OPT Baselines Logbook

- All baselines created using <redacted>.
- Occasionally hyperparameters were set differently for one model.
- All use the same data as the 175B model

Training Log

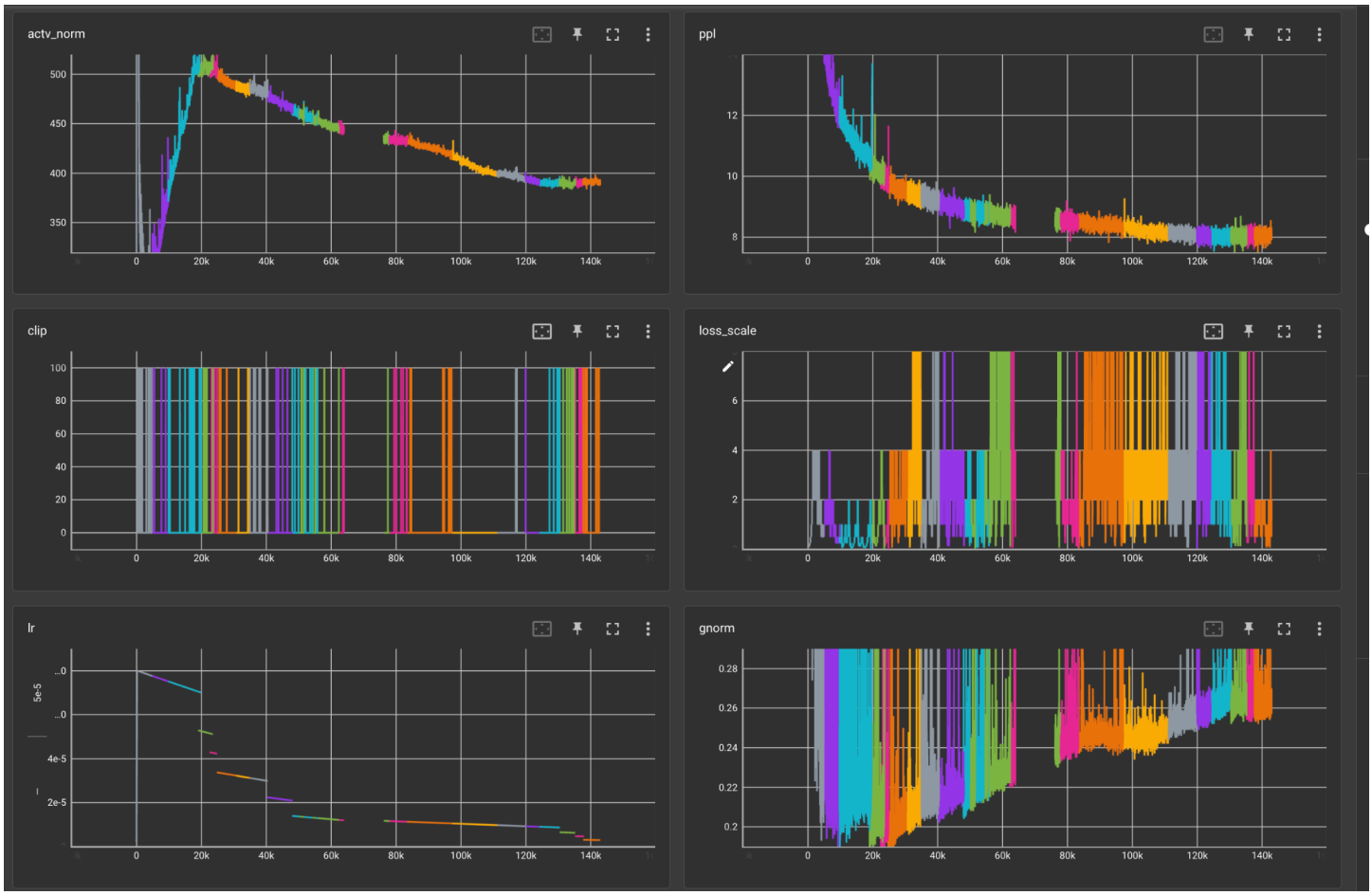
2022-05-17 [Susan] WE ARE DONE!!!

- One more pdsh run to push on the checkpoints to blob. Hooray!

Copying blobs over to final blob path:

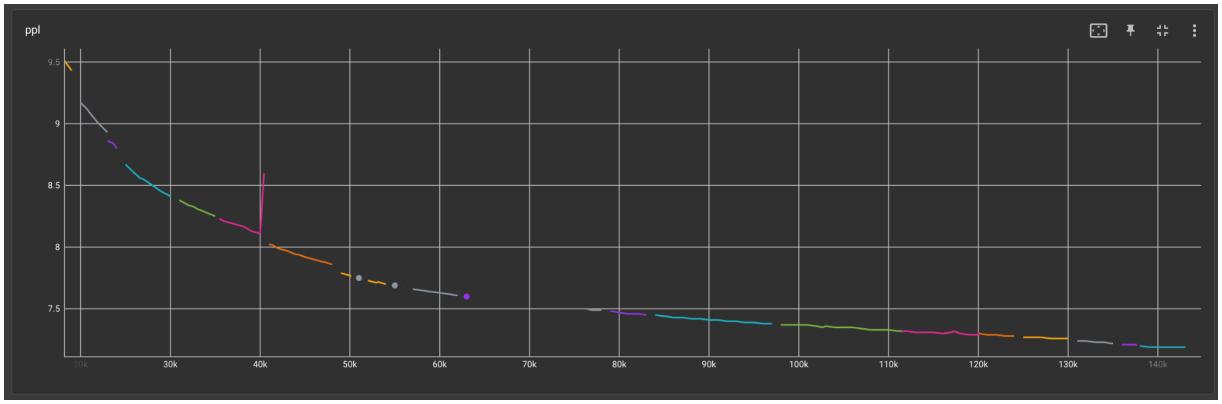
azcopy copy "<redacted>" <redacted>" --include-pattern "checkpoint_49_143000*.pt" --recursive

Final view of run16-43 (missing run36 from home dir issue and no backups):



Tensorboard

- Takeaways:
    - 66B seemed more difficult to train with 2M batch size than 175B with 2M batch size
        - BF16 likely would work better here (but doesn't explain the 66B vs 175B instability)
    - LR may have been too low for too long, despite validation on wikipedia_en continuing to "improve"



Tensorboard with wikipedia_en validation ppl

- Adjusting dynamic loss scaling window (to be a function of loss scalar value) seemed to have helped with stability, but won't be needed at all when switching to bf16

2022-05-15 [Susan] Recover failed uploads, restart with LR at 6e-6

PDSH_RCMD_TYPE=ssh pdsh -w hpc-pg0-[9-12,14-31,34-43,45-48,50-69,71-78] '/shared/home/susanz/bin/azcopy copy "<redacted>/*.pt" "<redacted>"'

- Kicked off run 43

```
BLOB_PREFIX1="<redacted>/66B_run42"
BLOB_PREFIX2=<redacted>/66B_run43"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_48_137750.pt?${BLOB_AUTH}"
RUN_ID=66B_run43
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

- Logging hostlists given continued azcopy failures:

```
30279       hpc 66B_run4   susanz  R    8:42:52     64 hpc-pg0-[9-12,14-31,34-40,42-43,45-48,50-69,71-78,84]
```

2022-05-15 [Susan] Nan grads, lowered LR to 9e-6

- End LR is kept at 0.5*LR = 4.5e-6

```
BLOB_PREFIX1="<redacted>/66B_run40"
BLOB_PREFIX2="<redacted>/66B_run41"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_47_135250.pt?${BLOB_AUTH}"
RUN_ID=66B_run41
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
```

--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE

- Node failures on host 13, did not auto-recover. Manually drained, and resuming again from the same point given intermittent azcopy failures too.
    - Added “args.requeue_on_fail = True” to sweep (wasn’t there before).
- Relaunched exactly the same as run 41, just incremented to run 42.
- Azcopy failures persist, tracking hostlist to recover from:

30152 hpc 66B_run4 susanz R 6:49:09 64 hpc-pg0-[9-12,14-31,34-43,45-48,50-69,71-78]

2022-05-13 [Susan] 65 hosts in drain for IB issues, lowered LR to 1.2e-5

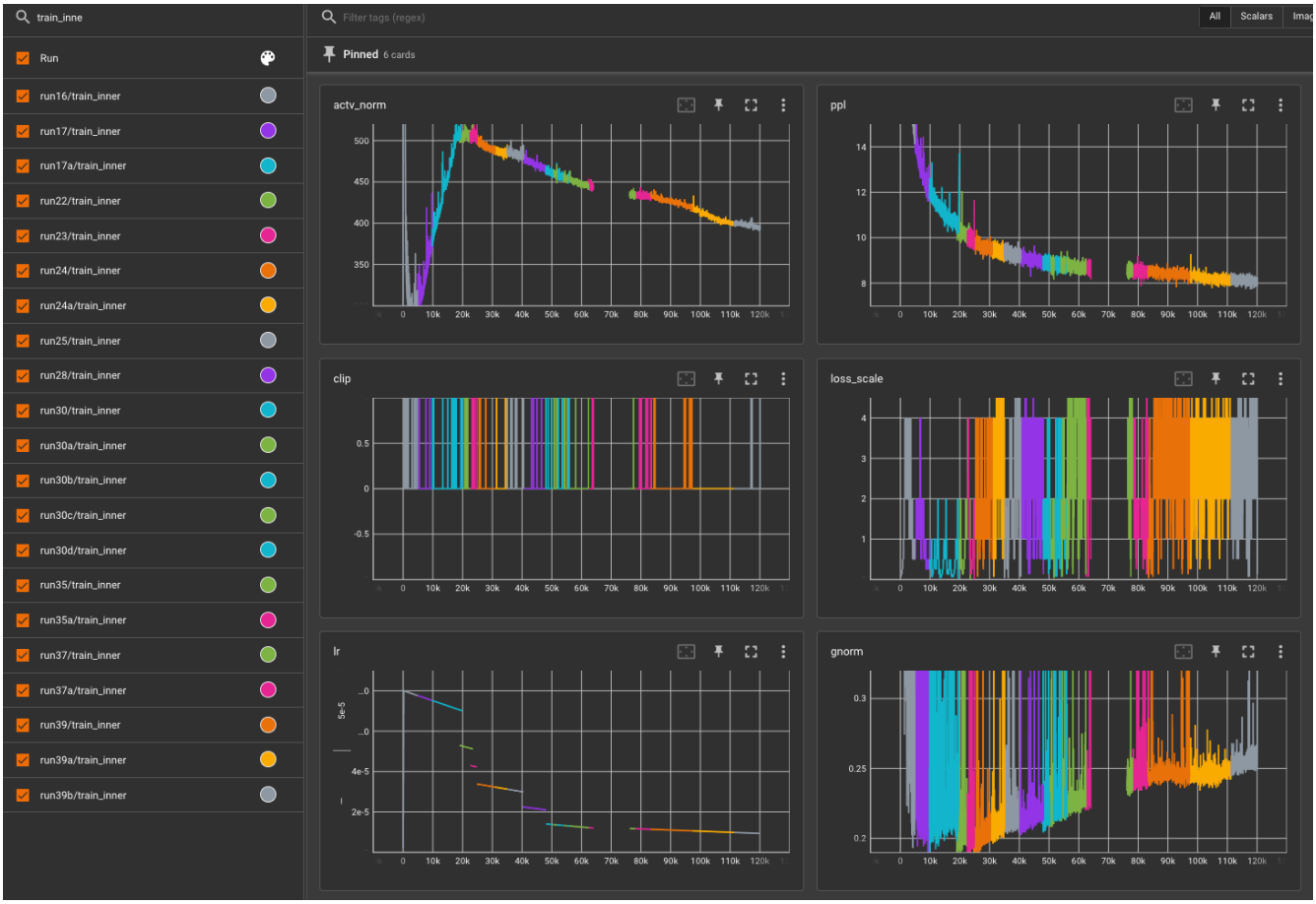
- End LR is kept at 0.5*LR = 6e-6

BLOB_PREFIX1="<redacted>66B_run39"
BLOB_PREFIX2=<redacted>/66B_run40"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_45_130500.pt?${BLOB_AUTH}"
RUN_ID=66B_run40
EXCLUDED_HOSTS=<redacted> \
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE

2022-05-11 [Susan] Nans in grad, lost GPU on node 32

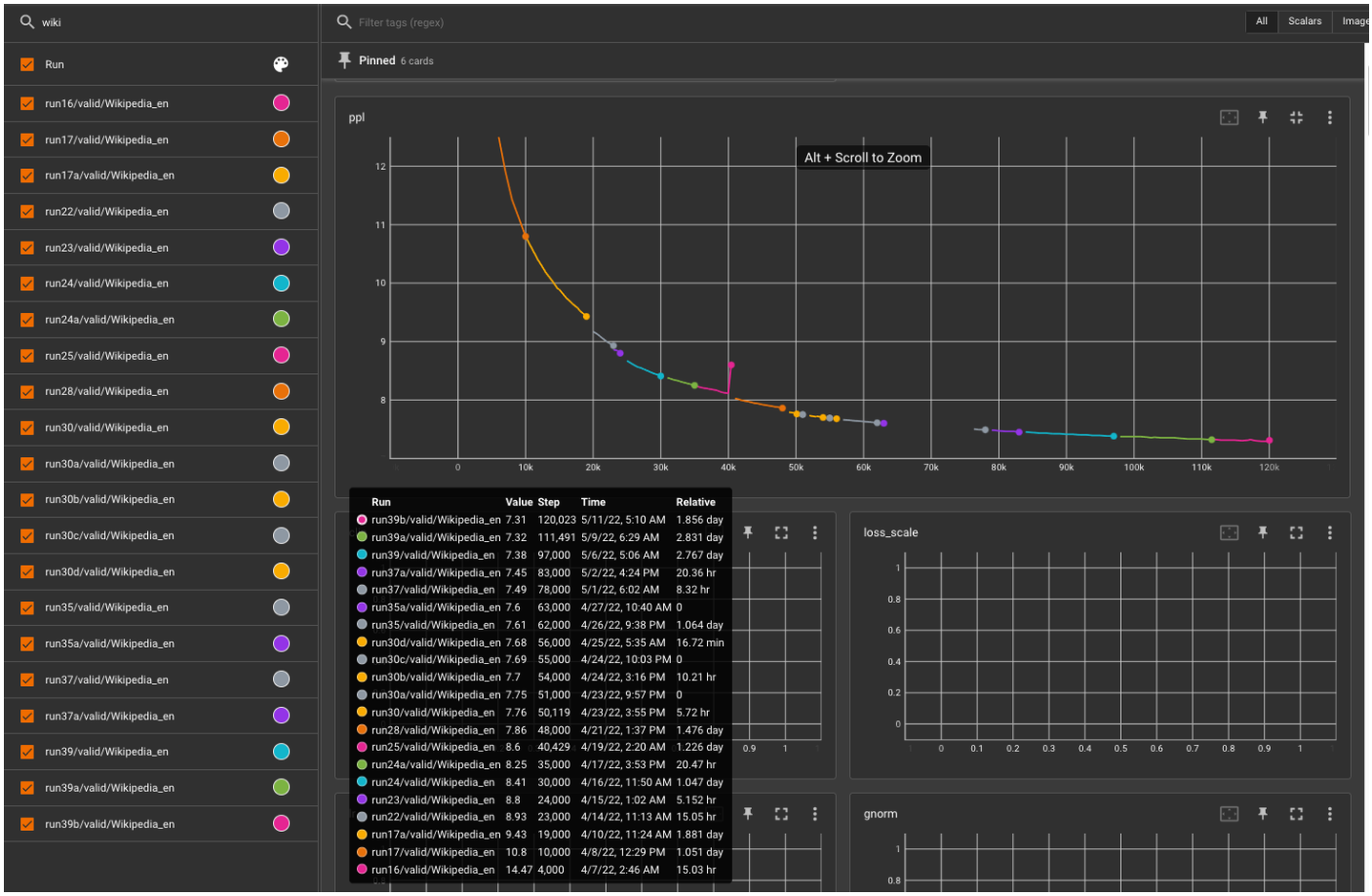
2022-05-11 11:22:12 | INFO | fairseq.trainer | NOTE: floating point error detected, ignoring gradient, Fatal error: gradients are inconsistent between workers. Try --ddp-backend=legacy_ddp. Or are you mixing up different generation of GPUs in training?

- fixmyazure drained node 32 for lost GPU error.
- Overall things look pretty healthy for now. Missing chunk of logs in the middle from when the clusters’ home directory got swamped / wiped by another team’s usage. Logs are now getting backed up periodically to cloud.



Tensorboard

- Validation ppl on wikipedia looks good too:



Tensorboard

2022-05-07 [Susan] Model is finally chugging along

- No more learning rate changes since last entry.
- A few restarts from hardware failures.
- We are about ~73% through.
- ETA for completion: ~9 days
- Spot-checked validation ppls: seems to still be improving, though more slowly now (LR is at 1e-5).

2022-04-27 [Susan] Lowering LR to 1.6e-5, restart @ 63,750

- Also increase end LR to be 0.5 * start LR

```
BLOB_PREFIX1="<redacted>/66B_run35"
BLOB_PREFIX2="<redacted>/66B_run36"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_22_63750.pt?${BLOB_AUTH}"
RUN_ID=66B_run36
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

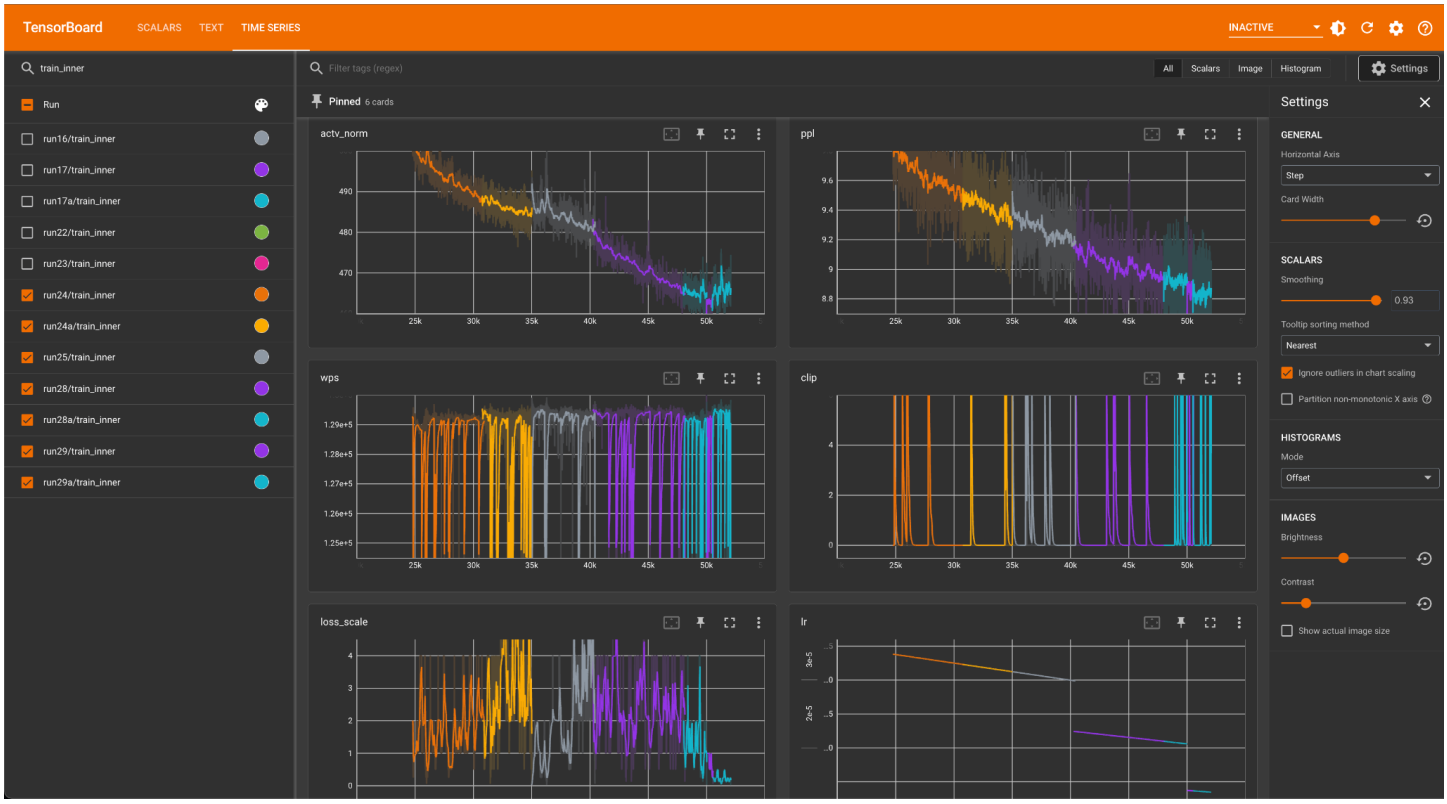
2022-04-25 [Susan] Nan/Inf grads @ 56029, skipping batches from swallowing FloatingPointError

- Next restart should increase end LR - maybe keep at original end LR (6e-6), or leave flat / try increasing end LR to warm up to a point where things crash
- Validation ppl on Wikipedia is still dropping, though now only 0.01 every 1k steps.
- Consistent crashing at 56029 step.
- Tried:
    - 1.6e-5 start lr, 8e-6 end lr, no good.
    - 1.2e-6 start lr, 6e-6 end lr, no good.
- Trying start lr of 1e-6 with end lr at 1e-5 (higher end LR than start, to try and “warm up LR” again).
    - Didn’t work.
- Tried swapping shard 20 and 29 (experiment 32), didn’t work.
    - Used: <redacted> as new data dir
- Drastically cutting LR (since the shard swapping indicates a 0 LR would also not be useful): trying now with starting LR of 2e-6 (end LR of 1e-6, for experiment 33).
    - Didn’t work.
    - Grad norm ends up being nan. Need clipping?
- Restarting with LR back to 2e-5, but increase clip to 0.25 (experiment 34).
    - Didn’t work (didn’t clip anything).
    - Trying with clip down to 0.22 (still experiment 34).
    - Still doesn’t work, stuck at same place (56029).
    - Trying with commenting out throwing FloatingPointError (and skip the batch?) - experiment 35
    - Ok this worked. -____-

```
BLOB_PREFIX1="<redacted>/66B_run30"
BLOB_PREFIX2="<redacted>/66B_run35"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_20_56000.pt?${BLOB_AUTH}"
RUN_ID=66B_run35
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

2022-04-22 [Susan] Keep lower LR of 2e-5, resume from 48k

- Things keep crashing
- Time to lower LR from an earlier checkpoint
- (wiki) validation looks “ok” though after lower LR of 2e-5 (continues to drop), so lowering it earlier shouldn’t hurt.
- Loss scales started looking rough after 48k, hence resuming from 48k with a lower LR.



[Tensorboard]

```
BLOB_PREFIX1="<redacted>/66B_run28"
BLOB_PREFIX2="<redacted>/66B_run30"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_17_48000.pt?${BLOB_AUTH}"
RUN_ID=66B_run30
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
```

```
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

2022-04-22 [Susan] Lower LR to ~~2.4e-5~~ 2e-5, resume from 50k

```
BLOB_PREFIX1="<redacted>/66B_run28"
BLOB_PREFIX2="<redacted>/66B_run29"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_17_50000.pt?${BLOB_AUTH}"
RUN_ID=66B_run29
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

2022-04-19 [Susan]

```
BLOB_PREFIX1="<redacted>/66B_run25"
BLOB_PREFIX2="<redacted>/66B_run26"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_14_40000.pt?${BLOB_AUTH}"
RUN_ID=66B_run26
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

- Hung after crashing log scale, didn't recover.
- Increasing LR to 8e-5, restarting from 41500.
    - Nope, bad idea. Should stick with the convention of lowering LR. Down to 3e-5. Still calling this run27 since the path will be different with different LR.

```
BLOB_PREFIX1="<redacted>/66B_run26"
BLOB_PREFIX2="<redacted>/66B_run27"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_15_41500.pt?${BLOB_AUTH}"
RUN_ID=66B_run27
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
```

```
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

- Rewinding some more to 40250:

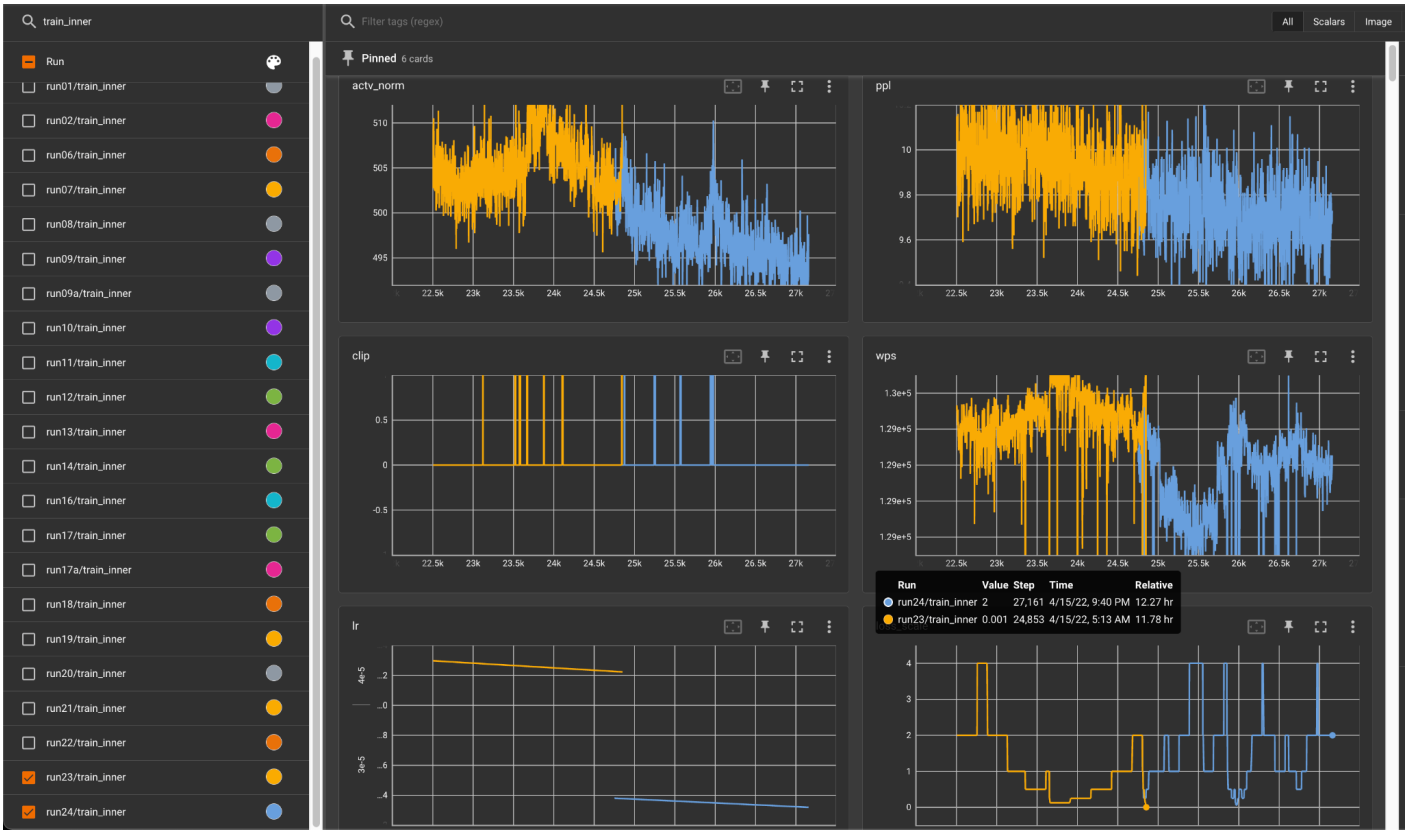
```
BLOB_PREFIX1="<redacted>/66B_run25"
BLOB_PREFIX2="<redacted>/66B_run28"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_14_40250.pt?${BLOB_AUTH}"
RUN_ID=66B_run28
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

2022-04-17 [Susan]: Gradient crashed and monitor did not kick in (given no lag in logging), restarting with no changes (outside of reducing logging)

```
BLOB_PREFIX1="<redacted>/66B_run24"
BLOB_PREFIX2="<redacted>/66B_run25"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_12_35000.pt?${BLOB_AUTH}"
RUN_ID=66B_run25
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

2022-04-15 [Susan]: Loss scale window logic change, LR 4e-5

- Change loss scale window to also scale down with loss scale, lower bounded loss scale to 0.03125 and commented out raising loss scale min threshold error (result is skipping those batches).
- Seems stable for the day, with activation norm slowly trending down:

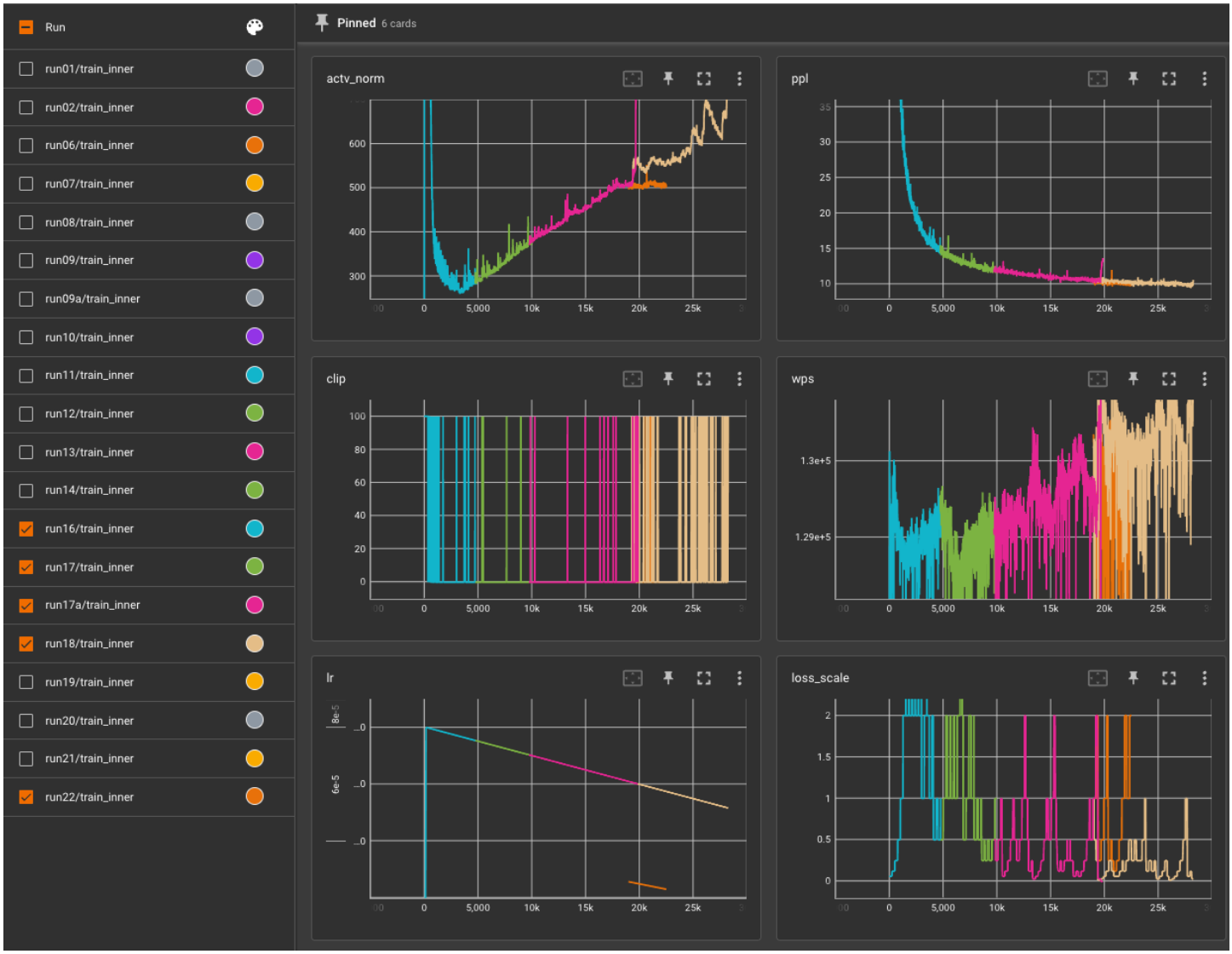


[Tensorboard]

```
BLOB_PREFIX1="<redacted>/66B_run23"
BLOB_PREFIX2="<redacted>/66B_run24"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_9_24750.pt?${BLOB_AUTH}"
RUN_ID=66B_run24
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

2022-04-14 [Susan]: Things looked good but nope.

- Activation norm curve looks more sane, loss scales look better too (orange vs pale orange before)

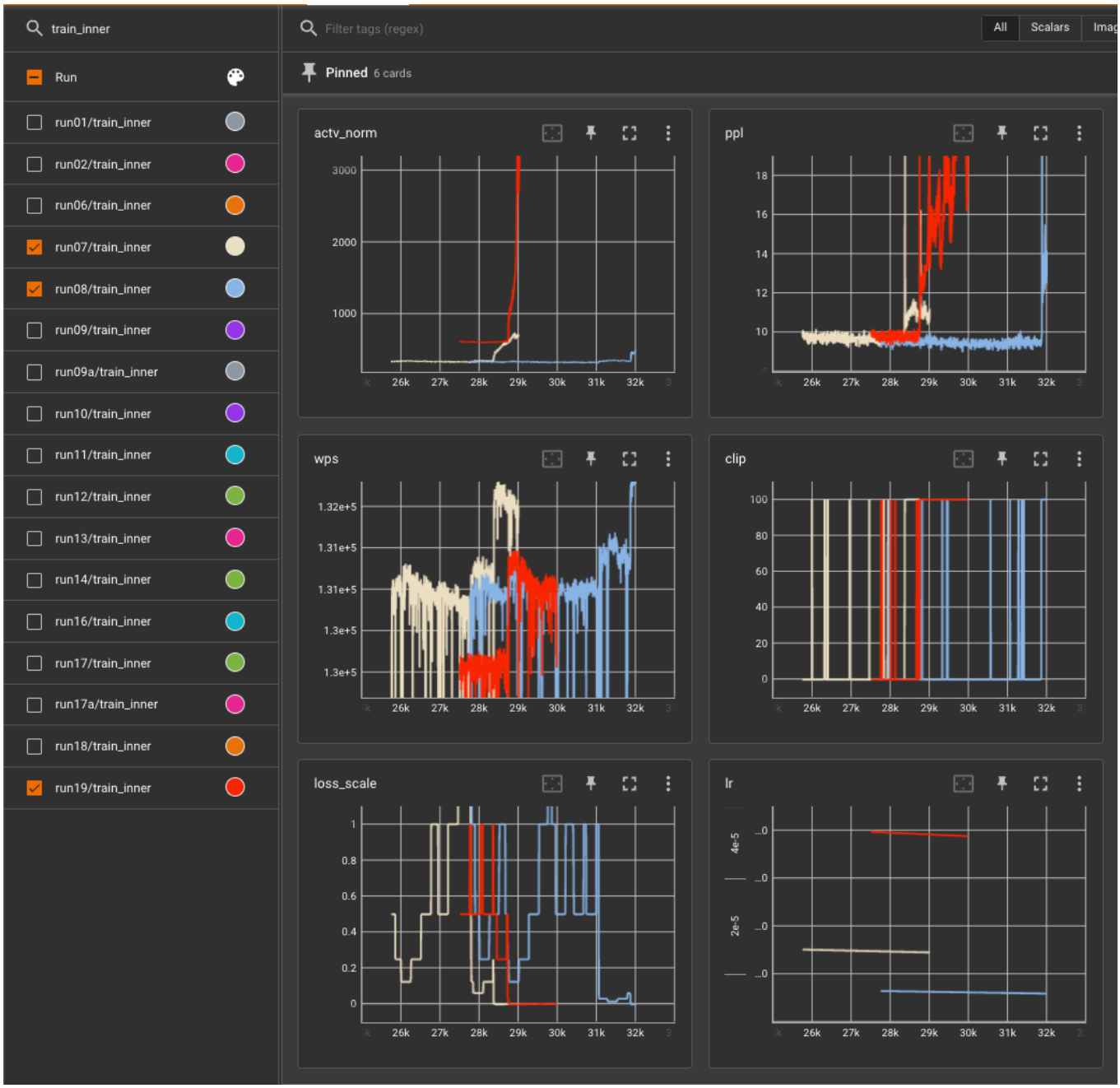


Tensorboard

- Update @ 4PM CDT: Spoke too soon - need to restart from 22.5k
    - Lowered LR to 5e-5

```
BLOB_PREFIX1="<redacted>/66B_run22"
BLOB_PREFIX2="<redacted>/66B_run23"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_8_22500?${BLOB_AUTH}"
RUN_ID=66B_run23
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

2022-04-13 [Susan]: Restart from 27.5k with lr of 4e-5, ppl diverged at ~28.7k



[Tensorboard]

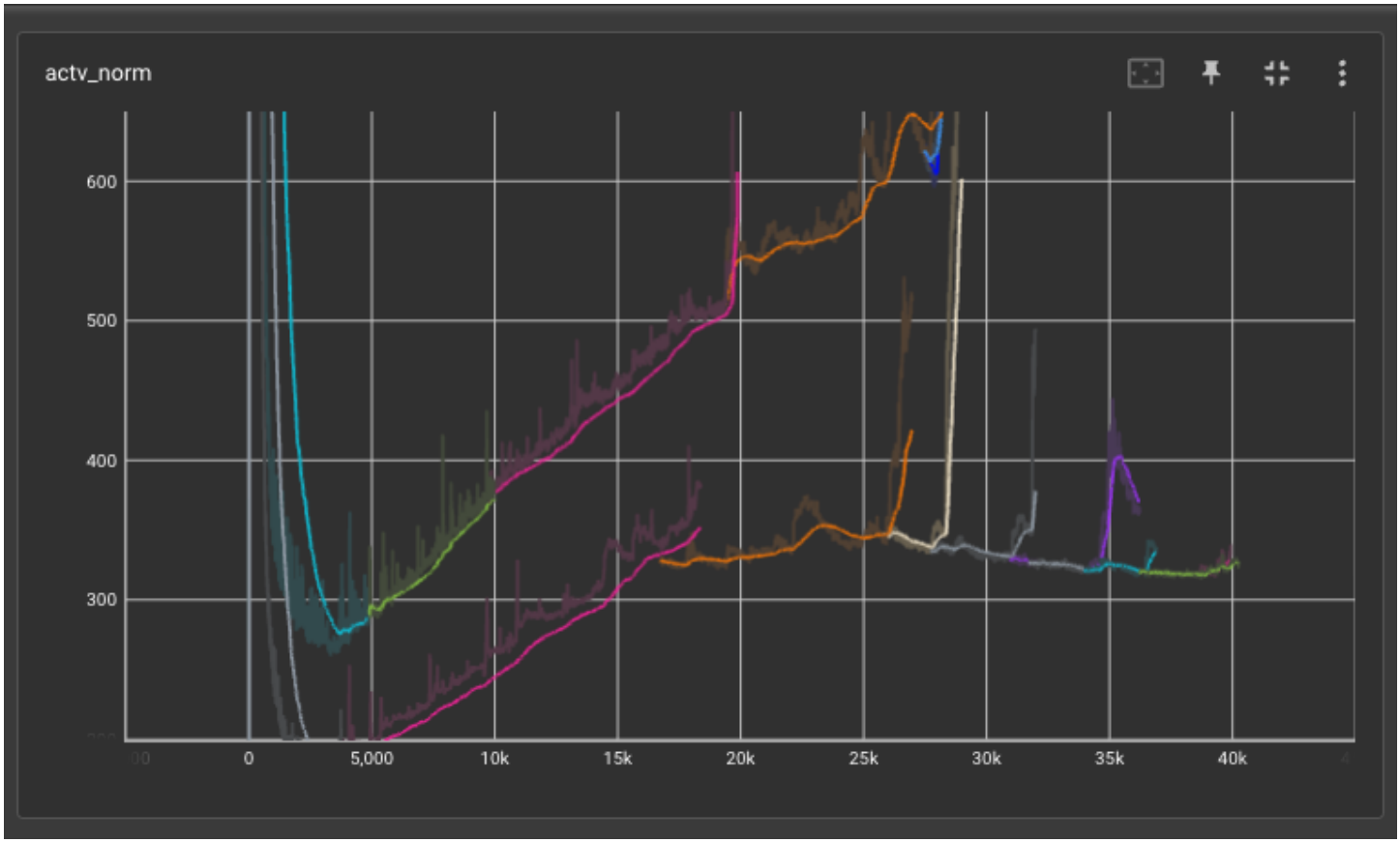
- LR was lowered around here previously as well
    - Bad data batch?

BLOB_PREFIX1="<redacted>/66B_run18"
BLOB_PREFIX2="<redacted>/66B_run20"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_10_27500.pt?${BLOB_AUTH}"
RUN_ID=66B_run20

./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \

--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE

- Lowered LR again to 3e-5 and launched 66B_run21
- Still no bueno



Tensorboard

- Seems like we should be restarting around 24k (or even ~20k) instead, before activation norm curve shot up and became convex.
- Going to restart run18 (starting from 19k) with a lower LR and see if it's more stable.

BLOB_PREFIX1="<redacted>/66B_run17"
BLOB_PREFIX2="<redacted>/66B_run22"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_7_19000.pt?${BLOB_AUTH}"
RUN_ID=66B_run22
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \

```
--restore-file $RESTORE_FILE
```

2022-04-12 [Susan]: Restart from 27.5k with lr of 6e-5, min loss scale reached at 28292

- Reducing LR from 8e-5 to 6e-5

```
BLOB_PREFIX1="<redacted>/66B_run18"
BLOB_PREFIX2="<redacted>/66B_run19"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_10_27500.pt?${BLOB_AUTH}"
RUN_ID=66B_run19
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

2022-04-10 [Susan]: Restart from 19k, clip 1.0 -> 0.3

- Activation norm + ppl diverged.

```
BLOB_PREFIX1="<redacted>/66B_run17"
BLOB_PREFIX2="<redacted>/66B_run18"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_7_19000.pt?${BLOB_AUTH}"
RUN_ID=66B_run18
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

2022-04-07 [Susan]: Restart from 4750, no changes.

- Had to fix the checkpointing path bug anyway, took the chance to upload remaining checkpoints from /mnt/scratch up to Azure blob.
- Restarted from 4750 (last checkpoint before failure).
- Run looks much more stable already. Seems like apex version is likely the culprit for instability.

```
BLOB_PREFIX1="<redacted>/66B_run16"
BLOB_PREFIX2="<redacted>/66B_run17"
BLOB_AUTH="<redacted>"
```

```
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_2_4750.pt?${BLOB_AUTH}"
RUN_ID=66B_run17
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

2022-04-06 [Susan]: Restart 66B from scratch. LR @ 8e-5, clip @ 1.0. Remove no-reshard-after-forward, remove padding (previous validation fix).

Branch susan/66b_apr6_restart from fairseq-big-internal.
Conda env: fairseq-20210913-old-apex

```
BLOB_PREFIX1="<redacted>/66B_run16"
BLOB_AUTH="<redacted>"
RUN_ID=66B_run16
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}"
```

- Tested batch size of 3M, slowed down WPS dramatically.
- Checkpoints not being uploaded, need to pull from scratch space on hosts:

```
 20938       hpc 66B_run1   susanz  R    7:03:28     64
hpc-pg0-[2-3,40-70,76-80,83-97,99-100,121,128-135]

PDSH_RCMD_TYPE=ssh pdsh -w hpc-pg0-[2-3,40-70,76-80,83-97,99-100,121,128-135]
'/shared/home/susanz/bin/azcopy copy "<redacted>/*.pt" "<redacted>/66B_run16?<redacted>"'
```

2022-04-05 [Susan]: revert apex version, keep LR @ 4e-6, move clip back up to 0.3, restart from 38.5k

(Susan's conda env: fairseq-20210913-old-apex ->
/shared/home/susanz/miniconda3/envs/fairseq-20210913-old-apex)

```
BLOB_PREFIX1="<redacted>/66B_run12"
BLOB_PREFIX2="<redacted>/66B_run13"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_14_38500.pt?${BLOB_AUTH}"
RUN_ID=66B_run13
```

./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE

- Still looks to be problematic. Reducing clip down to 0.25 (job was stalled from being held by someone on the cluster anyway).

BLOB_PREFIX1="<redacted>/66B_run12"
BLOB_PREFIX2="<redacted>/66B_run14"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_14_38500.pt?${BLOB_AUTH}"
RUN_ID=66B_run14
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE

Tensorboard

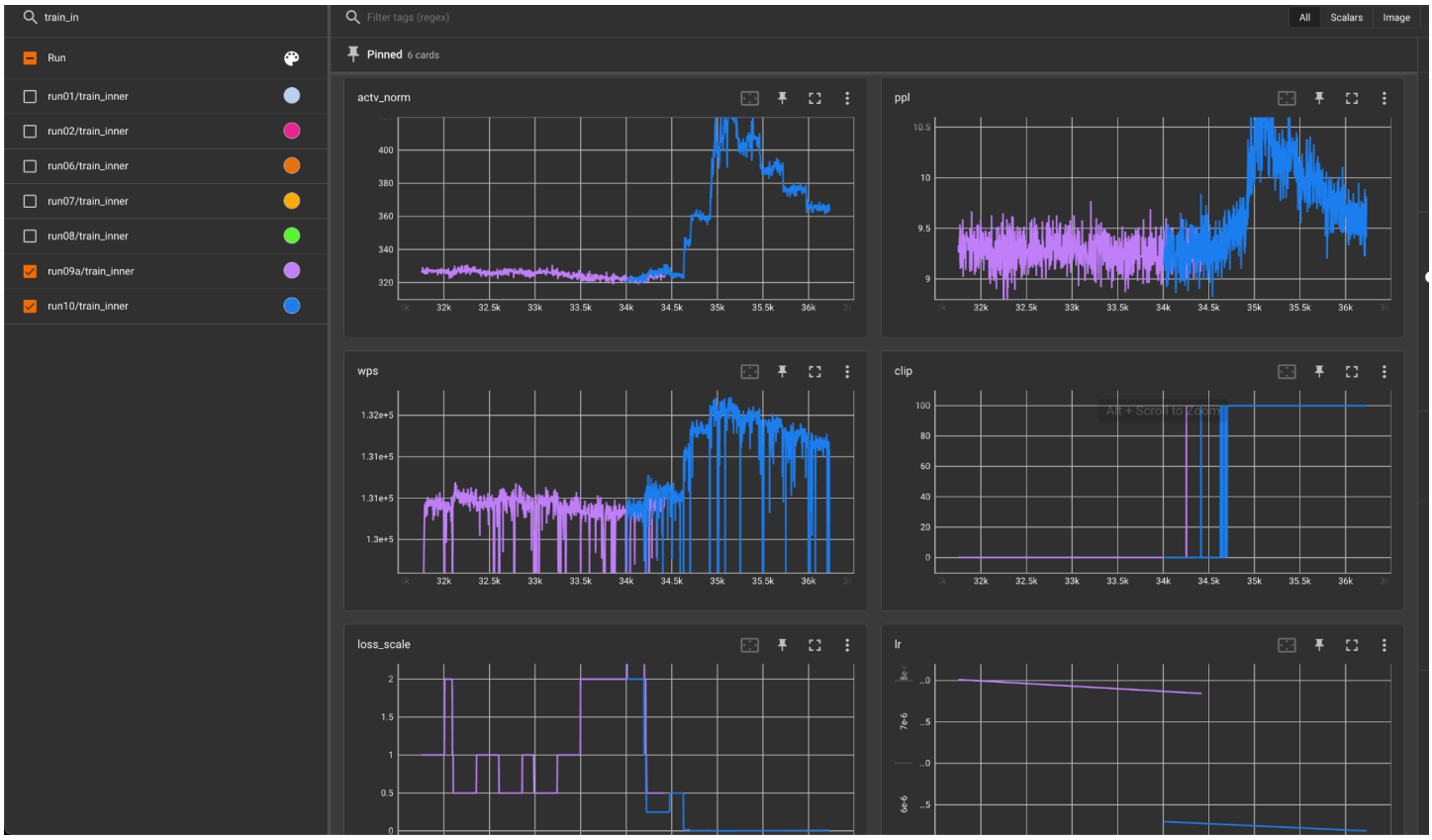
- Effect of changing clipping from 0.3 (pink) -> 0.25 (green)

- LR is now at 2e-6 already

2022-04-04 [Susan]: actv_norm exploding again, lowering LR to 4e-6, clip to 0.2, restart from 36,250

```
BLOB_PREFIX1="<redacted>/66B_run11"
BLOB_PREFIX2="<redacted>/66B_run12"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_13_36250.pt?${BLOB_AUTH}"
RUN_ID=66B_run12
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

2022-04-03 [Susan]: actv_norm exploding again, lowering LR to 6e-6, restart from 34k again

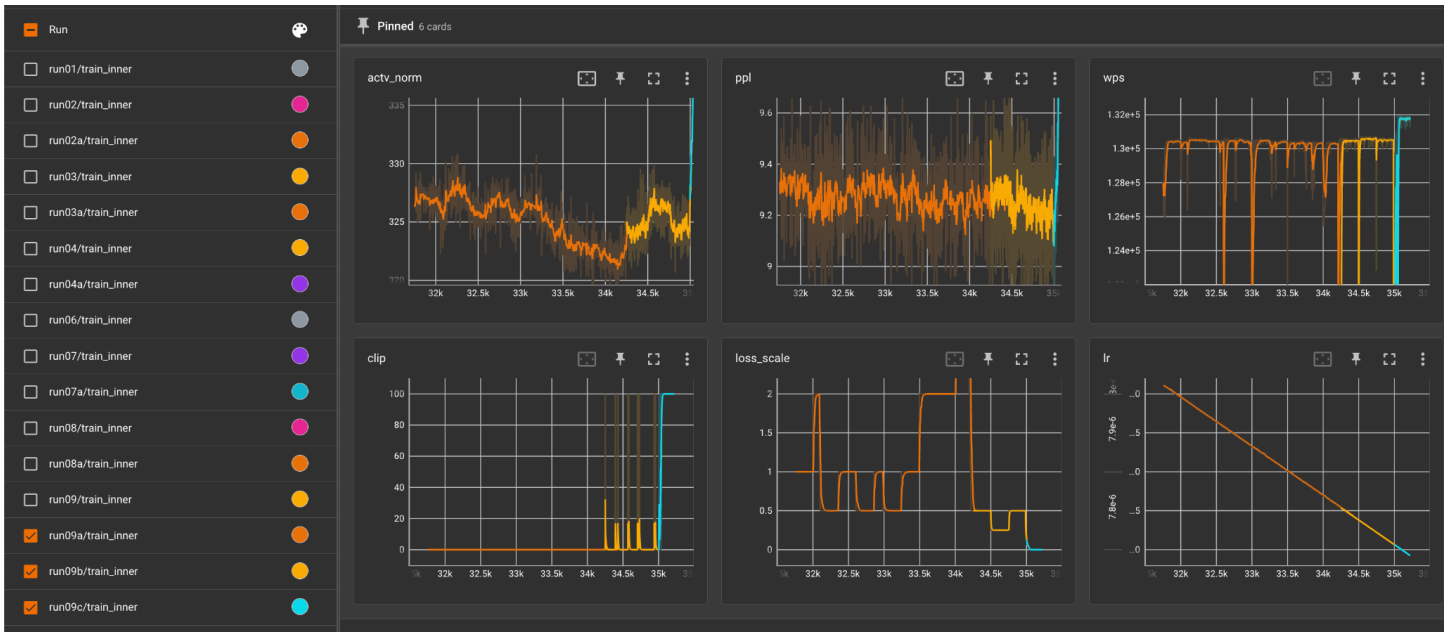


[Tensorboard]

```
BLOB_PREFIX1="<redacted>/66B_run09"
BLOB_PREFIX2="<redacted>/66B_run11"
BLOB_AUTH="<redacted>"
```

```
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_12_34000.pt?${BLOB_AUTH}"
RUN_ID=66B_run11
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

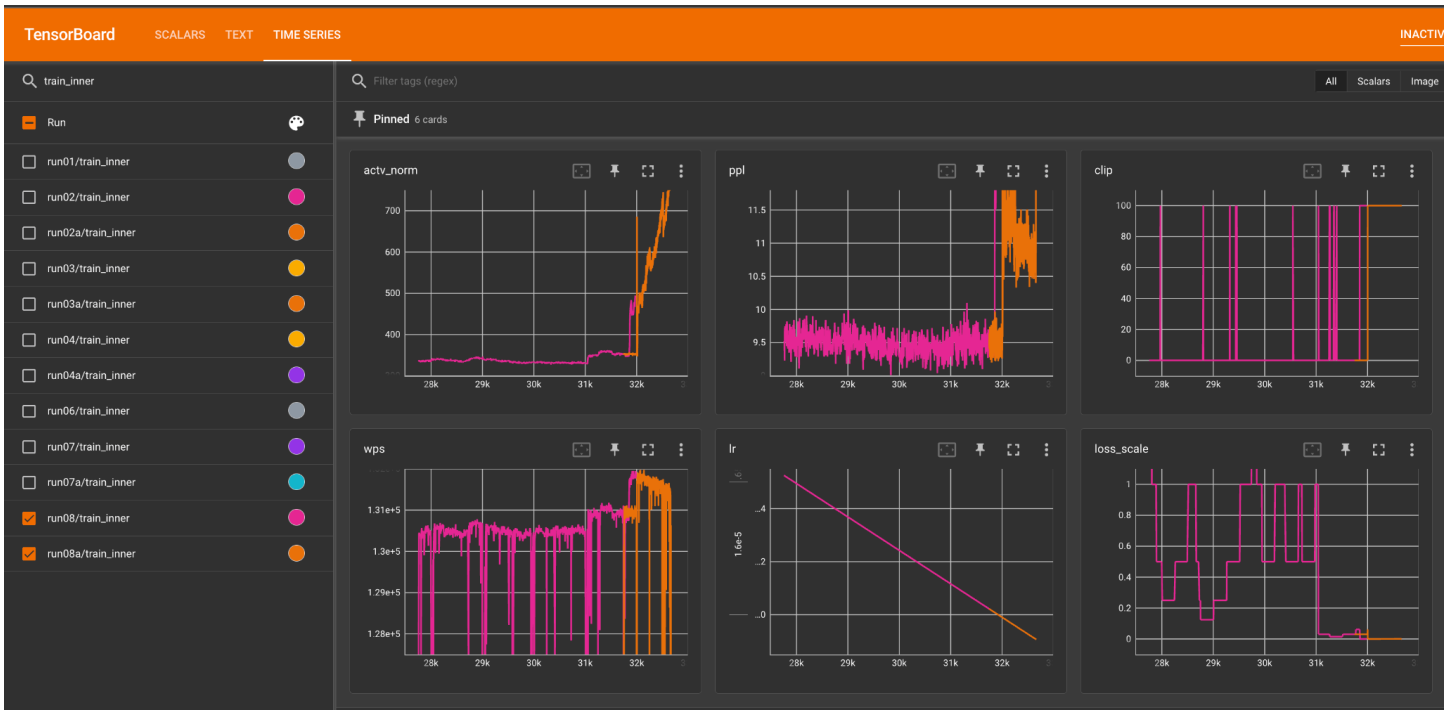
2022-04-01 [Susan]: actv_norm exploding again, lowering LR to 8e-6, restart from 34k



Tensorboard

```
BLOB_PREFIX1="<redacted>/66B_run09"
BLOB_PREFIX2="<redacted>/66B_run10"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_12_34000.pt?${BLOB_AUTH}"
RUN_ID=66B_run10
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

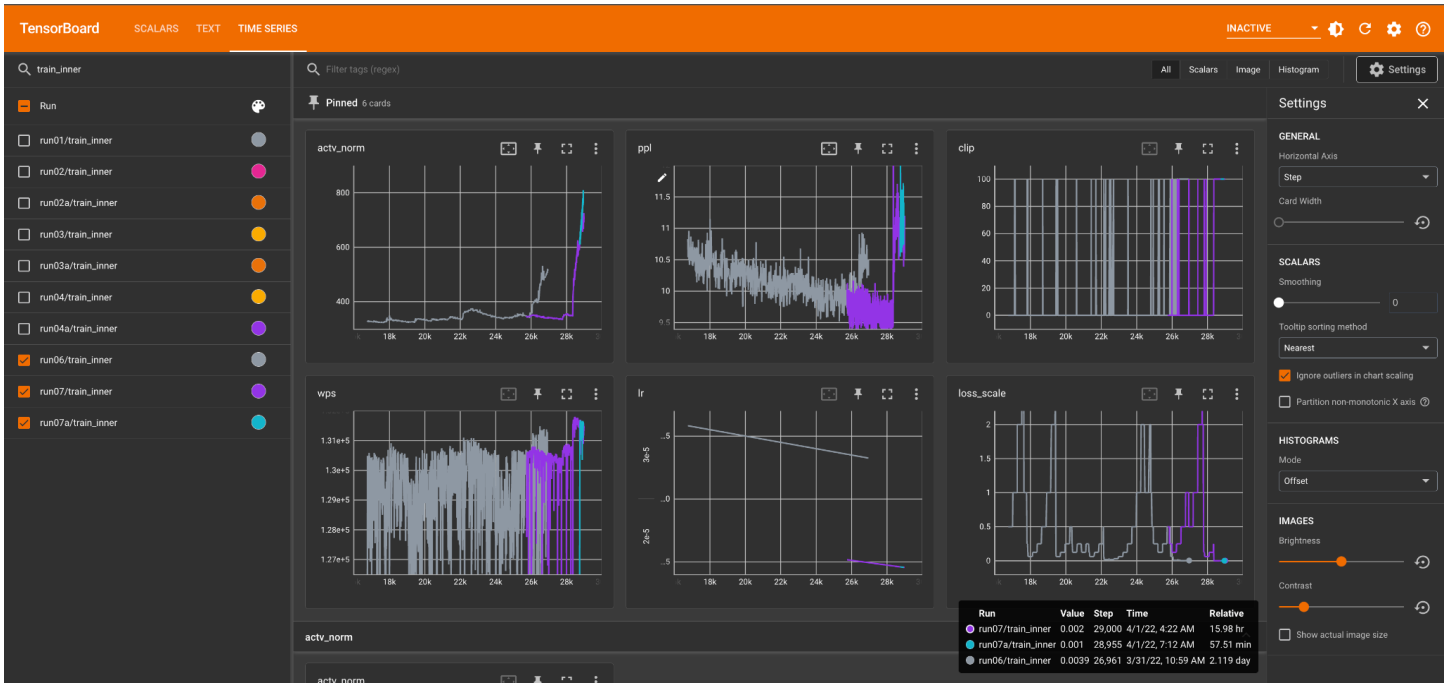
2022-04-01 [Susan]: actv_norm exploding again, lowering LR to 1e-5, restart from 31k



[Tensorboard]

BLOB_PREFIX1="<redacted>/66B_run08"
BLOB_PREFIX2="<redacted>/66B_run09"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_11_31000.pt?${BLOB_AUTH}"
RUN_ID=66B_run09
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE

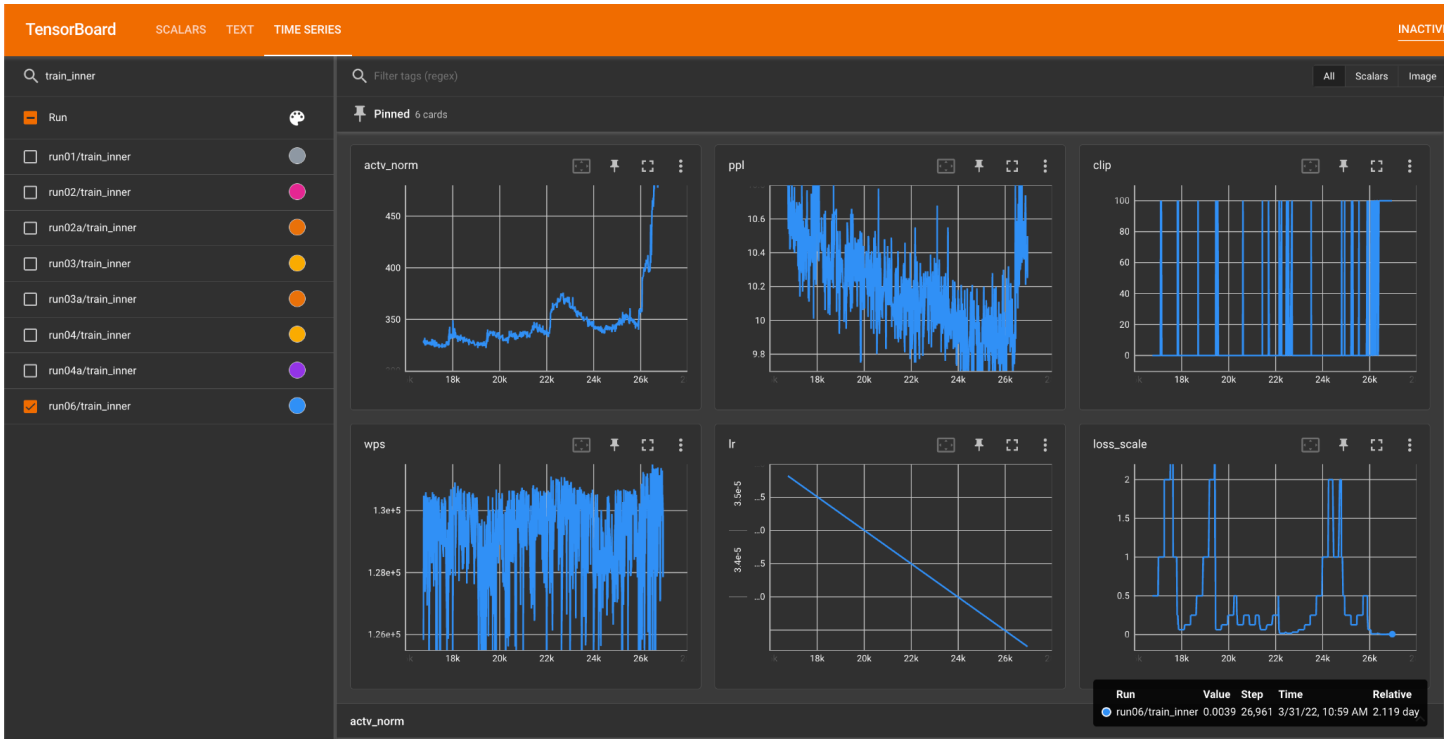
2022-04-01 [Susan]: actv_norm exploding again, lowering LR to 2e-5, restart from 27.75k



[Tensorboard]

```
BLOB_PREFIX1="<redacted>/66B_run07"
BLOB_PREFIX2="<redacted>/66B_run08"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_10_27750.pt?${BLOB_AUTH}"
RUN_ID=66B_run08
/<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

2022-03-31 [Susan]: actv_norm exploding again, lowering LR to 3e-5, restart from 25.75k

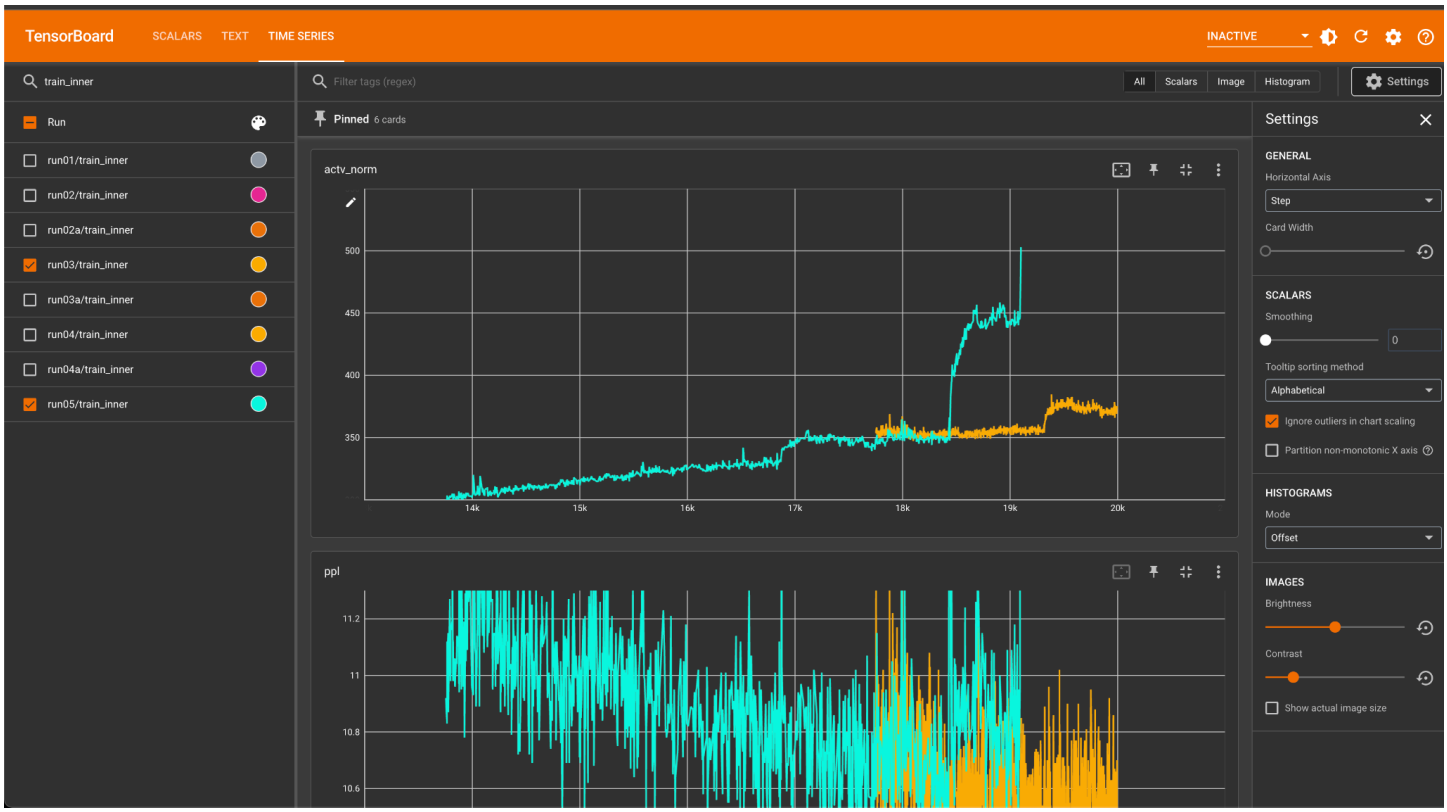


[Tensorboard]

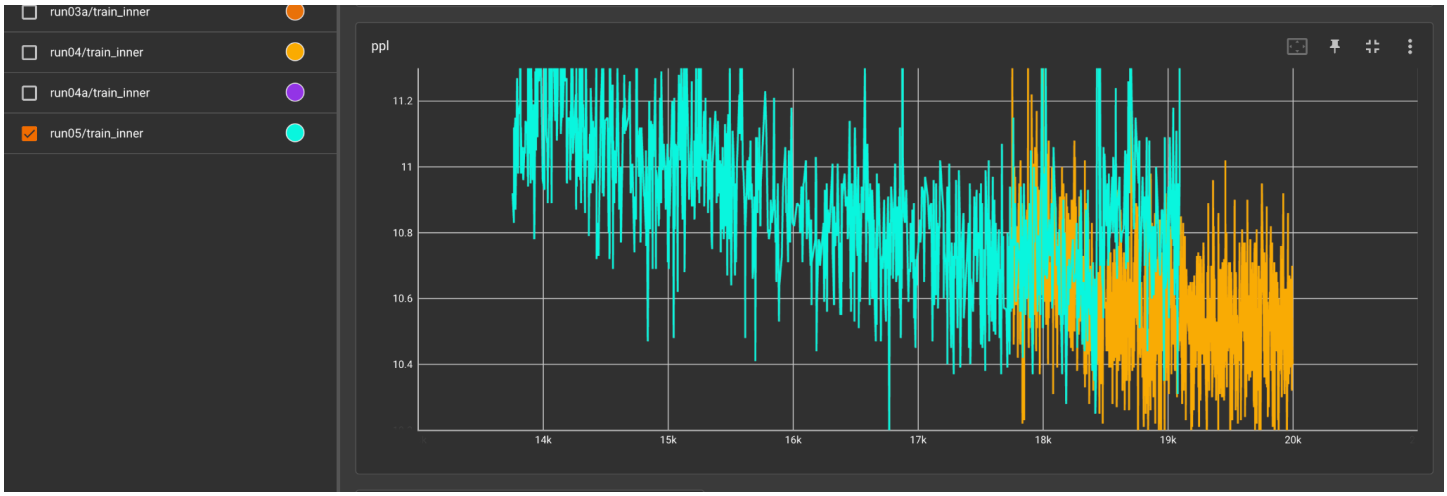
- Reverting to 25.75k when loss scale was still “healthy” at around 0.5, lowering LR again from 4e-5 to 3e-5

```
BLOB_PREFIX1="<redacted>/66B_run06"
BLOB_PREFIX2="<redacted>/66B_run07"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_9_25750.pt?${BLOB_AUTH}"
RUN_ID=66B_run07
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

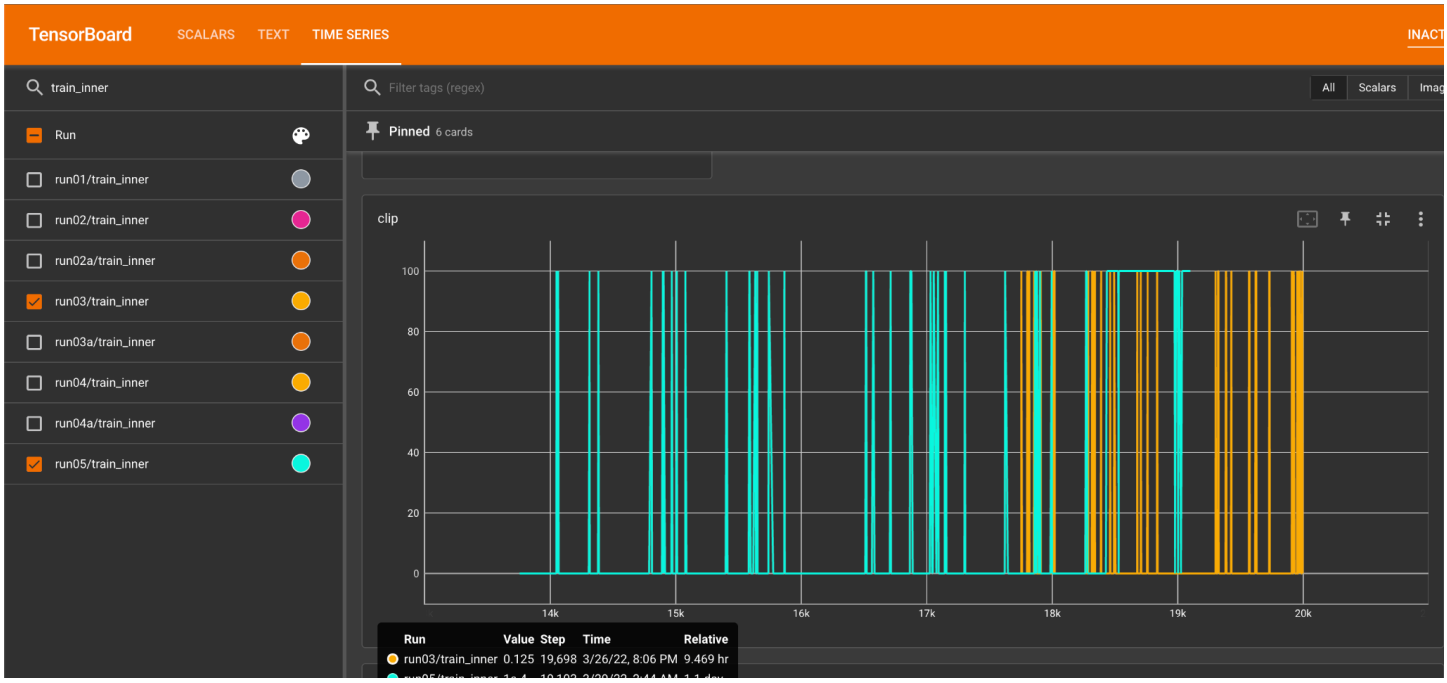
2022-03-29 [Susan]: actv_norm exploding again, lowering LR to 4e-5, restart from 16.75k



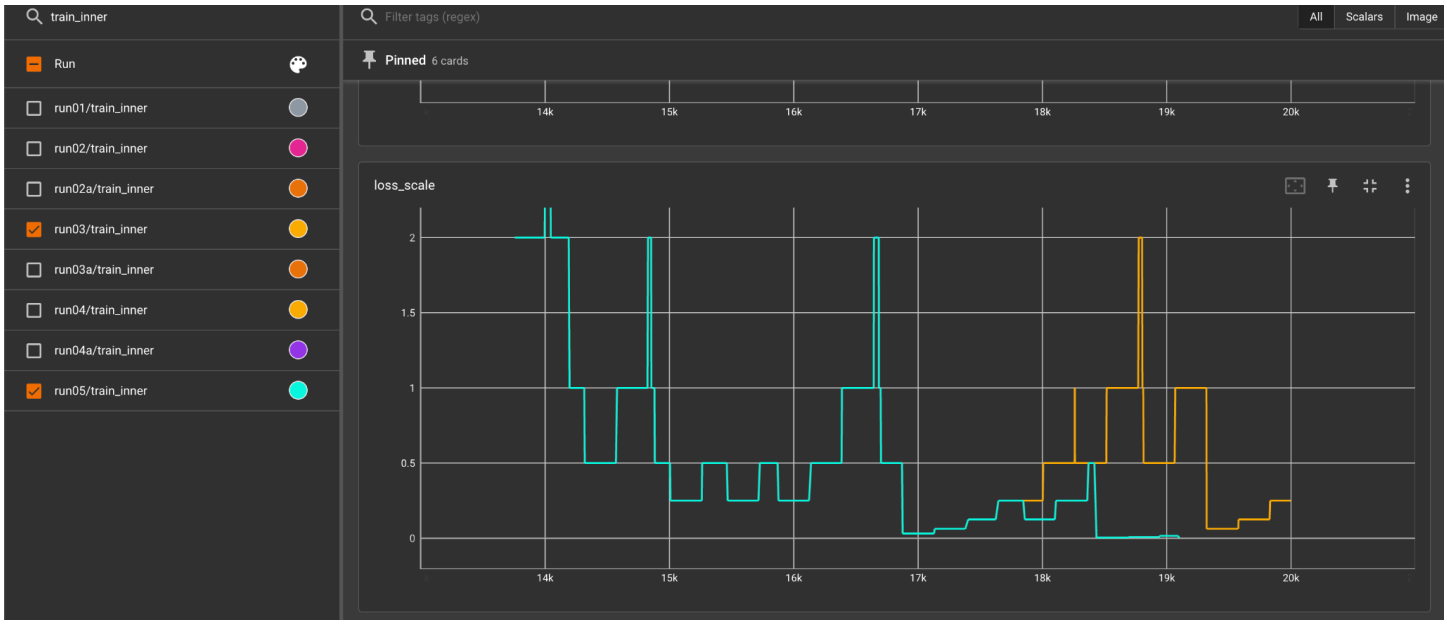
Tensorboard



Tensorboard



[Tensorboard]



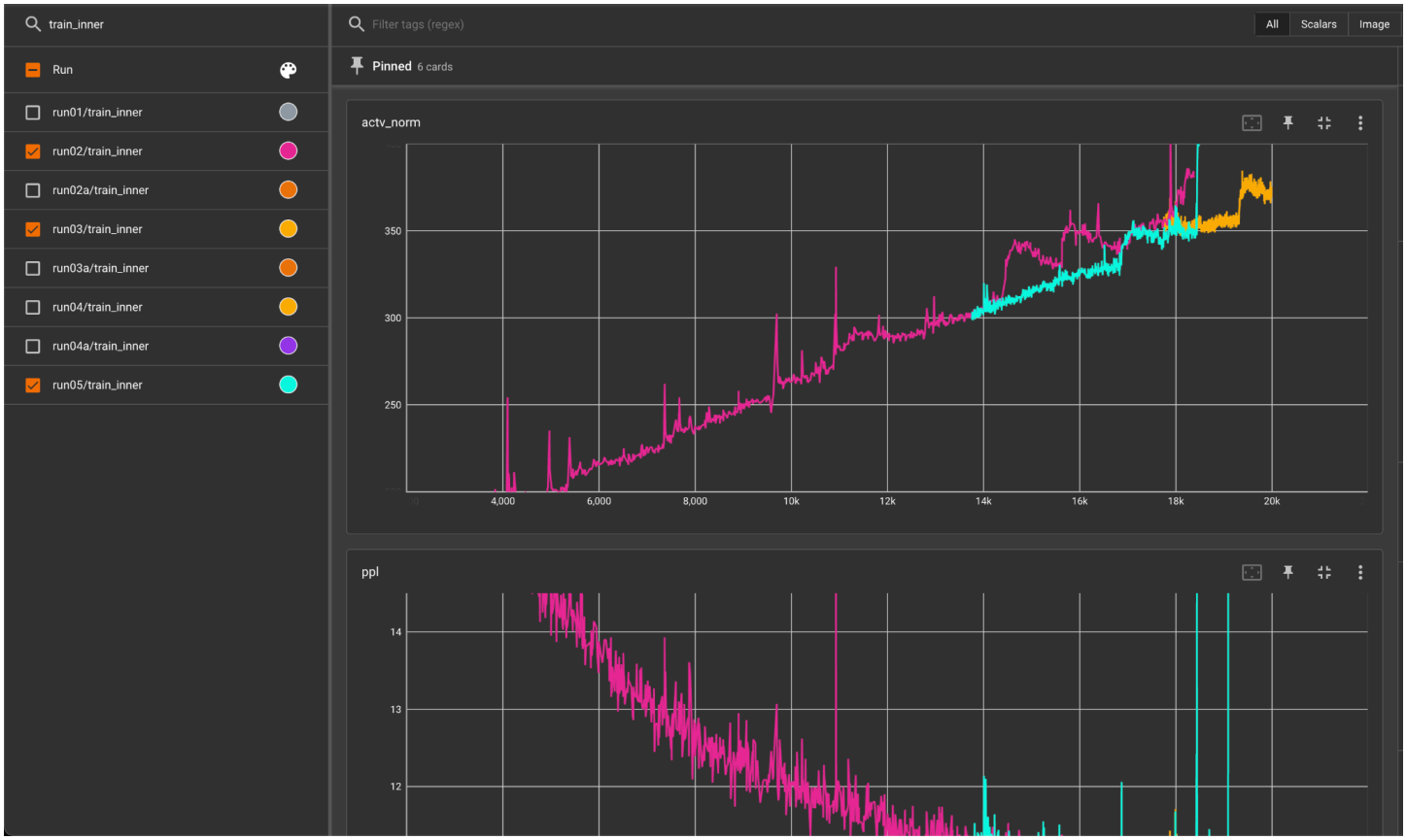
[Tensorboard]

- Accidentally deleted the run05 logs from the cluster, took above screenshots before tensorboard refreshed.
- Lowered LR to 4e-5
- Relaunched with:

```
BLOB_PREFIX1="<redacted>/66B_run05"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_6_16750.pt?${BLOB_AUTH}"
BLOB_PREFIX2="<redacted>/66B_run06"
BLOB_AUTH="<redacted>"
RUN_ID=66B_run06
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
```

--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE

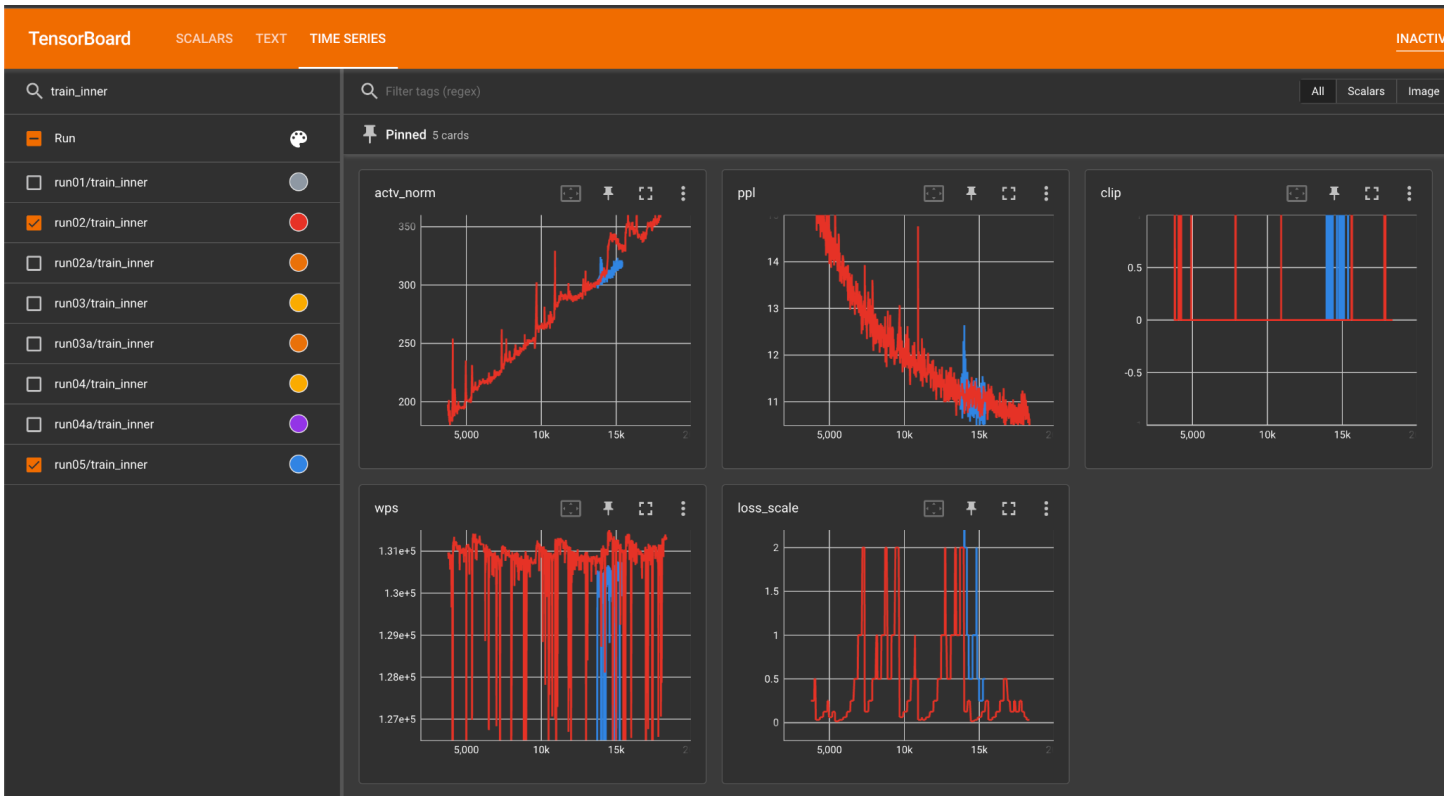
- Doing some more comparisons before data is refreshed in tb



Tensorboard

Tensorboard

2022-03-28 [Susan]: Checking in - turning on clip 0.3 earlier seems to have been necessary



[Tensorboard]

- From clip == 0.3, we start clipping much earlier (blue vs red, where clip == 1.0).
- Loss scales look healthier after clipping too (doesn't crash to < 0.25).
- Activation norm is also lowered.

2022-03-27 [Susan]: Restarting from 13,750, with clip == 0.3

- Activation norm at infinity

Tensorboard

- Seems like lowering clip to 0.3 and restarting from 17,750 wasn't drastic enough. Trying a rollback to even further back.
- LR is already set at the same value as what we had for the 175B. Batch size is also the same (2M).
    - We did lower LR to 3e-5 after ~91k in the 175B run. Worst case we do this earlier here too.

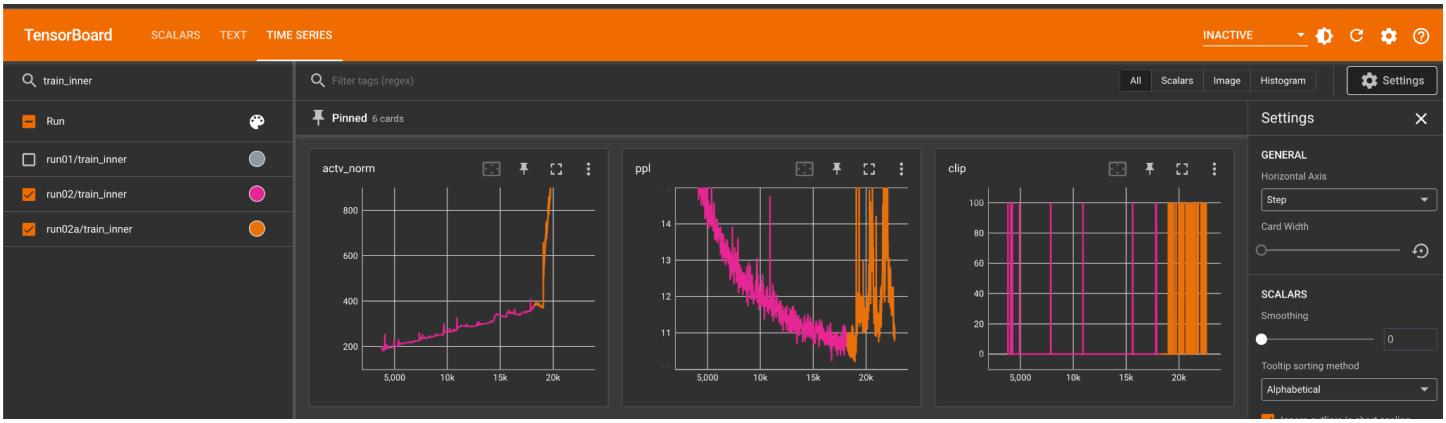
```
BLOB_PREFIX="<redacted>/66B"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX}/checkpoint_5_13750.pt?${BLOB_AUTH}"
RUN_ID=66B_run04
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

- Auto restart ended up resuming from the end of run03, given the same blob prefix. Needed to restart again with separating the restore file path from the upload path:

```
BLOB_PREFIX1="<redacted>/66B"
RESTORE_FILE="${BLOB_PREFIX1}/checkpoint_5_13750.pt?${BLOB_AUTH}"
BLOB_PREFIX2="<redacted>/66B_run05"
BLOB_AUTH="<redacted>"
RUN_ID=66B_run05
./<redacted> \
-n 64 -g 8 -t 1 \
```

-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX2}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE

2022-03-26 [Susan]: Restarting from 17,750, lowering clip to 0.3



Tensorboard

- Diverged and looks like it was unable to recover
- Restarting from 17,750, before we started clipping heavily when clip was 1.0

BLOB_PREFIX="<redacted>/66B"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX}/checkpoint_7_17750.pt?${BLOB_AUTH}"
RUN_ID=66B_run03
EXCLUDED_HOSTS=hpc-pg0-[9,11] \
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE

- Manual host exclusion due to https://github.com/fairinternal/cluster-health/pull/4#discussion_r835766331

2022-03-21 [Susan]: Restarting from 3.75k, validation on

- Rebased susan/66b_mar20_restart on top of main, after validation fix went in

LOB_PREFIX="<redacted>/66B"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX}/checkpoint_2_3750.pt?${BLOB_AUTH}"

```
RUN_ID=66B_run02
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX}/?${BLOB_AUTH}" \
--restore-file $RESTORE_FILE
```

Cluster maintenance

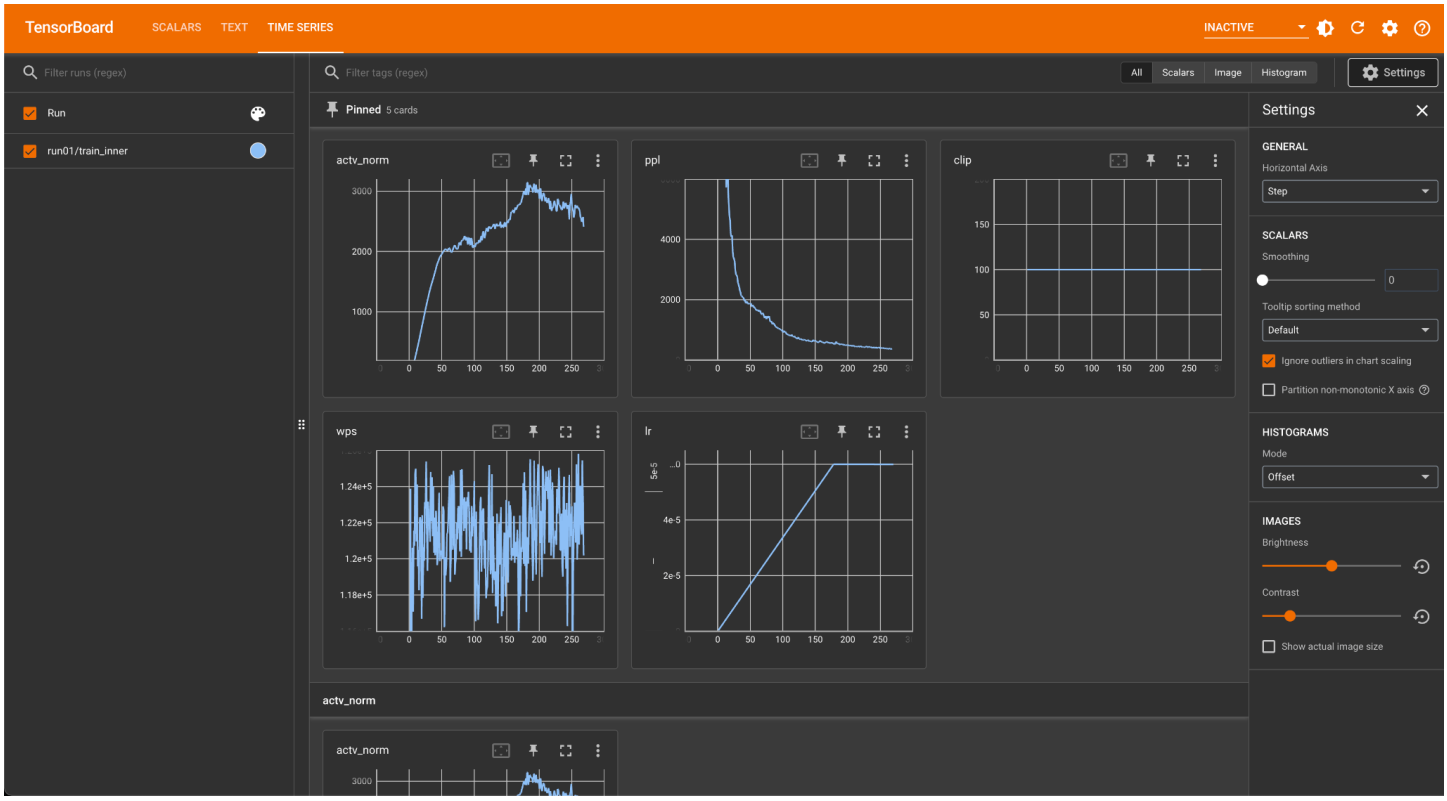
- 9 hosts in drain
    - hpc-pg0-6 seems to have ECC uncorrectable errors
    - hpc-pg0-114 seem to be diagnosed with bad IB as well
    - hpc-pg0-[46,55,68] all seem to have lost GPU
    - Rest were identified and reported as having bad IB already
- Pip installed new cluster-health module: https://github.com/fairinternal/cluster-health

2022-03-20 [Susan]: Relaunching 66B from scratch, from latest code, 6e-5 LR from start

- Bring us up to the point of latest fairseq-big-internal
- Branched off for 66B run: https://github.com/fairinternal/fairseq-big-internal/pull/128
- Discussions w/ Stephen & Anj on restarting clean, mainly to catch new code changes / potential env differences (wps diffs ended up being IB issues)

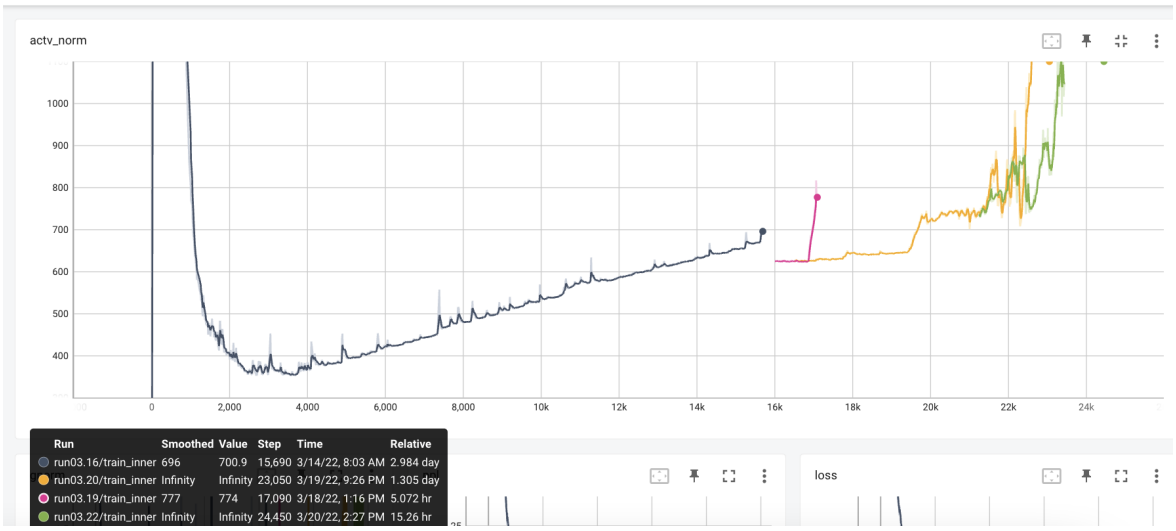
```
BLOB_PREFIX="<redacted>/66B"
BLOB_AUTH="<redacted>"
RUN_ID=66B_run01
./<redacted> \
-n 64 -g 8 -t 1 \
-p $RUN_ID \
--azure \
--model-size 66b \
--checkpoints-dir /shared/home/susanz/checkpoints/66B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--full-azure-upload-path "${BLOB_PREFIX}/?${BLOB_AUTH}"
```

- Activation norm drops ~200 steps in with 6e-5 LR



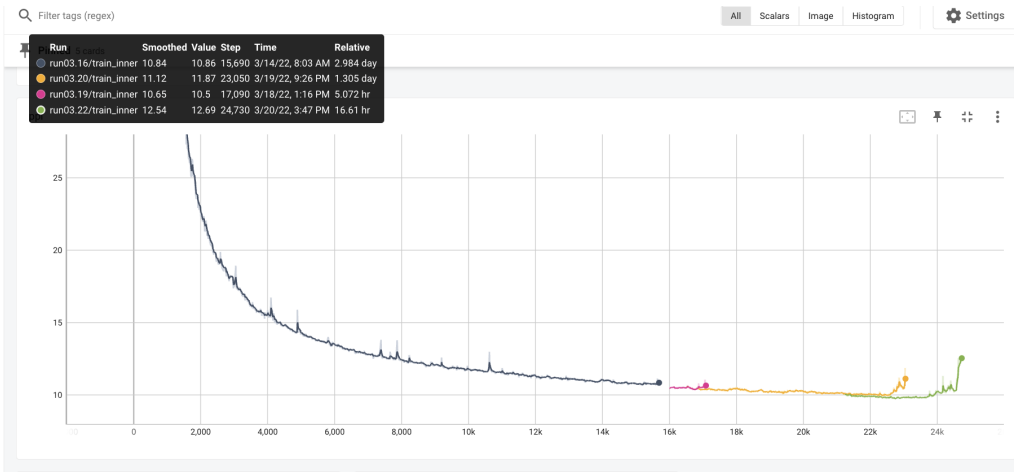
[Tensorboard]

2022-03-20 [Anjali] Made it past the last instability point, activ_norm is trending up



[Tensorboard]

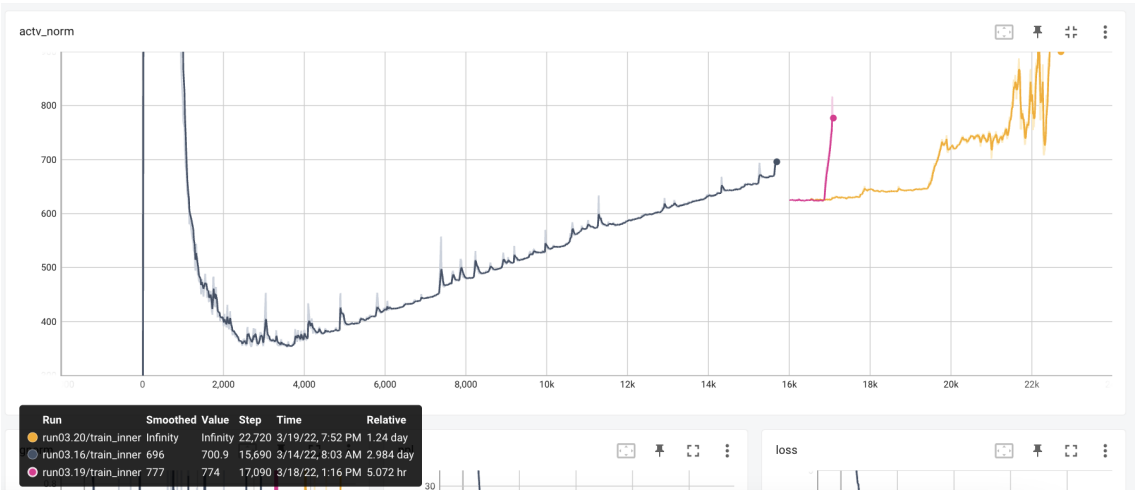
PPL went below lowest point but then increased again



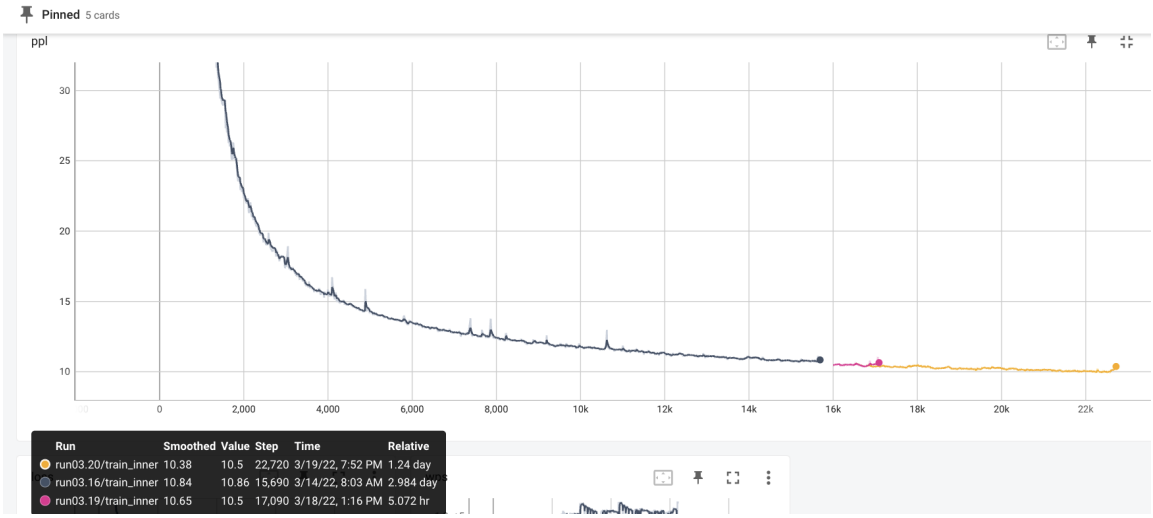
[Tensorboard]

We will need to restart the job with 6e-5 way before than we did i.e at 21.25k. Maybe at 10k or even earlier.

2022-03-19 [Anjali] Activ norm is blowing up. Restarting job with lower LR=6.0e-5



[Tensorboard]



[Tensorboard]

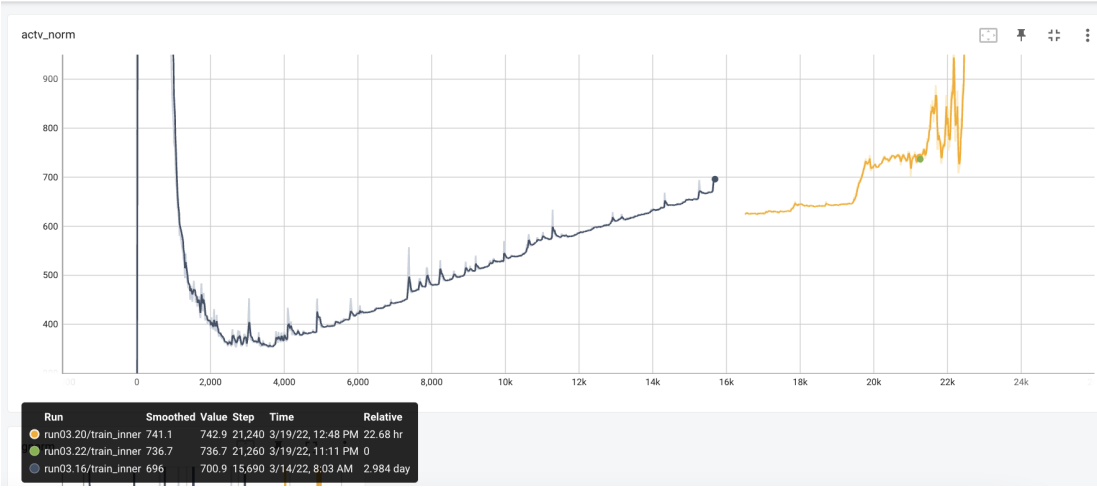
Restarting job with a lower LR of 6e-5.

RUN_ID=66B_run03.21
RESTORE_BLOB_PREFIX="<redacted>/66B_run03.20"
BLOB_PREFIX="<redacted>/66B_run03.21"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${RESTORE_BLOB_PREFIX}/checkpoint_8_21250.pt?${BLOB_AUTH}"

RUN_ID=66B_run03.22
RESTORE_BLOB_PREFIX="<redacted>/66B_run03.20"
BLOB_PREFIX="<redacted>/66B_run03.22"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${RESTORE_BLOB_PREFIX}/checkpoint_8_21250.pt?${BLOB_AUTH}"

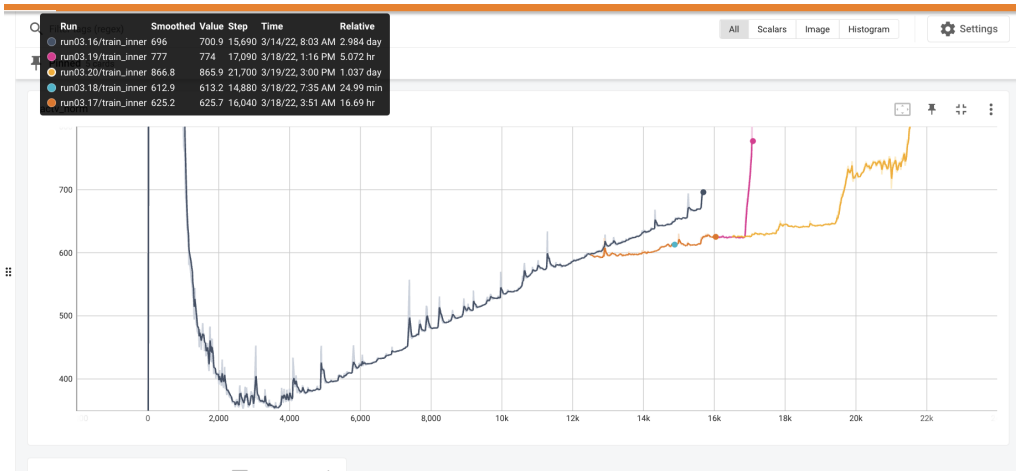
python -m <redacted baselines script> -n 64 -g 8 -t 1 -p $RUN_ID --checkpoints-dir /shared/home/anj/checkpoints/66B/ --local-checkpoints-dir /mnt/scratch/anj/checkpoints/$(date +%Y-%m-%d).$RUN_ID --full-azure-upload-path "${BLOB_PREFIX}/?${BLOB_AUTH}" --model-size 66b --restore-file $RESTORE_FILE

Restarted at the right point:



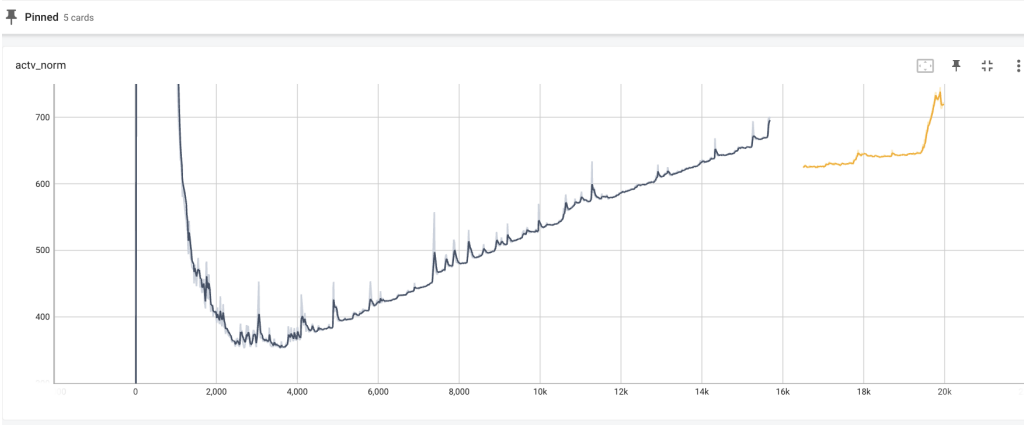
Tensorboard

2022-03-19 [Anjali] Spike again at 21.5k steps and monitoring



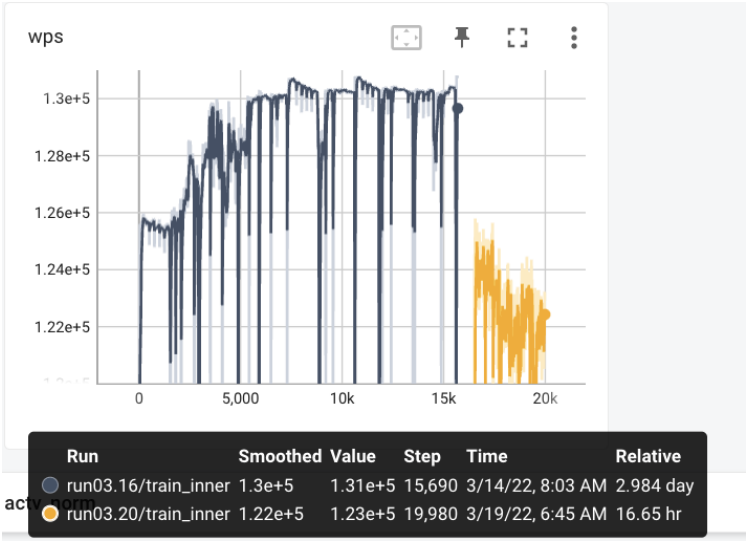
Tensorboard

2022-03-19 [Anjali] Spike of activation norm at 20k steps but trending down



Tensorboard

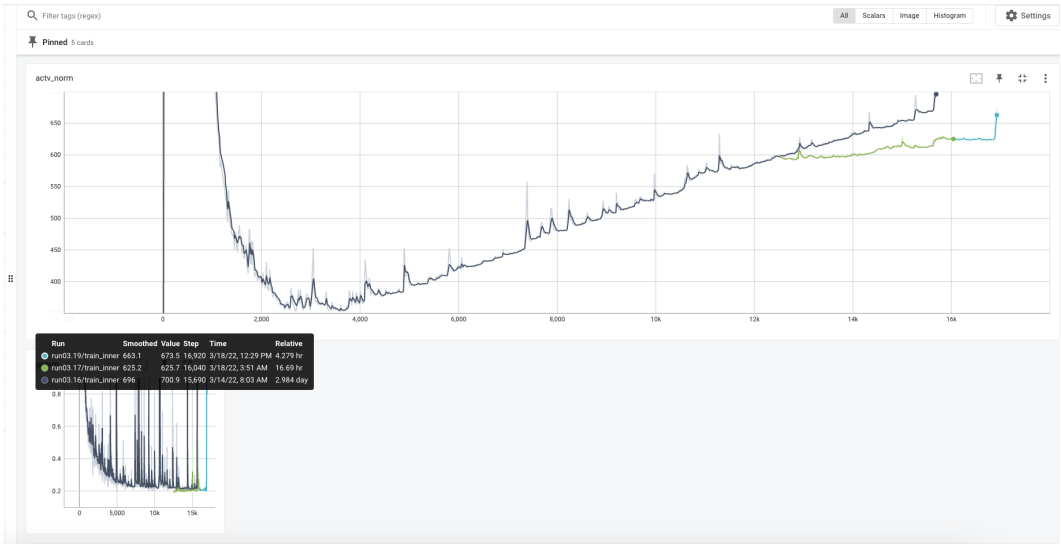
Another thing I observed is that post the clip-norm change the wps has dropped from 126k to 123k. Still within acceptable range but we should follow up:



Tensorboard

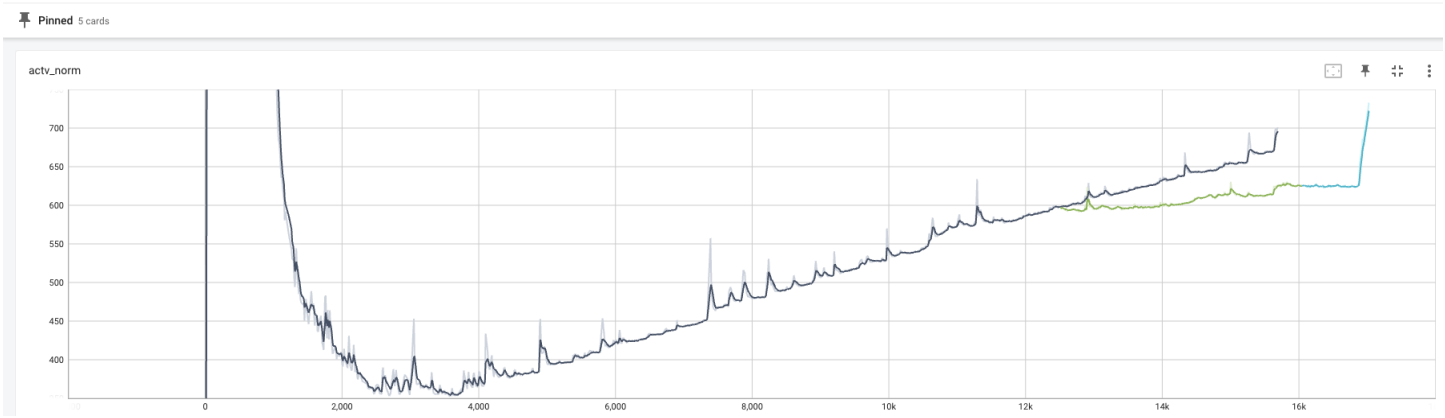
2022-03-18 [Anjali] Activ norm spiking, restarting job with clip-norm=0.3

Slight spike in the activation norm. PPL seems ok so monitoring for now since there have been spikes previously



[Tensorboard]

Looks like activ norm is getting unstable:



[Tensorboard]

Restarting job with clip-norm=0.3

RUN_ID=66B_run03.20
RESTORE_BLOB_PREFIX="<redacted>/66B_run03.19"
BLOB_PREFIX="<redacted>/66B_run03.20"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${RESTORE_BLOB_PREFIX}/checkpoint_6_16500.pt?${BLOB_AUTH}"

python -m <redacted baselines script> -n 64 -g 8 -t 1 -p $RUN_ID --checkpoints-dir /shared/home/anj/checkpoints/66B/ --local-checkpoints-dir /mnt/scratch/anj/checkpoints/$(date +%Y-%m-%d).$RUN_ID --full-azure-upload-path "${BLOB_PREFIX}/?${BLOB_AUTH}" --model-size 66b --restore-file $RESTORE_FILE

2022-03-18 [Anjali] Job failed, restarting again at 16k (or where the job died)

Checkpoint from 14k step.
Branch: anj/66b_gcmf_1
Same LR: 8.0e-5

RUN_ID=66B_run03.18
RESTORE_BLOB_PREFIX="<redacted>/66B_run03.01"

```
BLOB_PREFIX="<redacted>/66B_run03.18"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${RESTORE_BLOB_PREFIX}/checkpoint_last.pt?${BLOB_AUTH}"

 python -m <redacted baselines script>     -n 64 -g 8 -t 1     -p $RUN_ID     --checkpoints-dir /shared/home/anj/checkpoints/66B/     --local-checkpoints-dir /mnt/scratch/anj/checkpoints/$(date +%Y-%m-%d).$RUN_ID     --full-azure-upload-path "${BLOB_PREFIX}/?${BLOB_AUTH}"     --model-size 66b --restore-file $RESTORE_FILE
```

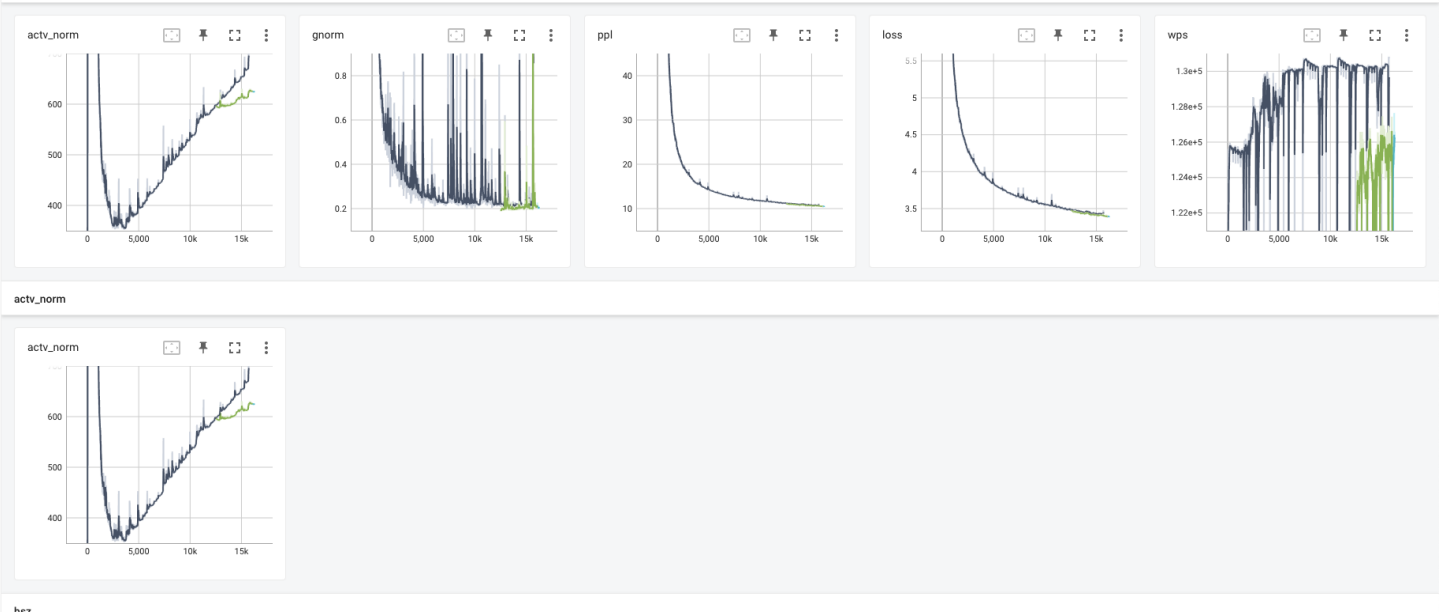
Looks like I started a little too behind. I am going to fast forward to 16k steps and restart the run.

```
RUN_ID=66B_run03.19
RESTORE_BLOB_PREFIX="<redacted>/66B_run03.01"
BLOB_PREFIX="<redacted>/66B_run03.19"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${RESTORE_BLOB_PREFIX}/checkpoint_6_16000.pt?${BLOB_AUTH}"

python -m <redacted baselines script>     -n 64 -g 8 -t 1     -p $RUN_ID     --checkpoints-dir /shared/home/anj/checkpoints/66B/     --local-checkpoints-dir /mnt/scratch/anj/checkpoints/$(date +%Y-%m-%d).$RUN_ID     --full-azure-upload-path "${BLOB_PREFIX}/?${BLOB_AUTH}"     --model-size 66b --restore-file $RESTORE_FILE
```

Logs look to be close to the last starting point.

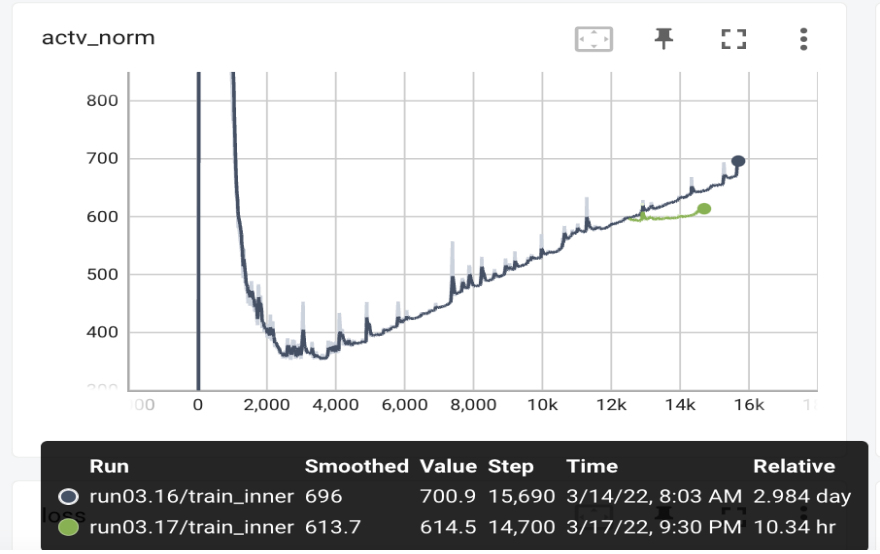
Looks ok at 9:18AM PST



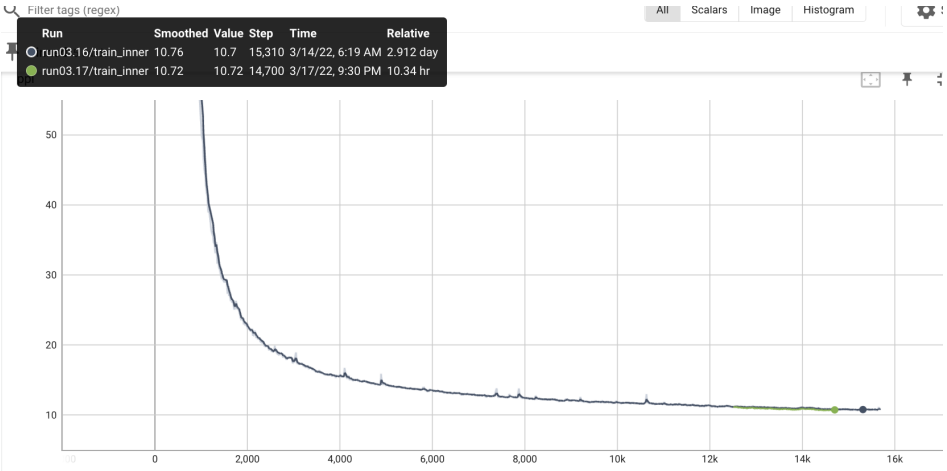
[Tensorboard]

2022-03-17 [Anjali] Update at ~14.5k steps

Graphs trending as below: activ norm is slowly increasing but lower than previous runs. Wps is about the same. Loss and PPL are following previous trends.

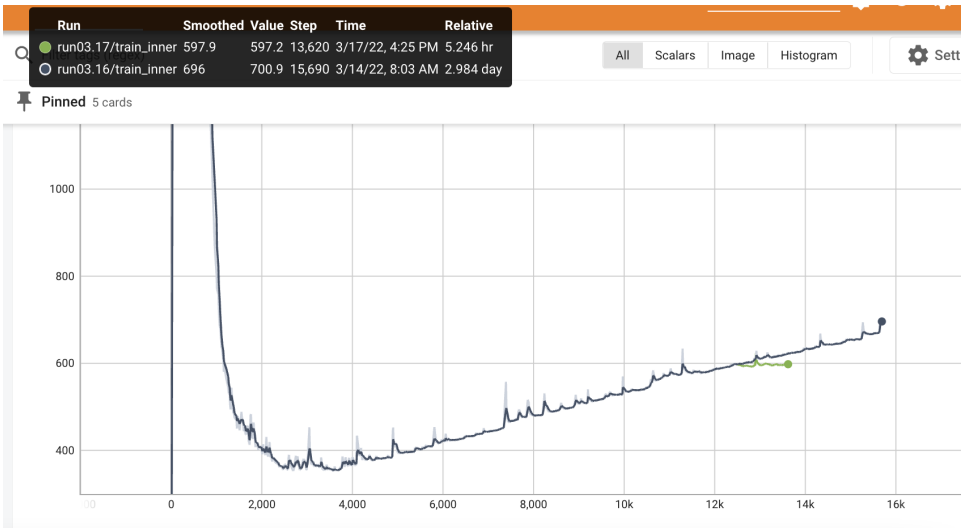


[Tensorboard]



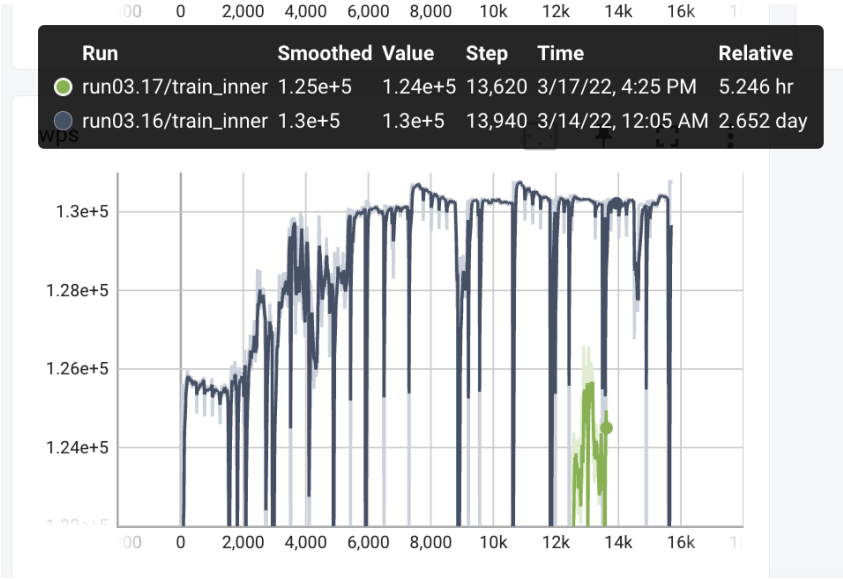
[Tensorboard]

2022-03-17 [Anjali] Update at 13.5k steps with new 8.0e-5 LR



[Tensorboard]

WPS is lower than before restarts:



[Tensorboard]

However the TFLOPs are still 133 so what we had initially gotten. It looks like it increased steadily after 2k steps so I'm going to see if we see the same behavior. Unsure why the speed was higher before.

2022-03-17 [Anjali]: Restart 66B run from checkpoint at 12k steps

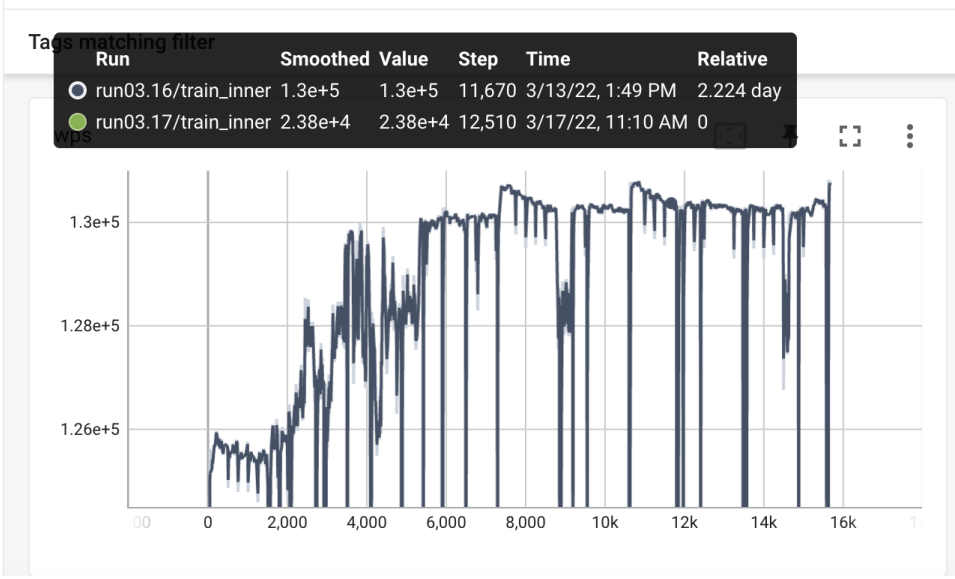
Run failed at 2022-03-17 13:41:15 - Loss exploded

Checkpoint from 12k step. Branch: anj/66b_gcmf_1.

RUN_ID=66B_run03.17
BLOB_PREFIX="<redacted>/66B_run03.01"
BLOB_AUTH="<redacted>"
RESTORE_FILE="${BLOB_PREFIX}/checkpoint_5_12500.pt?${BLOB_AUTH}"

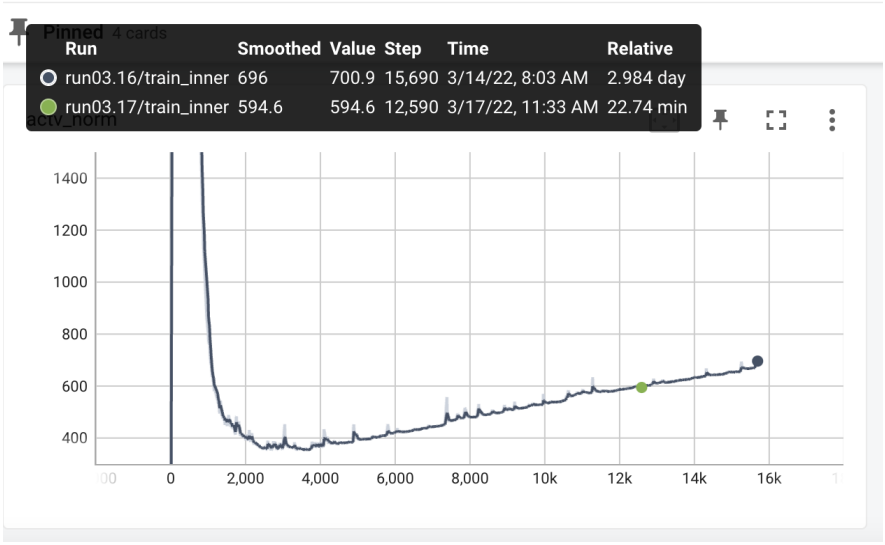
python -m <redacted baselines script> -n 64 -g 8 -t 1 -p $RUN_ID --checkpoints-dir /shared/home/anj/checkpoints/66B/ --local-checkpoints-dir /mnt/scratch/anj/checkpoints/$(date +%Y-%m-%d).$RUN_ID --full-azure-upload-path "${BLOB_PREFIX}/?${BLOB_AUTH}" --model-size 66b --restore-file $RESTORE_FILE

Output logs from restarted job show wps is back to ~124k
We started with ~124k throughput in the beginning of the run before it got higher. Trying to understand why we end up increasing the throughput over time.



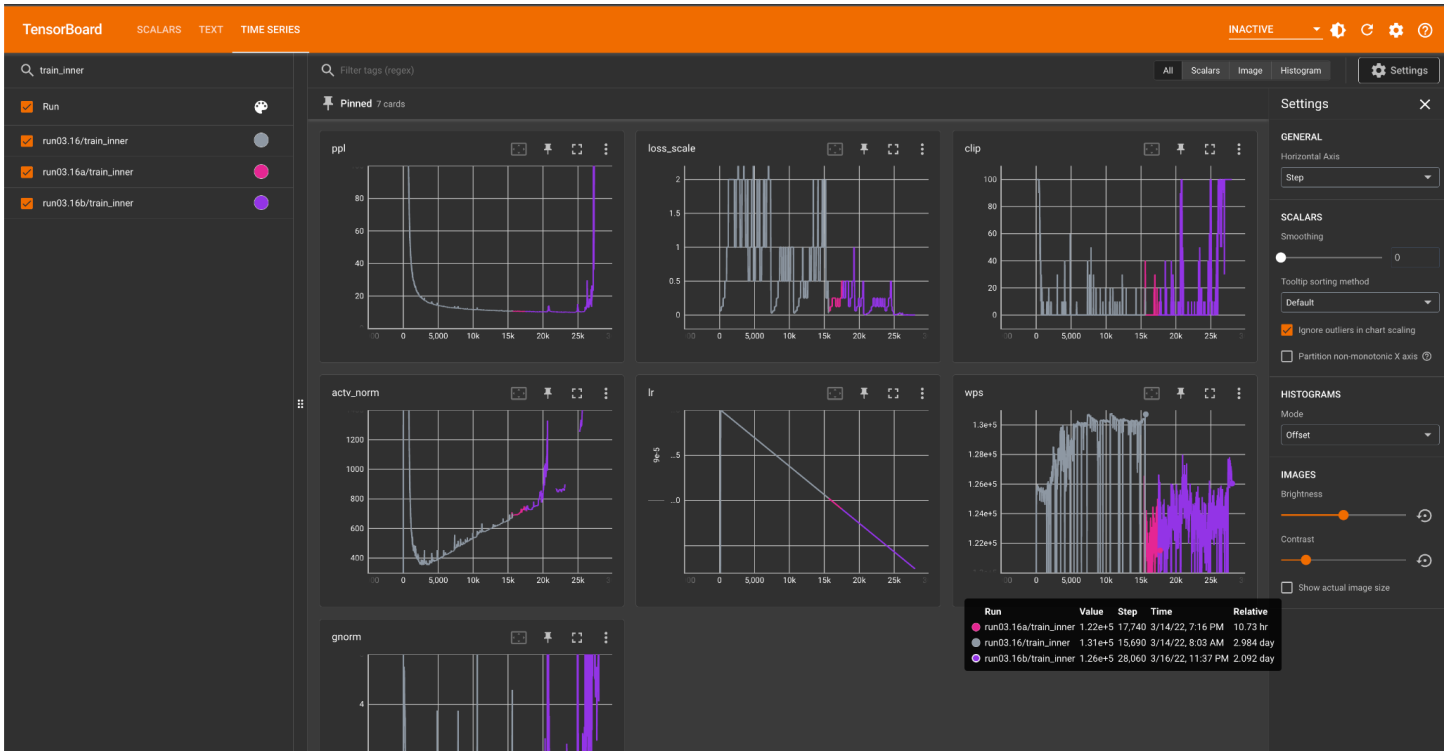
[Tensorboard]

Restarted from checkpoint where there was high WPS and stability:



[Tensorboard]

2021-03-15 [Susan]: 66B diverges, wps change upon restarts?



[Tensorboard]

Zooming in:

[Tensorboard]

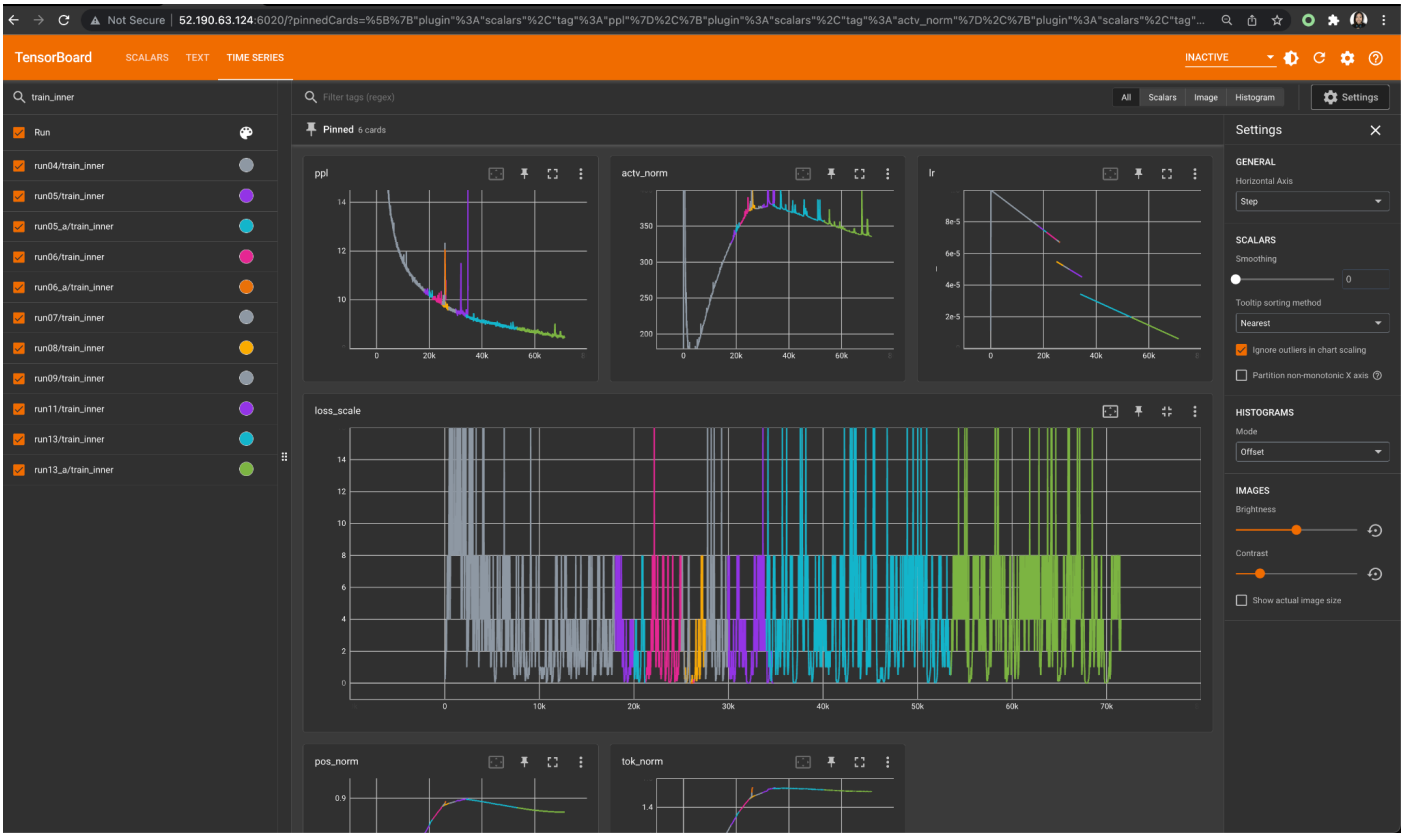
- Tensorboard is pointing to /shared/home/anj/checkpoints/66B/tensorboard/run03/tb, so added symlinks to the above symlinks there as well.
- Next steps:

- Restart from ???

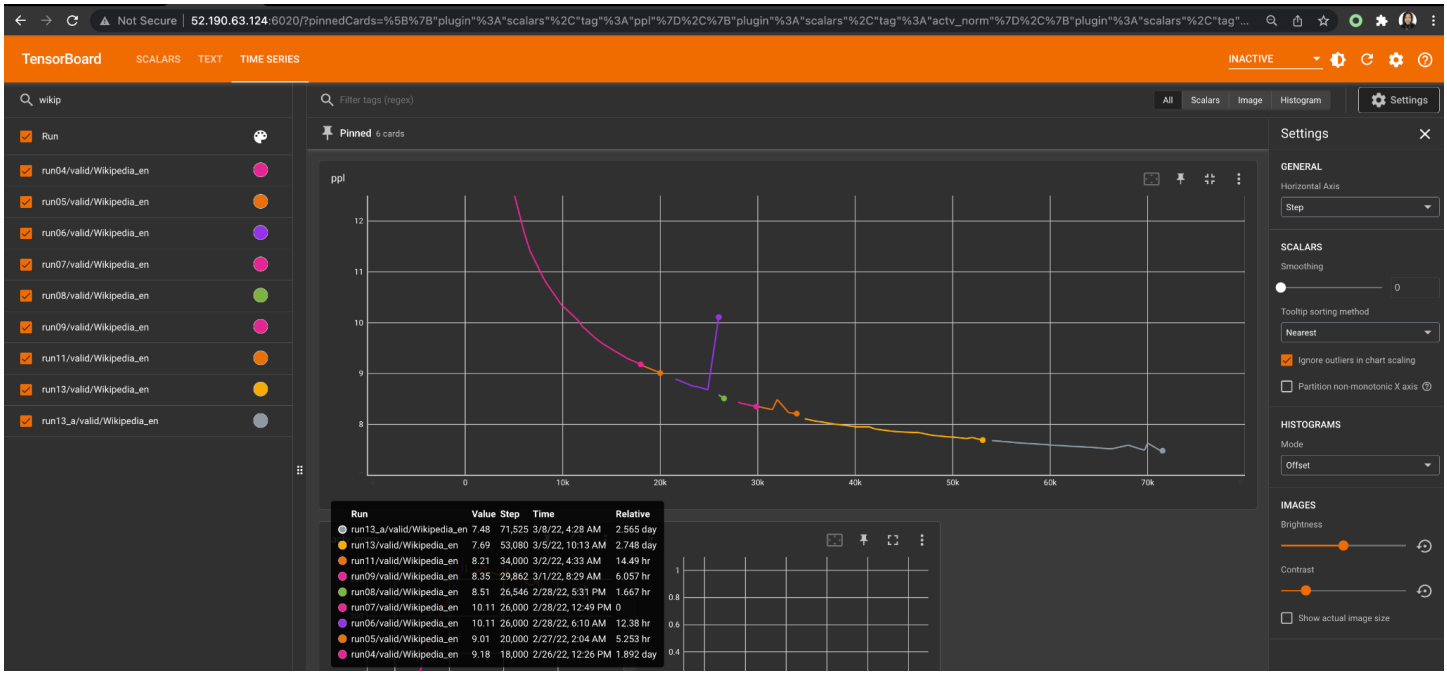
2022-03-11 [Anjali]: 66B Run starts

RUN_ID=66B_run03.16
MP=8 Batch size=2M
'66b': Size(64, 9216, 72, 128, int(2.0 * M), 1.0e-4, 8), # 66b

2021-03-08 [Susan]: 30B run completes



[Tensorboard]



[Tensorboard]

2021-03-01 [Susan]: 30B_run10, restart from 29.5k after CUDA error

- 30B_run09 failed with: RuntimeError: CUDA error: unknown error
- Relaunched from 29500:

```
BLOB_PREFIX="<redacted>/30B_run04"
BLOB_AUTH=???
RESTORE_FILE="${BLOB_PREFIX}/checkpoint_18_29500.pt?${BLOB_AUTH}"
RUN_ID=30B_run10
./<redacted> \
-n 112 -g 8 -t 1 \
-p $RUN_ID \
--model-size 30b \
--checkpoints-dir /shared/home/susanz/checkpoints/30B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--restore-file $RESTORE_FILE \
--full-azure-upload-path "${BLOB_PREFIX}/?${BLOB_AUTH}"
```

- Failed with division by zero? Relaunching from 29000 instead.

```
BLOB_PREFIX="<redacted>/30B_run04"
BLOB_AUTH=???
RESTORE_FILE="${BLOB_PREFIX}/checkpoint_18_29000.pt?${BLOB_AUTH}"
RUN_ID=30B_run11
./<redacted> \
-n 112 -g 8 -t 1 \
-p $RUN_ID \
--model-size 30b \
--checkpoints-dir /shared/home/susanz/checkpoints/30B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--restore-file $RESTORE_FILE \
--full-azure-upload-path "${BLOB_PREFIX}/?${BLOB_AUTH}"
```

2021-02-28 [Susan]: 30B_run07, restart from 25k after grad overflow

- Gradient overflow issues - see loss diverging, activation norm blowing up

Tensorboard

- Restarting from 25k.

```
BLOB_PREFIX="<redacted>/30B_run04"
BLOB_AUTH=???
RESTORE_FILE="${BLOB_PREFIX}/checkpoint_16_25000.pt?${BLOB_AUTH}"
RUN_ID=30B_run07
./<redacted> \
-n 112 -g 8 -t 1 \
-p $RUN_ID \
--model-size 30b \
--checkpoints-dir /shared/home/susanz/checkpoints/30B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--restore-file $RESTORE_FILE \
--full-azure-upload-path "${BLOB_PREFIX}/?${BLOB_AUTH}"
```

- If this blows up in the same spot again, will likely have to lower LR as the next step.
- Blew up in the same place again (~26k steps). Increasing checkpointing to be every 500 steps (instead of 1k), and lowering LR to 8e-5 (from 1e-4).

```
BLOB_PREFIX="<redacted>/30B_run04"
BLOB_AUTH=???
RESTORE_FILE="${BLOB_PREFIX}/checkpoint_16_25000.pt?${BLOB_AUTH}"
RUN_ID=30B_run08
./<redacted> \
-n 112 -g 8 -t 1 \
-p $RUN_ID \
--model-size 30b \
--checkpoints-dir /shared/home/susanz/checkpoints/30B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--restore-file $RESTORE_FILE \
--full-azure-upload-path "${BLOB_PREFIX}/?${BLOB_AUTH}"
```

- Lowering LR got us past 26k!

Tensorboard

2021-02-26 [Susan]: 30B Baseline Restart, 30B_run05, 06

- Run failed due to NCCL errors.
- Fixmyazure catches 75 hosts with 0 IB bandwidth
- Turns out IB isn't detected (UFM not up?).
- Azure folks alerted, fixed with rebooting VMs. 17 busy hosts will need to be rebooted later too, and more diagnostics needed on the 7 hosts in drain (+1 from running fixmyazure afterwards).
- Relaunched:

```
BLOB_PREFIX="<redacted>/30B_run04"
BLOB_AUTH=???
RESTORE_FILE="${BLOB_PREFIX}/checkpoint_11_18000.pt?${BLOB_AUTH}"
RUN_ID=30B_run05
./<redacted> \
-n 112 -g 8 -t 1 \
-p $RUN_ID \
--model-size 30b \
--checkpoints-dir /shared/home/susanz/checkpoints/30B/ \
--local-checkpoints-dir /mnt/scratch/$USER/checkpoints/$(date +%Y-%m-%d).$RUN_ID \
--restore-file $RESTORE_FILE \
--full-azure-upload-path "${BLOB_PREFIX}/?${BLOB_AUTH}"
```

2021-02-24 [Susan]: 30B, 66B Baseline

https://github.com/fairinternal/fairseq-py/pull/3131/files

30B [Azure]: 2x MP, 4M batch size, 112 hosts

- Launching from https://github.com/fairinternal/fairseq-py/tree/susan/30b
- Logs in /shared/home/susanz/checkpoints/30B/ , 2x
    - Got 272k WPS with 2M batch size (run03), 112 hosts
    - Got 418k WPS with 4M batch size (run04), 112 hosts
    - Keeping 4M batch size run: http://52.190.63.124:6020/ (Tensorboard)
        - TFLOP calculator shows only ~117 TFLOPs / GPU utilization - will need to improve: Susan TFLOPs Copy

66B [RSC]: 4x MP, 4M batch size, 128 hosts

- 66B baseline is waiting in queue in the RSC for benchmarking.

2021-01-16 & 2021-01-17 [Stephen]

- Also launched a 350M
- Weirdly, found it was having a really terrible time converging
- Lowered the LR a little bit (to 1e-3) and things went fine

2021-01-07 [Stephen]

- Naman's run crashed, so relaunched from scratch
    - Due to the larger batchsize, **I had to remove some of the validation sets**
    - **So the "combined" ppl is not directly comparable**
- Launched 6.7B with the other 512 GPUs. It's more than fast enough to finish before OSHA run.

2021-01-06 [Stephen]

- Launched 13B on azure with 512 GPUs. Observed it was very slow (25 days to finish)
    - Naman tweaked it and doubled the batchsize to 4B, will be done in 6 days or so

2021-12-26 [Stephen]

- Decided to pause the 6.7B, which was running too slow and isn't very important
- **Launched the 2.7B in its place.**

2021-12-24 [Stephen]

- Managed to resolve all cluster issues by going back to CUDA 11.3
- We'll have to downgrade all the clusters to a new AMI
- Launched 13B and 6.7B. 6.7B was running very slow

2021-12-18 [Stephen]

- Figured out what was up with the tokenizers segfaults
    - Something about using custom compiled NCCL

    - Switched to using the 2.11.4 version released by Nvidia + the CUDA/EFA libraries existing on the cluster.
    - TODO: Update the AWS instructions
    - Also noted that I needed to source a bunch of crap from /etc/profile.d
- Note: Got all my ssh connections suddenly disconnected at one point.
- Launched the 13b baseline
    - Doesn't bode well.
    - A node went down pretty fast before we even finished initializing: Large-25
        - With the one down, not sure 248 gpus is going to be enough to launch the 6.7b baseline.
        - Put it in drain but it wasn't magically replaced like before :(
        - Looks like it came back within an hour or so
    - Tried again and couldn't get past initialization.
        - RuntimeError: CUDA error: CUBLAS_STATUS_ALLOC_FAILED when calling `cublasCreate(handle)
    - It failed a 3rd time so I dropped back to NCCL 2.7.8(?). That's what's printed in the logs, but the DLAMI is supposed to be 2.11.4
        - This is easy to do by dropping LD_PRELOAD
        - That seemed to do it – past initialization
        - But I was getting 0.5x speed compared to num_workers 0.
        - Tried again with setting NCCL_NET_SHARED_BUFFERS=0, no beuno
    - Trying now with the nccl 2.11.4+cuda11.5 from DLAMI
        - This required me to set LD_PRELOAD to their libnccl.so
        - Also refused to start
    - LOL I TRIED COMPILING MYSELF AND WE ARE BACK TO SEGFAULTS
    - Back to DLAMI versions, but with my own compiled aws-ofi-nccl:
        - Off the shelf, segfaults
        - export FI_EFA_USE_DEVICE_RDMA=0 works.

2021-12-10 [Stephen] Set up baselines

- Set up PR for gpt-z baselines: https://github.com/fairinternal/fairseq-py/pull/2781
- Launched 125M for fun