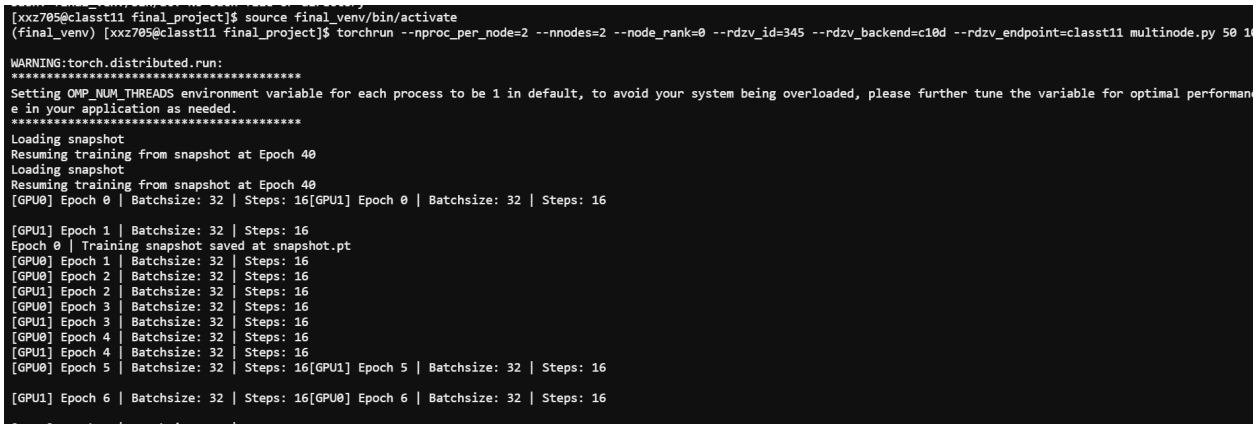
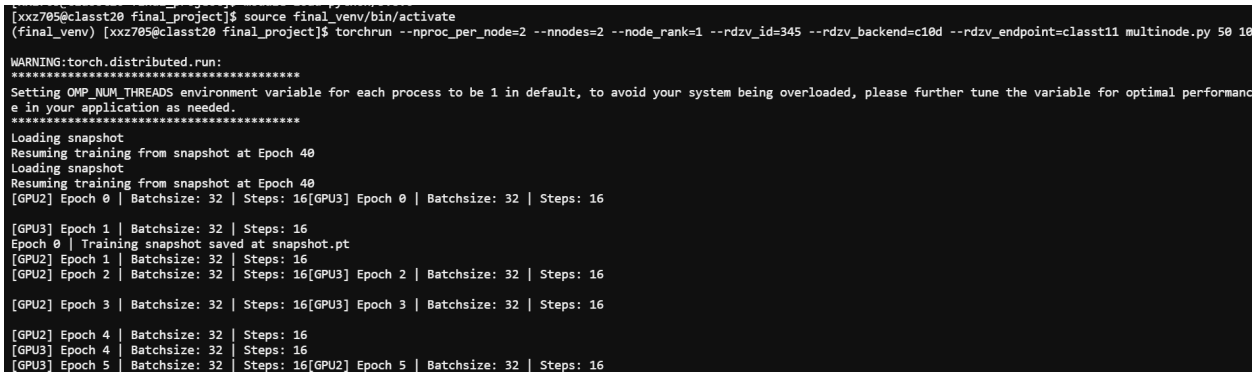
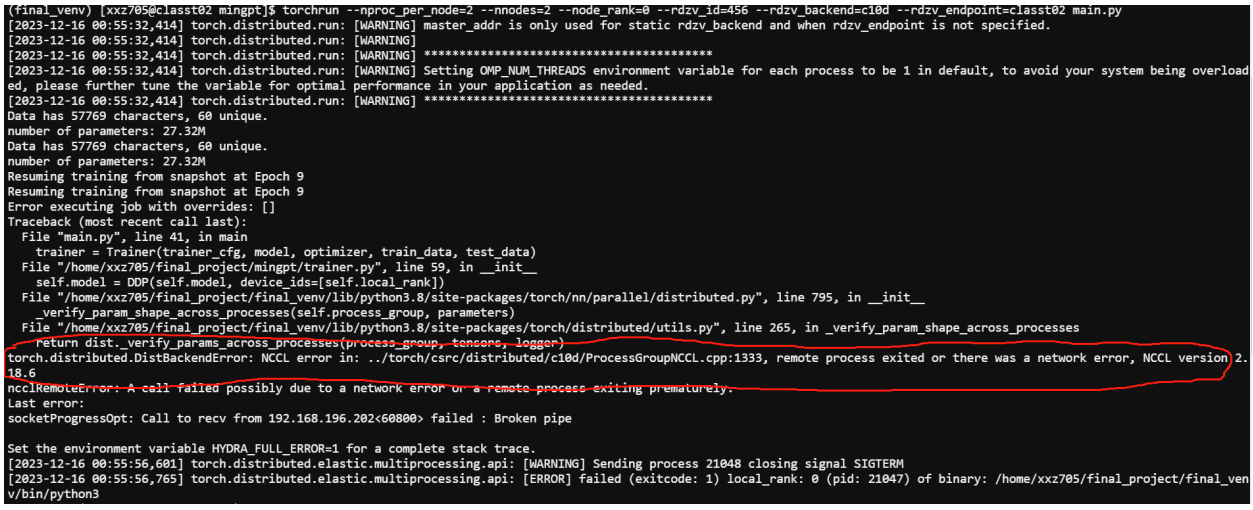
1.We believe that the implementation of the GPT and BERT using pytorch ddp is correct and can run on multiple nodes that have different number of gpus. The reason is because before updating the environment the GPT model is working, here is the screenshots of the console outputs on CASE hpc:





this is using the exact implementation we have right now on 12/14/2023, and you can see [gpu0], [gpu1], [gpu2], [gpu3] is the global rank of the gpus on different nodes

However, as the deadline approaches (it's technically past), when I want to rerun the GPT model after waiting for 6 hours for the resources. this is what I got:



I googled for a long time for an answer and here's the closest thing I got: https://discuss.pytorch.org/t/torch-distributed-distbackenderror-nccl-error/191509/10 because the person who had problem in this discussion thread didn't solve the problem at the end, but I have learned that this is some network error that I will never know.

Just like that, some network errors, I cannot even make the program works with multiple nodes. I have tried my best to fix this problem but so far there's no success. But I do want whoever grades this project understand that the program did work and failed mysteriously without me changing anything on my end.

for Bert model, the same error happens which strengthens my confidence in this belief: