Training optimization

Optimizing GPT-124M pretraining function for Google Colab A100 environment

1. Data loading
    a. using Streaming Datasets
    b. multiple workers
2. Optimizing training function
    a. enabling tensor cores
    b. using mixed precision and gradient scaling
    c. fuse operations - compile model
    d. avoid synchronizing CPU-GPU (print(), item())
3. Optimizing model
    a. adjusting layer sizes for training on tensor cores
    b. using flash attention
    c. weight tying (not an optimization exactly)

Baseline

GPU 37/40 GB (92.5%); batch size: 28; cycle: 10 + 1 batches

```
batches:11 | loss: 8.836/7.979| tok-seen: 157,696| time: 24.71| tok/sec: 6,381.04 | epoch time: 158.39 hrs
batches:22 | loss: 7.431/7.307| tok-seen: 315,392| time: 9.31| tok/sec: 16,947.16 | epoch time: 59.64 hrs
batches:33 | loss: 7.065/8.288| tok-seen: 473,088| time: 9.30| tok/sec: 16,952.96 | epoch time: 59.62 hrs
batches:44 | loss: 6.862/7.065| tok-seen: 630,784| time: 9.30| tok/sec: 16,964.41 | epoch time: 59.58 hrs
batches:55 | loss: 7.227/7.746| tok-seen: 788,480| time: 9.30| tok/sec: 16,964.17 | epoch time: 59.58 hrs
...
batches:319 | loss: 7.110/6.427| tok-seen: 4,573,184| time: 9.31| tok/sec: 16,939.79 | epoch time: 59.66 hrs
Tokens seen (train/val/total) 4,186,112 / 415,744 / 4,601,856
Total books seen 79
Saving checkpoint...79
Training completed in 4.84 minutes.
```

Enable Tensor Cores

https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf

```
torch.set_float_32_matmul_precision('high')
```

GPU 37/40 GB (92.5%); batch size: 28; cycle: 10 + 1 batches

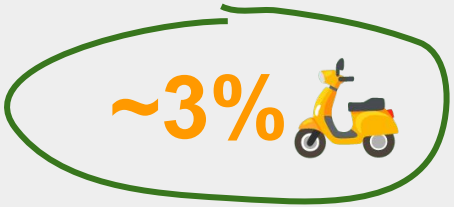
```
batches:11 | loss: 8.844/7.891| tok-seen: 157,696| time: 21.09| tok/sec: 7,477.15 | epoch time: 135.17 hrs
batches:22 | loss: 7.421/7.291| tok-seen: 315,392| time: 3.86| tok/sec: 40,848.71 | epoch time: 24.74 hrs
batches:33 | loss: 7.066/8.206| tok-seen: 473,088| time: 3.85| tok/sec: 40,975.57 | epoch time: 24.67 hrs
batches:44 | loss: 6.859/7.090| tok-seen: 630,784| time: 3.85| tok/sec: 40,992.17 | epoch time: 24.66 hrs
...

batches:748 | loss: 6.029/6.412| tok-seen: 10,723,328| time: 3.85| tok/sec: 40,909.33 | epoch time: 24.71 hrs
batches:759 | loss: 6.087/6.043| tok-seen: 10,881,024| time: 3.87| tok/sec: 40,753.20 | epoch time: 24.80 hrs
Tokens seen (train/val/total) 9,992,192 / 989,184 / 10,981,376
Total books seen 189
Saving checkpoint...189
Training completed in 4.83 minutes.
```

Avoid synchronizing CPU and GPU

```
print(cuda_tensor) | cuda_tensor.item() | if (cuda_tensor != 0).all() | cuda_tensor.nonzero()
```

GPU 37/40 GB (92.5%); batch size: 28; cycle: 200 + 20 batches

```
batches:220 | loss: 6.963/6.920| tok-seen: 3,153,920| time: 85.13| tok/sec: 37,047.10 | epoch time: 27.28 hrs
batches:440 | loss: 6.212/6.421| tok-seen: 6,307,840| time: 76.89| tok/sec: 41,019.40 | epoch time: 24.64 hrs
batches:660 | loss: 6.473/6.590| tok-seen: 9,461,760| time: 76.90| tok/sec: 41,012.20 | epoch time: 24.64 hrs
Tokens seen (train/val/total) 10,422,272 / 860,160 / 11,282,432
Total books seen 127
Saving checkpoint...127
Training completed in 4.82 minutes.
```



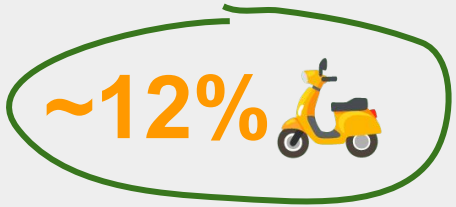
Mixed Precision - torch.amp + GradScaler

(use `dtype=torch.bfloat16` to avoid using GradScaler ?)

1st run: Slight speedup + GPU freed up 32.4/40 GB (81%) | batch size: 28

GPU 36.9/40 (92.25%) | batch_size: 32 | cycle: 200 + 20 batches

```
batches:220 | loss: 6.966/6.849| tok-seen: 3,604,480| time: 91.93| tok/sec: 39,209.78 | epoch: 25.78 hrs
batches:440 | loss: 6.409/6.432| tok-seen: 7,208,960| time: 82.33| tok/sec: 43,783.06 | epoch: 23.08 hrs
batches:660 | loss: 6.398/6.565| tok-seen: 10,813,440| time: 82.63| tok/sec: 43,623.05 | epoch: 23.17 hrs
Tokens seen: 11,665,408 / 983,040 / 12,648,448
Total books seen 140
Saving checkpoint...140
Training completed in 5.08 minutes.
```



Fuse operations with torch.compile()

1st run: Substantial speedup (~22M tok.seen/5min) + memory freed up GPU 25.4/40 GB (63.5%) | batch size: 32

2nd run: GPU 37.4/40 (93.5%) | batch_size: 48 | cycle: 200 + 20 batches

```
batches:220 | loss: 6.995/7.635| tok-seen: 5,406,720| time: 112.89| tok/sec: 47,894.34 | epoch: 21.10 hrs
batches:440 | loss: 6.611/6.483| tok-seen: 10,813,440| time: 48.29| tok/sec: 111,974.28 | epoch: 9.03 hrs
batches:660 | loss: 6.222/6.445| tok-seen: 16,220,160| time: 48.10| tok/sec: 112,412.16 | epoch: 8.99 hrs
batches:880 | loss: 6.092/6.306| tok-seen: 21,626,880| time: 48.22| tok/sec: 112,136.19 | epoch: 9.01 hrs
Tokens seen: 22,683,648 / 1,966,080 / 24,649,728
Total books seen 267
Saving checkpoint...267
Training completed in 4.83 minutes.
```

Flash attention

1st run: Substantial speedup (~30M tok.seen) + memory freed up GPU 29.6 / 40.0 GB (%) | batch size: 48

2nd run: GPU 38.9/40 (97.25%) | batch_size: 64 | cycle: 200 + 20 batches

```
batches:220 | loss: 7.042/6.927| tok-seen: 7,208,960| time: 118.08| tok/sec: 61,051.77 | epoch: 16.55 hrs

batches:440 | loss: 6.528/6.262| tok-seen: 14,417,920| time: 54.49| tok/sec: 132,292.42 | epoch: 7.64 hrs

batches:660 | loss: 6.188/6.297| tok-seen: 21,626,880| time: 54.57| tok/sec: 132,110.35 | epoch: 7.65 hrs

batches:880 | loss: 6.277/6.022| tok-seen: 28,835,840| time: 54.52| tok/sec: 132,226.77 | epoch: 7.64 hrs

Saving model after KeyboardInterrupt

Tokens seen: 27,262,976 / 2,621,440 / 29,884,416

Total books seen 304

Training completed in 4.98 minutes.
```

Adjusting Vocabulary Size (50,257 to 50,304)

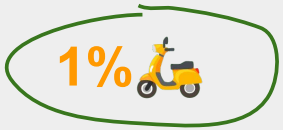
GPU **38.9/40 (97.25%)** | batch_size: 64 | cycle: 200 + 20 batches

```
batches:220 | loss: 7.041/6.790| tok-seen: 7,208,960| time: 107.98| tok/sec: 66,759.81 | epoch: 15.14 hrs

batches:440 | loss: 6.532/6.285| tok-seen: 14,417,920| time: 48.81| tok/sec: 147,693.92 | epoch: 6.84 hrs

batches:660 | loss: 6.195/6.295| tok-seen: 21,626,880| time: 48.92| tok/sec: 147,349.01 | epoch: 6.86 hrs

batches:880 | loss: 6.276/6.032| tok-seen: 28,835,840| time: 48.92| tok/sec: 147,366.79 | epoch: 6.86 hrs

Saving model after KeyboardInterrupt

Tokens seen: 31,391,744 / 2,621,440 / 34,013,184

Total books seen 337

Training completed in 4.87 minutes.
```

Tied weights in GPT model

GPU **38.7/40 (96.75%)** | batch_size: 64 | cycle: 200 + 20 batches

```
batches:220 | loss: 36.890/14.536| tok-seen: 7,208,960| time: 108.44| tok/sec: 66,477.84 | epoch: 15.20 hrs

batches:440 | loss: 12.458/9.232| tok-seen: 14,417,920| time: 48.27| tok/sec: 149,332.75 | epoch: 6.77 hrs

batches:660 | loss: 9.676/8.000| tok-seen: 21,626,880| time: 48.52| tok/sec: 148,572.49 | epoch: 6.80 hrs

batches:880 | loss: 9.132/7.153| tok-seen: 28,835,840| time: 48.50| tok/sec: 148,637.53 | epoch: 6.80 hrs

Saving model after KeyboardInterrupt

Tokens seen: 31,719,424 / 2,621,440 / 34,340,864

Total books seen 344

Training completed in 4.87 minutes.
```



Workers in DataLoader

GPU 38.7/40 (96.75%) | batch_size: 64 | cycle: 200 + 20 batches

```
batches:220 | loss: 32.894/13.998| tok-seen: 7,208,960| time: 110.54| tok/sec: 65,214.56 | epoch: 15.50 hrs
batches:440 | loss: 12.910/8.938| tok-seen: 14,417,920| time: 49.25| tok/sec: 146,384.86 | epoch: 6.90 hrs
batches:660 | loss: 9.607/7.965| tok-seen: 21,626,880| time: 48.88| tok/sec: 147,477.63 | epoch: 6.85 hrs
batches:880 | loss: 8.569/7.328| tok-seen: 28,835,840| time: 49.12| tok/sec: 146,766.49 | epoch: 6.89 hrs
Tokens seen: 30,277,632 / 2,621,440 / 32,899,072
Total books seen 350
Saving checkpoint...350
Training completed in 4.78 minutes.
```

?

Total improvement compared to the baseline

LINKS

PyTorch Performance Tuning Guide
https://docs.pytorch.org/tutorials/recipes/recipes/tuning_guide.html
Nvidia optimizing performance: https://docs.nvidia.com/deeplearning/performance/index.html#optimizing-performance
Transformers Case Study: https://docs.nvidia.com/deeplearning/performance/dl-performance-fully-connected/index.html#case-studies

IterableDataset
PyTorch: https://docs.pytorch.org/docs/stable/data.html
Huggingface: https://huggingface.co/docs/datasets/v4.0.0/en/about_mapstyle_vs_iterable

Tensor Cores:
PyTorch: https://docs.pytorch.org/docs/stable/notes/cuda.html#tensorfloat-32-tf32-on-ampere-and-later-devices
A100 Nvidia White Paper:
https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf

Mixed Precision:
Recommended by Karpathy: https://docs.pytorch.org/tutorials/recipes/recipes/amp_recipe.html
but there is the same thing here: https://docs.pytorch.org/docs/stable/amp.html

Flash Attention:
https://arxiv.org/pdf/2205.14135
Online normalizer calc. for softmax: https://arxiv.org/pdf/1805.02867