space on my computer to clone the repo, and DSMLP provides only 10 GB of space.

I chose to look into AWS resources because I have used their cloud services before and am familiar with the interface. However, the last time I used it, I had access through the company where I interned, so paying for the resources was not an issue. Now, however, I need to be more judicious and try to be precise about my resource usage to avoid spending too much money.

Task 2: Generate 2 questions for discussion

Question 1: Does anyone have any suggestions as to how I can run the code for the paper?

Question 2: In quantization-aware training, how exactly is the quantization alphabet constraint enforced during training? During gradient descent, how are the weights constrained to the alphabet set?