Past Projects

Yuqiong Li
2023/4/14

ML Infrastructure: training

- Maintain and improve a distributed reinforcement learning training system
    - System scale
        - 50+ cloud machines
        - 40+ threads / processes per machine
            - each process hosts a simulator which is a collection of microservices
        - async update to 1 learner
        - training episodes 20k sps * 3-7 days
    - System complexity
        - async update: locks, local files, message queues
        - data decoupling + caching: raw (heterogeneous), processed, replay (weighted sampling)
        - distributed training: deadlock, race condition
        - message passing: protobuf, C++, Python, Cython, Redis, gRPC, pub-sub
        - optimize cpu / memory : fork

ML Infrastructure: training

- Example accomplishments
    - Add new features from the simulator for RL training
        - challenge: maintaining processing speed
        - data compression and decompression - CPython
    - Improve machine utilization
        - tuning replay ratio and queue size
    - Adding metrics to monitor training system health
        - msgs sent / received, # of live threads, steps per second, etc
    - Integrating GCP neural architecture search service with internal training pipeline
        - negavating various storage / network permissions

ML Infra: monitoring and engineering productivity

- Streamline cloud model training and evaluation process using Airflow, Kubernetes and Terraform
- Built an alert system for testing and monitoring model training and quality
  - Devise metrics for system health: training duration, # of live threads, sps, msgs received, etc
  - Early exit mechanisms
  - Notification Bigquery, Dashboard, Emails, Slacks
- Saved $300k per year on ML training expenses
  - Migrate RL pods to share other node pools, stabilize usage with discounted pricing
  - Investigate cluster level cpu utlization and trim cpu requests for offending models

Deep Learning: Object Detection (2019-2020)

- 3D object detection on point cloud
  - PIXOR: https://arxiv.org/abs/1902.06326
  - PointPillar: https://arxiv.org/abs/1812.05784
- Data collection -> parsing -> labeling (open source tool / vendors) -> curation
- Modeling -> training -> deployment -> finetuning
- Tensorflow, TensorRT, C++

Deep Learning: Autoencoder (2020)

- Autoencode and produce fingerprints for buildings
- OpenGL rendering -> panoramic view -> autoencoder
- Some CUDA parallel processing

Deep Learning and Machine Learning (pre-2019)

- 3D generative models on city building shapes
    - Lots of work in data curation and processing: CityGML -> mesh / point cloud / voxel / heightmap
    - For a brief while I was a “power user” of CityGML: my StackOverflow answers , blog post
    - Variational autoencoders, GANs
    - CVPR workshop oral presentation in 2019
- Natural Language Processing (NYU course)
- Undergraduate thesis on using the EM algorithm to solve a mixture Poisson distribution

Misc

- GPU Programming: example cuda code for aforementioned voxelization
  - https://github.com/yuqli/voxRefactor
  - https://github.com/yuqli/vox2dem/blob/master/src/main.cu
- Robotics tools: OpenCV, ROS
- Big data: PySpark, various databases, MapReduce type of parallization algos, MPI