Scaling ML: model selection

Let’s pick at least one

Proposed Milestones

1. **Identify target ML models in collaboration with experiments**
2. Port, train, and run at least two target models on two different HPC systems
3. Compare two data parallel training solutions for at least one target model
4. Compare two hyperparameter optimization tools on at least one target model
5. Setting up a prototype Inference as-a-service platform on at least one DOE HPC system

Task list sign up sheet

Email list: CCE-SML@anl.gov

First meeting in January

Selection criteria

- Maturity of model
  - Should be (near) production ready
- Representative
  - Should be a model that is commonly used and addresses needs that are common among experiments and will likely be needed in the long term
- Impactful scaling
  - Scaling the model or performing an HPO will improve the model significantly
- Person power
  - Should be a model with names attached to it. People who want to work on scaling that particular model

2 different scaling tasks:
Inference as a service (IaaS)
And
Training

Possible target ML models

Next step: request repos and input data (but do these really need to be public? If we have enough, e.g., CMS people available to work on a model, isn't that enough?)

Categories:

- **Simulation (training and maybe IaaS)**
  - FastCaloGAN -> a lot of human intervention to make the GANs converge. LBANN has multi-generator, multi-discriminator framework that is only possible with scaling. (asked about publicly available repo)
  - Cosmological simulations, DES adversarial domain adaptation (haven't requested repo)
- **Reconstruction**
  - Flavor tagging (scaling focused on training): models and training framework (which are both mature) are public but training data is not. Is not having training data be public really a problem when we have ATLAS people in HEP-CCE?
  - Tracking (training and IaaS)
  - DUNE reconstruction (training and IaaS) (sent request but no response yet)
- **Analysis (training and IaaS)**
  - Simulation-based inference (haven't requested repo)
  - LSST image processing (haven't requested repo)
- **Resource constrained (FPGA/ASIC) model (training and maybe Iaas) (haven't requested model or repo)**
  - **HPO** is more important vs offline models
  - Size of model vs performance
  - Quantization slows down training
  - Smart pixels (6-layer CNN) takes 3 days (tracking related)