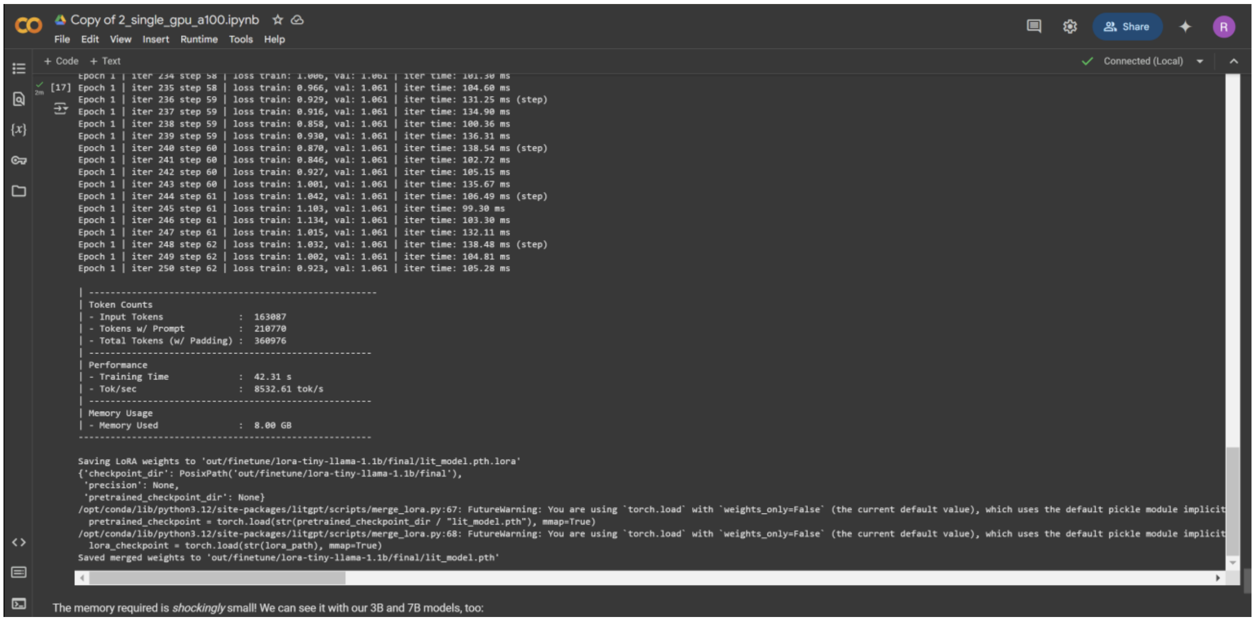
Parameter efficient fine tuning


Copy of 2_single_gpu_a100.ipynb
File Edit View Insert Runtime Tools Help
+ Code + Text
Share
Connected (Local)
[17]
2m

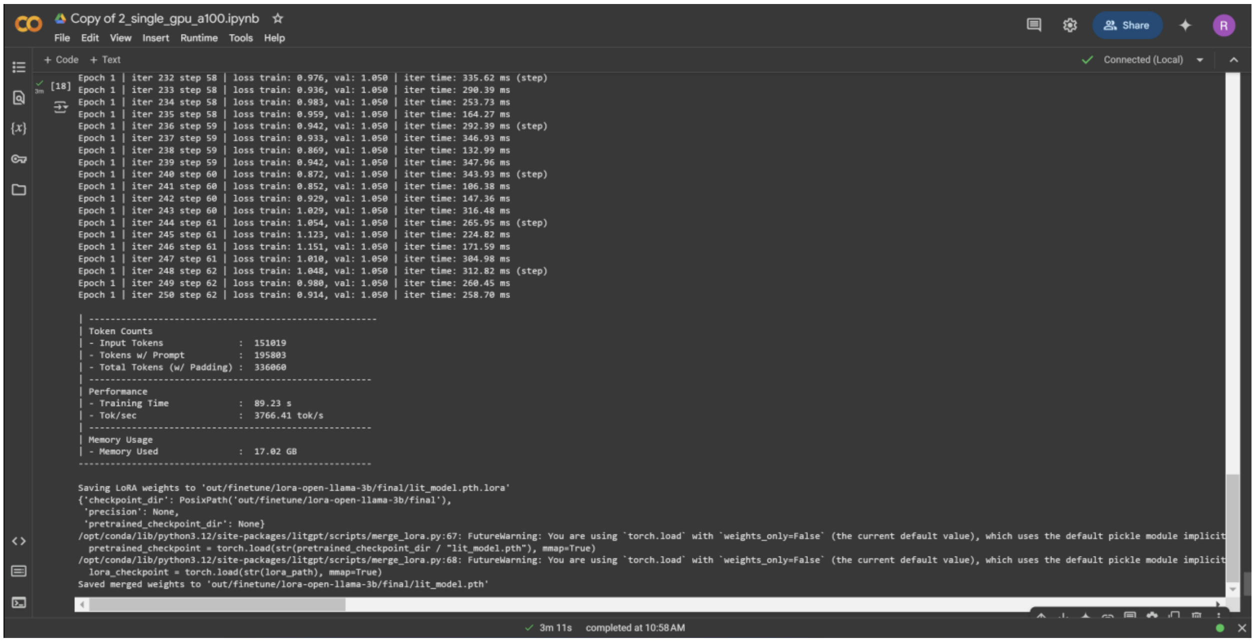
```
Epoch 1 | iter 234 step 58 | loss train: 1.006, val: 1.061 | iter time: 101.30 ms
Epoch 1 | iter 235 step 58 | loss train: 0.966, val: 1.061 | iter time: 104.60 ms
Epoch 1 | iter 236 step 59 | loss train: 0.929, val: 1.061 | iter time: 131.25 ms (step)
Epoch 1 | iter 237 step 59 | loss train: 0.916, val: 1.061 | iter time: 134.90 ms
Epoch 1 | iter 238 step 59 | loss train: 0.858, val: 1.061 | iter time: 100.36 ms
Epoch 1 | iter 239 step 59 | loss train: 0.930, val: 1.061 | iter time: 136.31 ms
Epoch 1 | iter 240 step 60 | loss train: 0.870, val: 1.061 | iter time: 138.54 ms (step)
Epoch 1 | iter 241 step 60 | loss train: 0.846, val: 1.061 | iter time: 102.72 ms
Epoch 1 | iter 242 step 60 | loss train: 0.927, val: 1.061 | iter time: 105.15 ms
Epoch 1 | iter 243 step 60 | loss train: 1.001, val: 1.061 | iter time: 135.67 ms
Epoch 1 | iter 244 step 61 | loss train: 1.042, val: 1.061 | iter time: 106.49 ms (step)
Epoch 1 | iter 245 step 61 | loss train: 1.103, val: 1.061 | iter time: 99.30 ms
Epoch 1 | iter 246 step 61 | loss train: 1.134, val: 1.061 | iter time: 103.30 ms
Epoch 1 | iter 247 step 61 | loss train: 1.015, val: 1.061 | iter time: 132.11 ms
Epoch 1 | iter 248 step 62 | loss train: 1.032, val: 1.061 | iter time: 138.48 ms (step)
Epoch 1 | iter 249 step 62 | loss train: 1.002, val: 1.061 | iter time: 104.81 ms
Epoch 1 | iter 250 step 62 | loss train: 0.923, val: 1.061 | iter time: 105.28 ms

| ----------------------------------------------------
| Token Counts
| - Input Tokens              :  163087
| - Tokens w/ Prompt          :  210770
| - Total Tokens (w/ Padding) :  360976
| ---------------------------------------------------
| Performance
| - Training Time             :  42.31 s
| - Tok/sec                   :  8532.61 tok/s
| ---------------------------------------------------
| Memory Usage
| - Memory Used               :  8.00 GB
----------------------------------------------------

Saving LoRA weights to 'out/finetune/lora-tiny-llama-1.1b/final/lit_model.pth.lora'
{'checkpoint_dir': PosixPath('out/finetune/lora-tiny-llama-1.1b/final'),
 'precision': None,
 'pretrained_checkpoint_dir': None}
/opt/conda/lib/python3.12/site-packages/litgpt/scripts/merge_lora.py:67: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicit
  pretrained_checkpoint = torch.load(str(pretrained_checkpoint_dir / "lit_model.pth"), mmap=True)
/opt/conda/lib/python3.12/site-packages/litgpt/scripts/merge_lora.py:68: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicit
  lora_checkpoint = torch.load(str(lora_path), mmap=True)
Saved merged weights to 'out/finetune/lora-tiny-llama-1.1b/final/lit_model.pth'
```

The memory required is *shockingly* small! We can see it with our 3B and 7B models, too:

1.1B model


Copy of 2_single_gpu_a100.ipynb
File Edit View Insert Runtime Tools Help
+ Code + Text
Share
Connected (Local)
[18]
3m

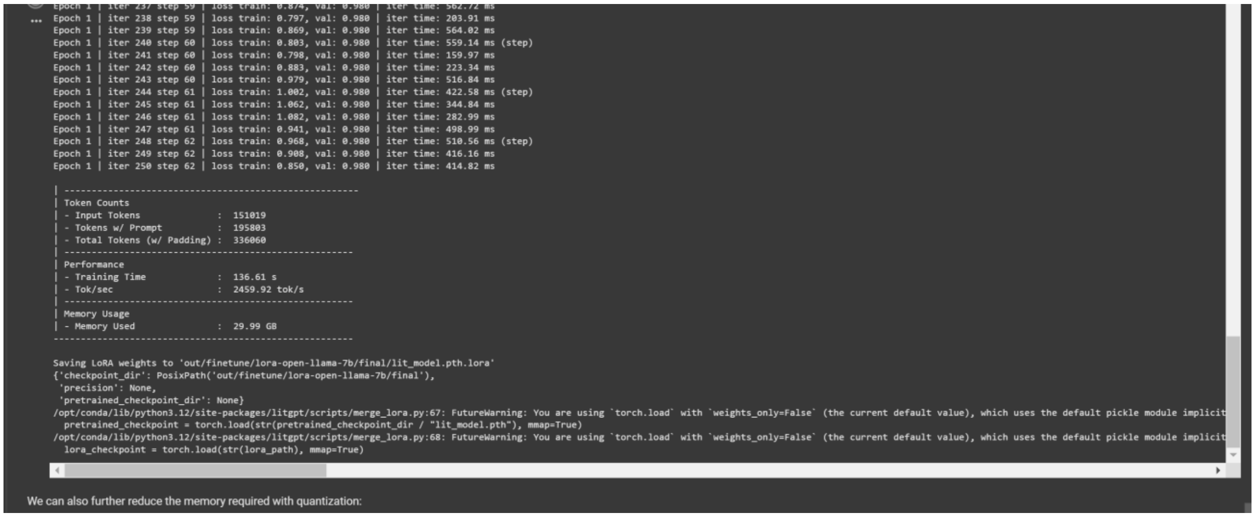
```
Epoch 1 | iter 232 step 58 | loss train: 0.976, val: 1.050 | iter time: 335.62 ms (step)
Epoch 1 | iter 233 step 58 | loss train: 0.936, val: 1.050 | iter time: 290.39 ms
Epoch 1 | iter 234 step 58 | loss train: 0.983, val: 1.050 | iter time: 253.73 ms
Epoch 1 | iter 235 step 58 | loss train: 0.959, val: 1.050 | iter time: 164.27 ms
Epoch 1 | iter 236 step 59 | loss train: 0.942, val: 1.050 | iter time: 292.39 ms (step)
Epoch 1 | iter 237 step 59 | loss train: 0.933, val: 1.050 | iter time: 346.93 ms
Epoch 1 | iter 238 step 59 | loss train: 0.869, val: 1.050 | iter time: 132.99 ms
Epoch 1 | iter 239 step 59 | loss train: 0.942, val: 1.050 | iter time: 347.96 ms
Epoch 1 | iter 240 step 60 | loss train: 0.872, val: 1.050 | iter time: 343.93 ms (step)
Epoch 1 | iter 241 step 60 | loss train: 0.852, val: 1.050 | iter time: 106.38 ms
Epoch 1 | iter 242 step 60 | loss train: 0.929, val: 1.050 | iter time: 147.36 ms
Epoch 1 | iter 243 step 60 | loss train: 1.029, val: 1.050 | iter time: 316.48 ms
Epoch 1 | iter 244 step 61 | loss train: 1.054, val: 1.050 | iter time: 265.95 ms (step)
Epoch 1 | iter 245 step 61 | loss train: 1.123, val: 1.050 | iter time: 224.82 ms
Epoch 1 | iter 246 step 61 | loss train: 1.151, val: 1.050 | iter time: 171.59 ms
Epoch 1 | iter 247 step 61 | loss train: 1.010, val: 1.050 | iter time: 304.98 ms
Epoch 1 | iter 248 step 62 | loss train: 1.048, val: 1.050 | iter time: 312.82 ms (step)
Epoch 1 | iter 249 step 62 | loss train: 0.980, val: 1.050 | iter time: 260.45 ms
Epoch 1 | iter 250 step 62 | loss train: 0.914, val: 1.050 | iter time: 258.70 ms

| ----------------------------------------------------
| Token Counts
| - Input Tokens              :  151019
| - Tokens w/ Prompt          :  195803
| - Total Tokens (w/ Padding) :  336060
| ---------------------------------------------------
| Performance
| - Training Time             :  89.23 s
| - Tok/sec                   :  3766.41 tok/s
| ---------------------------------------------------
| Memory Usage
| - Memory Used               :  17.02 GB
----------------------------------------------------

Saving LoRA weights to 'out/finetune/lora-open-llama-3b/final/lit_model.pth.lora'
{'checkpoint_dir': PosixPath('out/finetune/lora-open-llama-3b/final'),
 'precision': None,
 'pretrained_checkpoint_dir': None}
/opt/conda/lib/python3.12/site-packages/litgpt/scripts/merge_lora.py:67: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicit
  pretrained_checkpoint = torch.load(str(pretrained_checkpoint_dir / "lit_model.pth"), mmap=True)
/opt/conda/lib/python3.12/site-packages/litgpt/scripts/merge_lora.py:68: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicit
  lora_checkpoint = torch.load(str(lora_path), mmap=True)
Saved merged weights to 'out/finetune/lora-open-llama-3b/final/lit_model.pth'
```

✓ 3m 11s completed at 10:58 AM

3B model



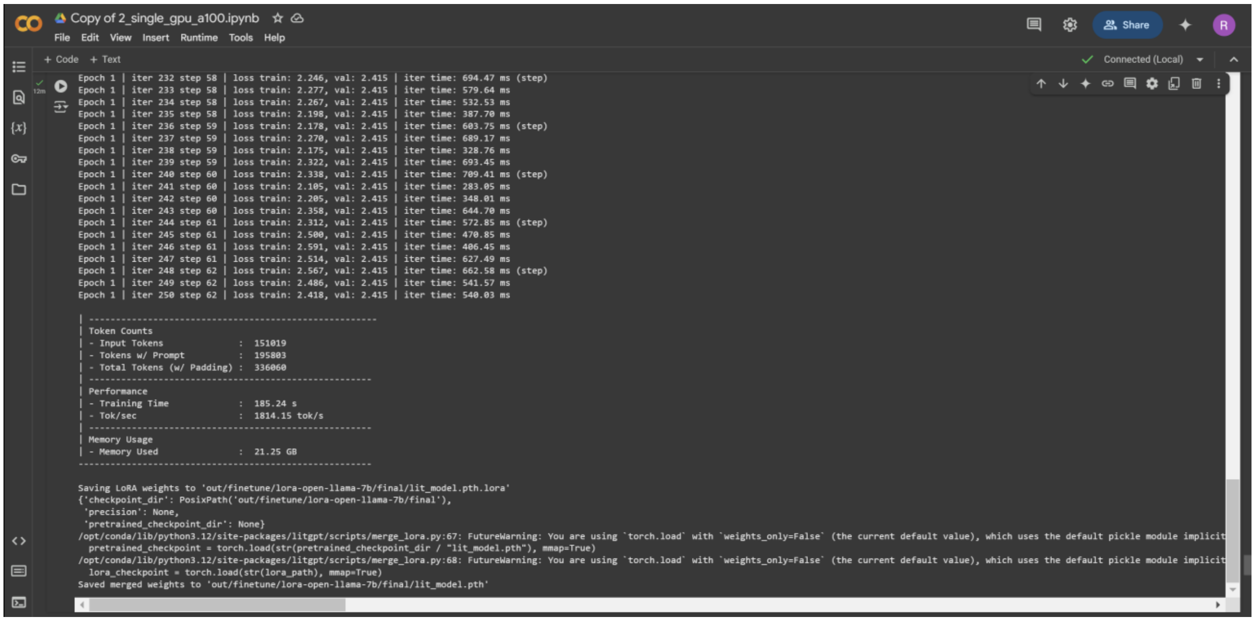
```
Epoch 1 | iter 237 step 59 | loss train: 0.874, val: 0.980 | iter time: 562.72 ms
Epoch 1 | iter 238 step 59 | loss train: 0.797, val: 0.980 | iter time: 203.91 ms
Epoch 1 | iter 239 step 59 | loss train: 0.869, val: 0.980 | iter time: 564.02 ms
Epoch 1 | iter 240 step 60 | loss train: 0.803, val: 0.980 | iter time: 559.14 ms (step)
Epoch 1 | iter 241 step 60 | loss train: 0.798, val: 0.980 | iter time: 159.97 ms
Epoch 1 | iter 242 step 60 | loss train: 0.883, val: 0.980 | iter time: 223.34 ms
Epoch 1 | iter 243 step 60 | loss train: 0.979, val: 0.980 | iter time: 516.84 ms
Epoch 1 | iter 244 step 61 | loss train: 1.002, val: 0.980 | iter time: 422.58 ms (step)
Epoch 1 | iter 245 step 61 | loss train: 1.062, val: 0.980 | iter time: 344.84 ms
Epoch 1 | iter 246 step 61 | loss train: 1.082, val: 0.980 | iter time: 282.99 ms
Epoch 1 | iter 247 step 61 | loss train: 0.941, val: 0.980 | iter time: 498.99 ms
Epoch 1 | iter 248 step 62 | loss train: 0.968, val: 0.980 | iter time: 510.56 ms (step)
Epoch 1 | iter 249 step 62 | loss train: 0.908, val: 0.980 | iter time: 416.16 ms
Epoch 1 | iter 250 step 62 | loss train: 0.850, val: 0.980 | iter time: 414.82 ms

| ----------------------------------------------------
| Token Counts
| - Input Tokens              :  151019
| - Tokens w/ Prompt          :  195803
| - Total Tokens (w/ Padding) :  336060
| ---------------------------------------------------
| Performance
| - Training Time             :  136.61 s
| - Tok/sec                   :  2459.92 tok/s
| ---------------------------------------------------
| Memory Usage
| - Memory Used               :  29.99 GB
----------------------------------------------------

Saving LoRA weights to 'out/finetune/lora-open-llama-7b/final/lit_model.pth.lora'
{'checkpoint_dir': PosixPath('out/finetune/lora-open-llama-7b/final'),
 'precision': None,
 'pretrained_checkpoint_dir': None}
/opt/conda/lib/python3.12/site-packages/litgpt/scripts/merge_lora.py:67: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicit
  pretrained_checkpoint = torch.load(str(pretrained_checkpoint_dir / "lit_model.pth"), mmap=True)
/opt/conda/lib/python3.12/site-packages/litgpt/scripts/merge_lora.py:68: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicit
  lora_checkpoint = torch.load(str(lora_path), mmap=True)
```

We can also further reduce the memory required with quantization:

7B model

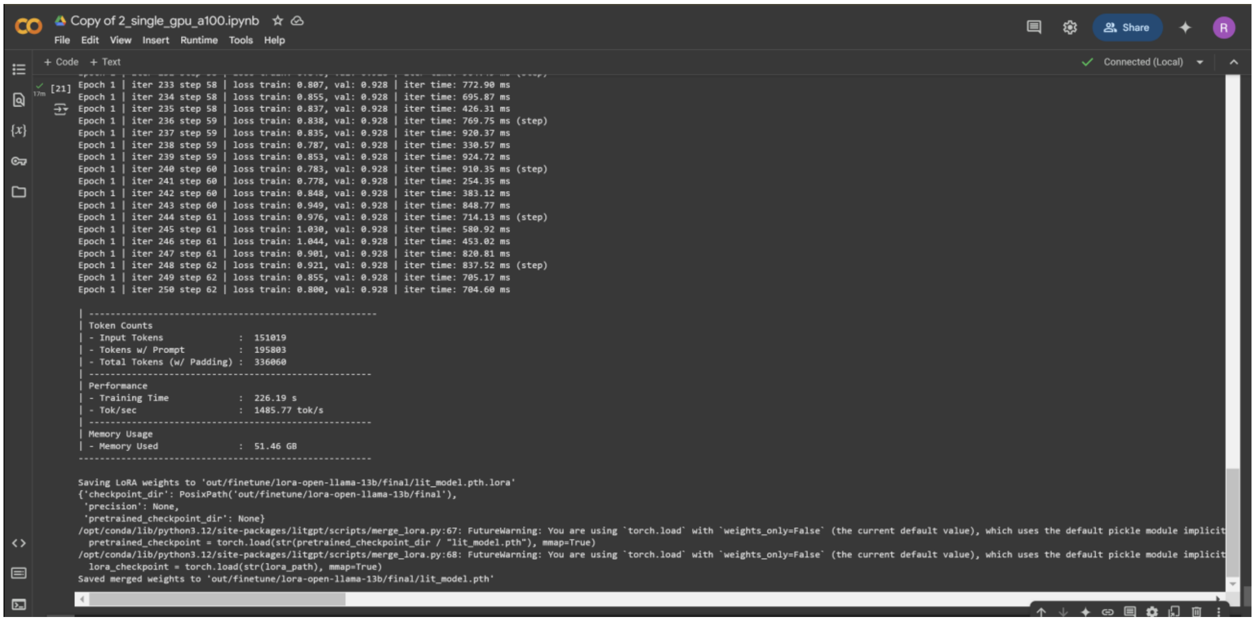


```
Epoch 1 | iter 232 step 58 | loss train: 2.246, val: 2.415 | iter time: 694.47 ms (step)
Epoch 1 | iter 233 step 58 | loss train: 2.277, val: 2.415 | iter time: 579.64 ms
Epoch 1 | iter 234 step 58 | loss train: 2.267, val: 2.415 | iter time: 532.53 ms
Epoch 1 | iter 235 step 58 | loss train: 2.198, val: 2.415 | iter time: 387.70 ms
Epoch 1 | iter 236 step 59 | loss train: 2.178, val: 2.415 | iter time: 603.75 ms (step)
Epoch 1 | iter 237 step 59 | loss train: 2.270, val: 2.415 | iter time: 689.17 ms
Epoch 1 | iter 238 step 59 | loss train: 2.175, val: 2.415 | iter time: 328.76 ms
Epoch 1 | iter 239 step 59 | loss train: 2.322, val: 2.415 | iter time: 693.45 ms
Epoch 1 | iter 240 step 60 | loss train: 2.338, val: 2.415 | iter time: 709.41 ms (step)
Epoch 1 | iter 241 step 60 | loss train: 2.105, val: 2.415 | iter time: 283.05 ms
Epoch 1 | iter 242 step 60 | loss train: 2.205, val: 2.415 | iter time: 348.01 ms
Epoch 1 | iter 243 step 60 | loss train: 2.358, val: 2.415 | iter time: 644.70 ms
Epoch 1 | iter 244 step 61 | loss train: 2.312, val: 2.415 | iter time: 572.85 ms (step)
Epoch 1 | iter 245 step 61 | loss train: 2.500, val: 2.415 | iter time: 470.85 ms
Epoch 1 | iter 246 step 61 | loss train: 2.591, val: 2.415 | iter time: 406.45 ms
Epoch 1 | iter 247 step 61 | loss train: 2.514, val: 2.415 | iter time: 627.49 ms
Epoch 1 | iter 248 step 62 | loss train: 2.567, val: 2.415 | iter time: 662.58 ms (step)
Epoch 1 | iter 249 step 62 | loss train: 2.486, val: 2.415 | iter time: 541.57 ms
Epoch 1 | iter 250 step 62 | loss train: 2.418, val: 2.415 | iter time: 540.03 ms

| ----------------------------------------------------
| Token Counts
| - Input Tokens              :  151019
| - Tokens w/ Prompt          :  195803
| - Total Tokens (w/ Padding) :  336060
| ---------------------------------------------------
| Performance
| - Training Time             :  185.24 s
| - Tok/sec                   :  1814.15 tok/s
| ---------------------------------------------------
| Memory Usage
| - Memory Used               :  21.25 GB
----------------------------------------------------

Saving LoRA weights to 'out/finetune/lora-open-llama-7b/final/lit_model.pth.lora'
{'checkpoint_dir': PosixPath('out/finetune/lora-open-llama-7b/final'),
 'precision': None,
 'pretrained_checkpoint_dir': None}
/opt/conda/lib/python3.12/site-packages/litgpt/scripts/merge_lora.py:67: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicit
  pretrained_checkpoint = torch.load(str(pretrained_checkpoint_dir / "lit_model.pth"), mmap=True)
/opt/conda/lib/python3.12/site-packages/litgpt/scripts/merge_lora.py:68: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicit
  lora_checkpoint = torch.load(str(lora_path), mmap=True)
Saved merged weights to 'out/finetune/lora-open-llama-7b/final/lit_model.pth'
```

7B model with quantization


Copy of 2_single_gpu_a100.ipynb
File Edit View Insert Runtime Tools Help
Share
+ Code + Text
Connected (Local)
[21]
17m

```
Epoch 1 | iter 233 step 58 | loss train: 0.807, val: 0.928 | iter time: 772.90 ms
Epoch 1 | iter 234 step 58 | loss train: 0.855, val: 0.928 | iter time: 695.87 ms
Epoch 1 | iter 235 step 58 | loss train: 0.837, val: 0.928 | iter time: 426.31 ms
Epoch 1 | iter 236 step 59 | loss train: 0.838, val: 0.928 | iter time: 769.75 ms (step)
Epoch 1 | iter 237 step 59 | loss train: 0.835, val: 0.928 | iter time: 920.37 ms
Epoch 1 | iter 238 step 59 | loss train: 0.787, val: 0.928 | iter time: 330.57 ms
Epoch 1 | iter 239 step 59 | loss train: 0.853, val: 0.928 | iter time: 924.72 ms
Epoch 1 | iter 240 step 60 | loss train: 0.783, val: 0.928 | iter time: 910.35 ms (step)
Epoch 1 | iter 241 step 60 | loss train: 0.778, val: 0.928 | iter time: 254.35 ms
Epoch 1 | iter 242 step 60 | loss train: 0.848, val: 0.928 | iter time: 383.12 ms
Epoch 1 | iter 243 step 60 | loss train: 0.949, val: 0.928 | iter time: 848.77 ms
Epoch 1 | iter 244 step 61 | loss train: 0.976, val: 0.928 | iter time: 714.13 ms (step)
Epoch 1 | iter 245 step 61 | loss train: 1.030, val: 0.928 | iter time: 580.92 ms
Epoch 1 | iter 246 step 61 | loss train: 1.044, val: 0.928 | iter time: 453.02 ms
Epoch 1 | iter 247 step 61 | loss train: 0.901, val: 0.928 | iter time: 820.81 ms
Epoch 1 | iter 248 step 62 | loss train: 0.921, val: 0.928 | iter time: 837.52 ms (step)
Epoch 1 | iter 249 step 62 | loss train: 0.855, val: 0.928 | iter time: 705.17 ms
Epoch 1 | iter 250 step 62 | loss train: 0.800, val: 0.928 | iter time: 704.60 ms

| ------------------------------------------------------
| Token Counts
| - Input Tokens              :  151019
| - Tokens w/ Prompt          :  195803
| - Total Tokens (w/ Padding) :  336060
| -----------------------------------------------------
| Performance
| - Training Time             :  226.19 s
| - Tok/sec                   :  1485.77 tok/s
| -----------------------------------------------------
| Memory Usage
| - Memory Used               :  51.46 GB
-------------------------------------------------------

Saving LoRA weights to 'out/finetune/lora-open-llama-13b/final/lit_model.pth.lora'
{'checkpoint_dir': PosixPath('out/finetune/lora-open-llama-13b/final'),
 'precision': None,
 'pretrained_checkpoint_dir': None}
/opt/conda/lib/python3.12/site-packages/litgpt/scripts/merge_lora.py:67: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicit
  pretrained_checkpoint = torch.load(str(pretrained_checkpoint_dir / "lit_model.pth"), mmap=True)
/opt/conda/lib/python3.12/site-packages/litgpt/scripts/merge_lora.py:68: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicit
  lora_checkpoint = torch.load(str(lora_path), mmap=True)
Saved merged weights to 'out/finetune/lora-open-llama-13b/final/lit_model.pth'
```

13B model