Large Scale Scientific Simulation Systematics GAN

LS4GAN: an CSI/EDG BNL LDRD

Brett Viren

May 7, 2021

Typical application of AI/ML to NPP analyses

Step 1 Train a network using **fake, simulated data**.

Step 2 Apply trained network to infer about **real, detector data**.

Step 3 ???

Step 4 Profit!!!*

* **Profit**: verb. Introduce and ignore systematic uncertainties while still achieving those sweet sweet publications.

Why the concern?

NPP simulations provide good but imperfect models of reality.

- Resulting systematic errors need estimating.
- Not obvious how to do this in train-on-fake/infer-on-real model.

AI/ML compared to "conventional domain-knowledge techniques":

- AI/ML has higher precision, it's why we use it.
  - Cynically: much more precisely does the **wrong thing**.
- Training can prioritize apparently subtle features which are idiosyncratic to the simulation (not in data).
- Inference can misinterpret features that are idiosyncratic to real data.

$\rightarrow$ LS4GAN: use AI/ML techniques to attack this inherently AI/ML problem.

The LS4GAN idea

Translate data samples (events) between domains via GAN

- Generally: $d_i \rightarrow d'_i = \text{ls4gan}(d_i)$, *eg*, simulated $\rightarrow$ real detector.
- Retain **specific features** of each sample but modify **general style**.
  - "Make simulation more like real data while keeping 'truth' features."

Phase 1 exploitation: simulation improvements

- Examine event-by-event $d_i - d'_i$, *eg* "pixel-level" differences.
- Attempt to determine cause and fix/improve the simulation.
- Repeat until reach some point of diminishing returns.

Phase 2 exploitation: propagation of residual systematics

- Send both $\{d_i\}$ and $\{d'_i\}$ through **any** downstream processing
  - Assign difference as a systematic due to imperfect simulation.

Existence proof: CycleGAN

https://arxiv.org/abs/1703.10593

Photograph
("nominal idealistic simulation")

Monet transfer
("more realistic simulation")

Challenges to take artistic GAN styling → LS4GAN

Larger image size, tighter GPU RAM limitations

- Pixels: "cat" photos ≈ 1k × 1k → DUNE event ≈ 10k × 10k
  - Attack with image tiling/fragmenting and/or sparse data techniques.

Scientific data requires objective precision and accuracy

- GAN-styled photo results enjoy friendly subjective human bias: "looks good enough".
  - With scientific data we can not rely on "artistic interpretation" for success.
- NPP real/fake differences are subtle compared to artistic GAN-styled photos.
  - Monet is a brutal artist compared to LArTPC.

Expect larger training sets

- 2PB data / ≈ 8M events from ProtoDUNE Run 1, more expected
  - Dominated by cosmic-μ background, training bias concern.
- Simulation likely provides a bottleneck.
  - Sim GPU acceleration from CSI/EDG via HEP-CCE-PPS work.
  - Requires GPU in production processing facilities → availability and scaling issues.

LS4GAN development plan

- 3 year ("Type A") LDRD in CSI
  - PIs: Meifeng Lin + Yihui (Ray) Ren (CSI)
  - Jin Huang, Haiwang Yu, me (Phys)
  - 2 new post-docs (CSI + EDG)
- First two years: DUNE/LArTPC
  - Develop/validate concept and software tools.
  - Assume focus is systematics of LArTPC detector response.
  - First step: 1D $\rightarrow$ 2D detector response model as "fake" $\rightarrow$ "real"
  - Second step: 2D model $\rightarrow$ real data from single-phase ProtoDUNE
- Third year: sPHENIX
  - Generalize tools/technique to novel domain.
  - Assume focus is jet features vs form factors and detector response.

Related opportunities

LS4GAN LDRD does not cover actually running (big) jobs!

Need to understand and contend with a ratio-of-ratios

$$\frac{\text{Software required CPU/GPU}}{\text{Hardware provided CPU/GPU}}$$

- Example Wire-Cell Toolkit simulation requirement: $\approx \frac{100 \text{ CPU cores}}{1 \text{ GPU card}}$
- Numerator varies job-to-job and during on job's life cycle.

Possible additional areas of development:

- Apply NPPS experience with PanDA for jobs in HPC to run LS4GAN during development and/or eventual production processing.
- Exploit/partner-with BNL facilities (ideally) but also external ones.
- Further the development of Wire-Cell Toolkit support for concurrent distributed GPU sharing and HDF5 file I/O.