1. Download anaconda3: <https://www.anaconda.com/products/distribution>

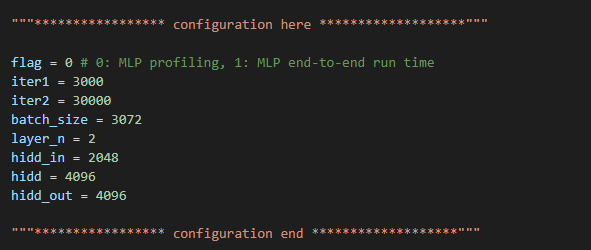
2. create new env, named as “bert” and export the environment from end-to-end.yaml: conda env create -f environment.yml. The python version is up to you, I used 3.7.13.

3. Before profiling, turn on or turn off the tf32 as needed:

<https://developer.nvidia.com/blog/profiling-and-optimizing-deep-neural-networks-with-dlprof-and-pyprof/>

<https://docs.nvidia.com/cuda/cutensor/user_guide.html>

4. for BERT, MLP, NCF, ViT, I left the configuration part labeled as below:



Flag = 0 is for profiling for each kernel, flag = 1 is for profiling total running time.

5. After setting up the correct configuration, you can direct run BERT.sh/MLP.sh/NCF.sh/ViT.sh and the profiling and power file will be saved.