| Name | Report Link | Description | Best result |
|---|---|---|---|
| H/14 | wandb logs | **Main result**: H/14 with big batch size works well, but unstable, and very hard to recover<br><br>Planned for 256 * 135M epochs from 2B-en<br>Spike at epoch 122. Tried a lot of stuff to recover<br>Only one thing worked: decreasing lr fast for 8 epoch, got 74% that way<br>Batch size 79k, starting lr 5e-4<br>1 week to train + many days to try to figure it out<br>800gpus<br>Doing 8 epoch with batch size 158k gave 75.4%<br><br>Finished up to 256 at batch size 79k in bfloat16 and reached 78.0% | 78.0 |
| B/32 | wandb logs | **Main result**: Environment is working fully, only ml problem<br><br>256 * 135M epochs from 2B-en<br>Starting lr 1e-3<br>Batch size 79k<br>Run till the end. Best B/32 thus far<br>60h to train<br>800 gpus | 66.5 |
| Fast g/14 | wandb logs | **Main result**: we can train a decent clip fast by decreasing lr fast<br><br>12 * 135M epoch from 2B-en<br>Starting lr 1e-3, went to 0 in 4 epochs. Batch size 79k<br>Restarted 2 times by increasing epoch count, hence increasing cosine lr schedule<br>Reached 69% fast<br>20h to train<br>800 gpus | 68.9 |
| Slow g/14 | wandb logs | **Main result:** bfloat16 is faster and more stable, solves instabilities<br><br>idea: high batch size is not stable, trying smaller one<br><br>90 * 135M epoch from 2B-en, around 12B samples seen which is about 32 times 400M, like initial clip sample seen<br>Starting lr 5e-4 batch size 32k<br>Should take about 8 days to train<br>1600 gpus<br>-> exploded at epoch 59<br><br>Then tried to continue with bfloat16 -> finished the 90 epochs<br>With 800 GPUs | 76.6 |

160k batch size small run exploded fast with float16