

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:

```
[xxz705@classt11 final_project]$ source final_venv/bin/activate
(final_venv) [xxz705@classt11 final_project]$ torchrun --nproc_per_node=2 --nnodes=2 --node_rank=0 --rdzv_id=345 --rdzv_backend=c10d --rdzv_endpoint=classt11 multinode.py 50 10

WARNING:torch.distributed.run:
*****
Setting OMP_NUM_THREADS environment variable for each process to be 1 in default, to avoid your system being overloaded, please further tune the variable for optimal performance in your application as needed.
*****
Loading snapshot
Resuming training from snapshot at Epoch 40
Loading snapshot
Resuming training from snapshot at Epoch 40
[GPU0] Epoch 0 | Batchsize: 32 | Steps: 16[GPU1] Epoch 0 | Batchsize: 32 | Steps: 16

[GPU1] Epoch 1 | Batchsize: 32 | Steps: 16
Epoch 0 | Training snapshot saved at snapshot.pt
[GPU0] Epoch 1 | Batchsize: 32 | Steps: 16
[GPU0] Epoch 2 | Batchsize: 32 | Steps: 16
[GPU1] Epoch 2 | Batchsize: 32 | Steps: 16
[GPU0] Epoch 3 | Batchsize: 32 | Steps: 16
[GPU1] Epoch 3 | Batchsize: 32 | Steps: 16
[GPU0] Epoch 4 | Batchsize: 32 | Steps: 16
[GPU1] Epoch 4 | Batchsize: 32 | Steps: 16
[GPU0] Epoch 5 | Batchsize: 32 | Steps: 16[GPU1] Epoch 5 | Batchsize: 32 | Steps: 16

[GPU1] Epoch 6 | Batchsize: 32 | Steps: 16[GPU0] Epoch 6 | Batchsize: 32 | Steps: 16

[xxz705@classt20 final_project]$ source final_venv/bin/activate
(final_venv) [xxz705@classt20 final_project]$ torchrun --nproc_per_node=2 --nnodes=2 --node_rank=1 --rdzv_id=345 --rdzv_backend=c10d --rdzv_endpoint=classt11 multinode.py 50 10

WARNING:torch.distributed.run:
*****
Setting OMP_NUM_THREADS environment variable for each process to be 1 in default, to avoid your system being overloaded, please further tune the variable for optimal performance in your application as needed.
*****
Loading snapshot
Resuming training from snapshot at Epoch 40
Loading snapshot
Resuming training from snapshot at Epoch 40
[GPU2] Epoch 0 | Batchsize: 32 | Steps: 16[GPU3] Epoch 0 | Batchsize: 32 | Steps: 16

[GPU3] Epoch 1 | Batchsize: 32 | Steps: 16
Epoch 0 | Training snapshot saved at snapshot.pt
[GPU2] Epoch 1 | Batchsize: 32 | Steps: 16
[GPU2] Epoch 2 | Batchsize: 32 | Steps: 16[GPU3] Epoch 2 | Batchsize: 32 | Steps: 16

[GPU2] Epoch 3 | Batchsize: 32 | Steps: 16[GPU3] Epoch 3 | Batchsize: 32 | Steps: 16

[GPU2] Epoch 4 | Batchsize: 32 | Steps: 16
[GPU3] Epoch 4 | Batchsize: 32 | Steps: 16
[GPU3] Epoch 5 | Batchsize: 32 | Steps: 16[GPU2] Epoch 5 | Batchsize: 32 | Steps: 16
```

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:

```
(final_venv) [xxz705@classt02 mingpt]$ torchrun --nproc_per_node=2 --nnodes=2 --node_rank=0 --rdzv_id=456 --rdzv_backend=c10d --rdzv_endpoint=classt02 main.py
[2023-12-16 00:55:32,414] torch.distributed.run: [WARNING] master_addr is only used for static rdzv_backend and when rdzv_endpoint is not specified.
[2023-12-16 00:55:32,414] torch.distributed.run: [WARNING]
[2023-12-16 00:55:32,414] torch.distributed.run: [WARNING] *****
[2023-12-16 00:55:32,414] torch.distributed.run: [WARNING] Setting OMP_NUM_THREADS environment variable for each process to be 1 in default, to avoid your system being overloaded, please further tune the variable for optimal performance in your application as needed.
[2023-12-16 00:55:32,414] torch.distributed.run: [WARNING] *****
Data has 57769 characters, 60 unique.
number of parameters: 27.32M
Data has 57769 characters, 60 unique.
number of parameters: 27.32M
Resuming training from snapshot at Epoch 9
Resuming training from snapshot at Epoch 9
Error executing job with overrides: {}
Traceback (most recent call last):
  File "main.py", line 41, in main
    trainer = Trainer(trainer_cfg, model, optimizer, train_data, test_data)
  File "/home/xxz705/final_project/mingpt/trainer.py", line 59, in __init__
    self.model = DDP(self.model, device_ids=[self.local_rank])
  File "/home/xxz705/final_project/final_venv/lib/python3.8/site-packages/torch/nn/parallel/distributed.py", line 795, in __init__
    _verify_param_shape_across_processes(self.process_group, parameters)
  File "/home/xxz705/final_project/final_venv/lib/python3.8/site-packages/torch/distributed/utils.py", line 265, in _verify_param_shape_across_processes
    return dist._verify_params_across_processes(process_group, tensors, logger)
torch.distributed.DistBackendError: NCCL error in: ./torch/csrc/distributed/c10d/ProcessGroupNCCL.cpp:1333, remote process exited or there was a network error, NCCL version 2.18.6
ncclRemoteError: A call failed possibly due to a network error or a remote process exiting prematurely.
Last error:
socketProgressOpt: Call to recv from 192.168.196.202<60800> failed : Broken pipe

Set the environment variable HYDRA_FULL_ERROR=1 for a complete stack trace.
[2023-12-16 00:55:56,601] torch.distributed.elastic.multiprocessing.api: [WARNING] Sending process 21048 closing signal SIGTERM
[2023-12-16 00:55:56,765] torch.distributed.elastic.multiprocessing.api: [ERROR] failed (exitcode: 1) local_rank: 0 (pid: 21047) of binary: /home/xxz705/final_project/final_venv/bin/python3
```

<https://discuss.pytorch.org/t/torch-distributed-distbackenderror-nccl-error/191509/10>

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.

```

path to a single file or url is deprecated and won't be possible anymore in v5. Use a model identifier or the path to a directory instead.
warnings.warn(
/home/xxz705/final_project/final_venv/lib/python3.8/site-packages/transformers/tokenization_utils_base.py:1929: FutureWarning: Calling BertTokenizer.from_pretrained() with the
path to a single file or url is deprecated and won't be possible anymore in v5. Use a model identifier or the path to a directory instead.
warnings.warn(
Traceback (most recent call last):
  File "main.py", line 62, in <module>
    main()
  File "main.py", line 47, in main
    bert_trainer = BERTTrainer(bert_lm, train_loader)
  File "/home/xxz705/final_project/BERT/trainer.py", line 37, in __init__
Traceback (most recent call last):
  File "main.py", line 62, in <module>
    main()
  File "main.py", line 47, in main
    bert_trainer = BERTTrainer(bert_lm, train_loader)
  File "/home/xxz705/final_project/BERT/trainer.py", line 37, in __init__
    self.model = DDP(self.model, device_ids=[self.local_rank])
  File "/home/xxz705/final_project/final_venv/lib/python3.8/site-packages/torch/nn/parallel/distributed.py", line 795, in __init__
    self.model = DDP(self.model, device_ids=[self.local_rank])
  File "/home/xxz705/final_project/final_venv/lib/python3.8/site-packages/torch/nn/parallel/distributed.py", line 795, in __init__
    _verify_param_shape_across_processes(self.process_group, parameters)
  File "/home/xxz705/final_project/final_venv/lib/python3.8/site-packages/torch/nn/parallel/distributed.py", line 795, in __init__
    _verify_param_shape_across_processes(self.process_group, parameters)
  File "/home/xxz705/final_project/final_venv/lib/python3.8/site-packages/torch/distributed/utils.py", line 265, in _verify_param_shape_across_processes
    return dist._verify_params_across_processes(process_group, tensors, logger)
  File "/home/xxz705/final_project/final_venv/lib/python3.8/site-packages/torch/distributed/utils.py", line 265, in _verify_param_shape_across_processes
    return dist._verify_params_across_processes(process_group, tensors, logger)
torch.distributed.DistBackendError: torch.distributed: .NCCL error in: ../torch/csrc/distributed/c10d/ProcessGroupNCCL.cpp:1333, remote process exited or there was a network error
r, NCCL version 2.18.6
ncclRemoteError: A call failed possibly due to a network error or a remote process exiting prematurely.
Last error:
socketProgressOpt: Call to recv from 192.168.196.212:54702 failed : Broken pipeDistBackendError
: NCCL error in: ../torch/csrc/distributed/c10d/ProcessGroupNCCL.cpp:1333, internal error - please report this issue to the NCCL developers, NCCL version 2.18.6
ncclInternalError: Internal check failed.
Last error:
Socket recv failed while polling for opId=0x7fbd2c0de8d0
Socket recv failed while polling for opId=0x7fbd2c0de8d0

```