Distributed RL

Richard Liaw
Common Computational Patterns for RL

How can we better utilize our computational resources to accelerate RL progress?
History of large scale distributed RL

2013
- DQN
  Playing Atari with Deep Reinforcement Learning (Mnih 2013)

2015
- GORILA
  Massively Parallel Methods for Deep Reinforcement Learning (Nair 2015)

2016
- A3C
  Asynchronous Methods for Deep Reinforcement Learning (Mnih 2016)

2018
- IMPALA
  IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures (Espeholt 2018)
- Ape-X
  Distributed Prioritized Experience Replay (Horgan 2018)

2019
- R2D3
  Making Efficient Use of Demonstrations to Solve Hard Exploration Problems (Le Paine 2019)
2013/2015: DQN

```python
for i in range(T):
    s, a, s_1, r = evaluate()
    replay.store((s, a, s_1, r))
    minibatch = replay.sample()
    q_network.update(minibatch)
    if should_update_target():
        q_network.sync_with(target_net)
```
2015: General Reinforcement Learning Architecture (GORILA)
GORILA Performance
History of large scale distributed RL

- 2013: DQN
  Playing Atari with Deep Reinforcement Learning (Mnih 2013)

- 2015: GORILA
  Massively Parallel Methods for Deep Reinforcement Learning (Nair 2015)

- 2016: A3C
  Asynchronous Methods for Deep Reinforcement Learning (Mnih 2016)

- 2018: IMPALA
  IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures (Espeholt 2018)

- 2018: Ape-X
  Distributed Prioritized Experience Replay (Horgan 2018)

- 2019: R2D3
  Making Efficient Use of Demonstrations to Solve Hard Exploration Problems (Le Paine 2019)
2016: Asynchronous Advantage Actor Critic (A3C)

# Each worker:

while True:
    sync_weights_from_master()
    for i in range(5):
        collect sample from env
    grad = compute_grad(samples)
    async_send_grad_to_master()

Each has different exploration -> more diverse samples!
A3C Performance

Changes to GORILA:

1. **Faster updates**
2. **Removes** the replay buffer
3. **Moves** to Actor-Critic (from Q learning)

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Time</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQN</td>
<td>8 days on GPU</td>
<td>121.9%</td>
<td>47.5%</td>
</tr>
<tr>
<td>Gorilla</td>
<td>4 days, 100 machines</td>
<td>215.2%</td>
<td>71.3%</td>
</tr>
<tr>
<td>D-DQN</td>
<td>8 days on GPU</td>
<td>332.9%</td>
<td>110.9%</td>
</tr>
<tr>
<td>Dueling D-DQN</td>
<td>8 days on GPU</td>
<td>343.8%</td>
<td>117.1%</td>
</tr>
<tr>
<td>Prioritized DQN</td>
<td>8 days on GPU</td>
<td>463.6%</td>
<td>127.6%</td>
</tr>
<tr>
<td>A3C, FF</td>
<td>1 day on CPU</td>
<td>344.1%</td>
<td>68.2%</td>
</tr>
<tr>
<td>A3C, FF</td>
<td>4 days on CPU</td>
<td>496.8%</td>
<td>116.6%</td>
</tr>
<tr>
<td>A3C, LSTM</td>
<td>4 days on CPU</td>
<td>623.0%</td>
<td>112.6%</td>
</tr>
</tbody>
</table>

*Table 1. Mean and median human-normalized scores on 57 Atari games using the human starts evaluation metric. Supplementary*
Importance Weighted Actor-Learner Architectures (IMPALA)

Motivated by progress in distributed deep learning!
How to correct for Policy Lag? Importance Sampling!

Given an actor-critic model:

1. Apply importance-sampling to policy gradient

\[ \mathbb{E}_{a_s \sim \mu(\cdot|x_s)} \left[ \frac{\pi \bar{\rho}(a_s|x_s)}{\mu(a_s|x_s)} \nabla \log \pi \bar{\rho}(a_s|x_s)q_s|x_s \right] \]

2. Apply importance sampling to critic update

4.1. V-trace target

Consider a trajectory \((x_t, a_t, r_t)_{t=s}^{t=s+n}\) generated by the actor following some policy \(\mu\). We define the \(n\)-steps V-trace target for \(V(x_s)\), our value approximation at state \(x_s\), as:

\[ v_s \overset{\text{def}}{=} V(x_s) + \sum_{t=s}^{s+n-1} \gamma^{t-s} \left( \prod_{i=s}^{t-1} c_i \right) \delta_t V, \quad (1) \]
Scaling Off-Policy learning...

**Ape-X:**
1. Distributed DQN/DDPG/R2D2
2. Reintroduces replay
3. **Distributed Prioritization:** Unlike Prioritized DQN, initial priorities are not set to “max TD”
Ape-X Performance

Figure 2: Left: Atari results aggregated across 57 games, evaluated from random no-op starts. Right: Atari training curves for selected games, against baselines. Blue: Ape-X DQN with 360 actors; Orange: A3C; Purple: Rainbow; Green: DQN. See appendix for longer runs over all games.
With Demonstrations: R2D3 (2019)
Other interesting distributed architectures
Figure 14: Architecture of the QT-Opt distributed reinforcement learning algorithm.

Each model trained on 64 GPUs and 19 parameter servers!
Evolution Strategies

Algorithm 2 Parallelized Evolution Strategies

1: Input: Learning rate $\alpha$, noise standard deviation $\sigma$, initial policy parameters $\theta_0$
2: Initialize: $n$ workers with known random seeds, and initial parameters $\theta_0$
3: for $t = 0, 1, 2, \ldots$ do
4:   for each worker $i = 1, \ldots, n$ do
5:     Sample $\epsilon_i \sim N(0, I)$
6:     Compute returns $F_i = F(\theta_t + \sigma \epsilon_i)$
7:   end for
8: Send all scalar returns $F_i$ from each worker to every other worker
9: for each worker $i = 1, \ldots, n$ do
10:    Reconstruct all perturbations $\epsilon_j$ for $j = 1, \ldots, n$ using known random seeds
11:    Set $\theta_{t+1} \leftarrow \theta_t + \alpha \frac{1}{n \sigma} \sum_{j=1}^{n} F_j \epsilon_j$
12: end for
13: end for
Beyond RL: Population-based Training
Benefits of PBT

RLlib: Abstractions for Distributed Reinforcement Learning (ICML'18)

Eric Liang*, Richard Liaw*, Philipp Moritz, Robert Nishihara, Roy Fox, Ken Goldberg, Joseph E. Gonzalez, Michael I. Jordan, Ion Stoica
RL research scales with compute

Fig. courtesy Nvidia Inc.

Fig. courtesy OpenAI
How do we leverage this hardware?

(a) Supervised Learning

(b) Reinforcement Learning

scalable abstractions for RL?
Systems for RL today

• Many implementations (16000+ repos on GitHub!)
  – how general are they (and do they scale)?
    PPO: multiprocessing, MPI
    Evolution Strategies: Redis
    A3C: shared memory, multiprocessing, TF
    AlphaZero: custom systems
    IMPALA: Distributed TensorFlow

• Huge variety of algorithms and distributed systems used to implement, but little reuse of components
Challenges to reuse

1. Wide range of physical execution strategies for one "algorithm"
Challenges to reuse

2. Tight coupling with deep learning frameworks

Different parallelism paradigms:
- Distributed TensorFlow vs TensorFlow + MPI?
Challenges to reuse

3. Large variety of algorithms with different structures

<table>
<thead>
<tr>
<th>Algorithm Family</th>
<th>Policy Evaluation</th>
<th>Replay Buffer</th>
<th>Gradient-Based Optimizer</th>
<th>Other Distributed Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQNs</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Policy Gradient</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Off-policy PG</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Model-Based/Hybrid</td>
<td>X</td>
<td></td>
<td>X</td>
<td>Model-Based Planning</td>
</tr>
<tr>
<td>Multi-Agent</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Derivative-Free Optimization</td>
</tr>
<tr>
<td>Evolutionary Methods</td>
<td>X</td>
<td></td>
<td>X</td>
<td>MCTS, Derivative-Free Optimization</td>
</tr>
<tr>
<td>AlphaGo</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
We need abstractions for RL

Good abstractions decompose RL algorithms into reusable components.

Goals:
- Code reuse across deep learning frameworks
- Scalable execution of algorithms
- Easily compare and reproduce algorithms
Structure of RL computations

Agent

Policy: state $\rightarrow$ action

Environment

action ($a_{i+1}$)

state ($s_i$) (observation)

reward ($r_i$)
Structure of RL computations

Agent

Policy improvement (e.g., SGD)

Policy evaluation (state → action)

trajectory $X: s_0, (s_1, r_1), \ldots, (s_n, r_n)$

Environment

action ($a_{i+1}$)

state ($s_i$) (observation)

reward ($r_i$)
Many RL loop decompositions

Async DQN (Mnih et al; 2016)

Actor-Learner
Actor-Learner
Actor-Learner

Param Server

\[ X \leftarrow \text{rollout()} \]
\[ d\theta \leftarrow \text{grad}(L, X) \]
\[ \text{sync}(d\theta) \]

Ape-X DQN (Horgan et al; 2018)

Replay
Learner

Actor
Actor
Actor

\[ \theta \leftarrow \text{sync()} \]
\[ \text{rollout()} \]

\[ X \leftarrow \text{replay()} \]
\[ \text{apply}(\text{grad}(L, X)) \]
Common components

Async DQN (Mnih et al; 2016)

- Actor-Learner
- Actor-Learner
- Actor-Learner

- Policy $\pi_\theta(o_t)$
- Trajectory postprocessor $\rho_\theta(X)$
- Loss $L(\theta,X)$

Ape-X DQN (Horgan et al; 2018)

- Replay
- Learner
- Actor
- Actor
- Actor

Policy $\pi_\theta(o_t)$
Trajectory postprocessor $\rho_\theta(X)$
Loss $L(\theta,X)$
Common components

Async DQN (Mnih et al; 2016)

Actor-Learner

Actor-Learner

Actor-Learner

Policy $\pi_\theta(o_t)$

Trajectory postprocessor $\rho_\theta(X)$

Loss $L(\theta,X)$

Ape-X DQN (Horgan et al; 2018)

Replay

Actor

Actor

Actor

Actor

Actor
Structural differences

Async DQN (Mnih et al; 2016)
- Asynchronous optimization
- Replicated workers
- Single machine

Ape-X DQN (Horgan et al; 2018)
- Central learner
- Data queues between components
- Large replay buffers
- Scales to clusters

+ Population-Based Training
  (Jaderberg et al; 2017)
- Nested parallel computations
- Control decisions based on intermediate results

...and this is just one family!

→ No existing system can effectively meet all the varied demands of RL workloads.
Requirements for a new system

Goal: Capture a broad range of RL workloads with high performance and substantial code reuse

1. Support stateful computations
   - e.g., simulators, neural nets, replay buffers
   - big data frameworks, e.g., Spark, are typically stateless

2. Support asynchrony
   - difficult to express in MPI, esp. nested parallelism

3. Allow easy composition of (distributed) components
Ray System Substrate

- RLlib builds on Ray to provide higher-level RL abstractions
- Hierarchical parallel task model with stateful workers
  - flexible enough to capture a broad range of RL workloads (vs specialized sys.)

Hierarchical Task Model

- Single-node
- Cluster
- Synchronous
- Asynchronous
- Send experiences
- Send gradients
- Multiprocessing
- MPI
- Param-server
- GPU
- CPU
Hierarchical Parallel Task Model

1. Create Python class instances in the cluster (stateful workers)
2. Schedule short-running tasks onto workers
   - Challenge: High performance: 1e6+ tasks/s, ~200us task overhead

Top-level worker (Python process)

Sub-worker (process)

Sub-work processes

Sub-sub worker processes

"collect experiences"

"do model-based rollouts"

"run K steps of training"

"allreduce your gradients"

exchange weight shards through Ray object store

RAY Cluster
Unifying system enables RL Abstractions

Policy Optimizer Abstraction

SyncSamples    SyncReplay    AsyncGradients    AsyncSamples    MultiGPU    ...

send experiences

single-node

Policy Graph Abstraction

{s, r, J(θ,X)}

Examples:

Hierarchical Task Model

send gradients

synchronous

CPU

GPU

asynchronous

cluster

send experiences

synchronous

Hierarchical Task Model

http://rllib.io
RLlib Abstractions in Action

**Policy Optimizers**

- SyncSamples
- SyncReplay
- AsyncGradients
- AsyncSamples
- MultiGPU
- ...

**Policy Graphs**

- \{Q-func, n-step, Q-loss\}
- \{LSTM, adv. calc, PG loss\}
- +actor-critic loss, GAE
- +clipped obj.
- +V-trace

- DQN (2015)
- Async DQN (2016)
- Ape-X (2018)

- Policy Gradient (2000)
- A2C (2016)
- A3C (2016)

- PPO (2017)
- PPO (GPU-optimized)
- IMPALA (2018)

http://rllib.io
RLlib Reference Algorithms

- **High-throughput architectures**
  - Distributed Prioritized Experience Replay (Ape-X)
  - Importance Weighted Actor-Learner Architecture (IMPALA)

- **Gradient-based**
  - Advantage Actor-Critic (A2C, A3C)
  - Deep Deterministic Policy Gradients (DDPG)
  - Deep Q Networks (DQN, Rainbow)
  - Policy Gradients
  - Proximal Policy Optimization (PPO)

- **Derivative-free**
  - Augmented Random Search (ARS)
  - Evolution Strategies
Scale your algorithms with RLlib

• Beyond a "collection of algorithms",
• RLlib's abstractions let you easily implement and scale new algorithms (multi-agent, novel losses, architectures, etc)

http://rllib.io
Code example: training PPO

```python
import ray
import ray.rllib.agents.ppo as ppo
from ray.tune.logger import pretty_print

ray.init()
config = ppo.DEFAULT_CONFIG.copy()
config["num_gpus"] = 0
config["num_workers"] = 1
config["eager"] = False
trainer = ppo.PPOTrainer(config=config, env="CartPole-v0")

# Can optionally call trainer.restore(path) to load a checkpoint.

for i in range(1000):
    # Perform one iteration of training the policy with PPO
    result = trainer.train()
    print(pretty_print(result))

    if i % 100 == 0:
        checkpoint = trainer.save()
        print("checkpoint saved at", checkpoint)
```

http://rllib.io
Code example: hyperparam tuning

```python
import ray
import ray.tune as tune

ray.init()
tune.run_experiments({
    "my_experiment": {
        "run": "PPO",
        "env": "CartPole-v0",
        "stop": {"episode_reward_mean": 200},
        "config": {
            "num_gpus": 0,
            "num_workers": 1,
            "sgd_stepsizes": tune.grid_search([0.01, 0.001, 0.0001]),
        },
    },
})
```
Code example: hyperparam tuning

== Status ==
Using FIFO scheduling algorithm.
Resources requested: 4/4 CPUs, 0/0 GPUs
Result logdir: ~/ray_results/my_experiment

PENDING trials:
- PPO_CartPole-v0_2_sgd_stepsize=0.0001: PENDING

RUNNING trials:
- PPO_CartPole-v0_0_sgd_stepsize=0.01: RUNNING [pid=21940], 16 s, 4013 ts, 22 rew
- PPO_CartPole-v0_1_sgd_stepsize=0.001: RUNNING [pid=21942], 27 s, 8111 ts, 54.7 rew
Summary: Ray and RLlib addresses challenges in providing scalable abstractions for reinforcement learning.

RLlib is open source and available at http://rlllib.io
Thanks!
Ray distributed execution engine

- Ray provides **Task parallel** and **Actor** APIs built on **dynamic task graphs**

- These APIs are used to build distributed **applications, libraries** and **systems**
Ray distributed scheduler

• Faster than Python multi-processing on a single node
• Competitive with MPI in many workloads