Real-World Reinforcement Learning

Challenges and Opportunities

Hugo Penedones - DSPT Day - 26th of October 2019, Porto - Portugal

Talk Layout

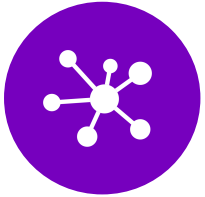
20 minutes

RL refresher

Reinforcement Learning

Applied RL Challenges

Open problems



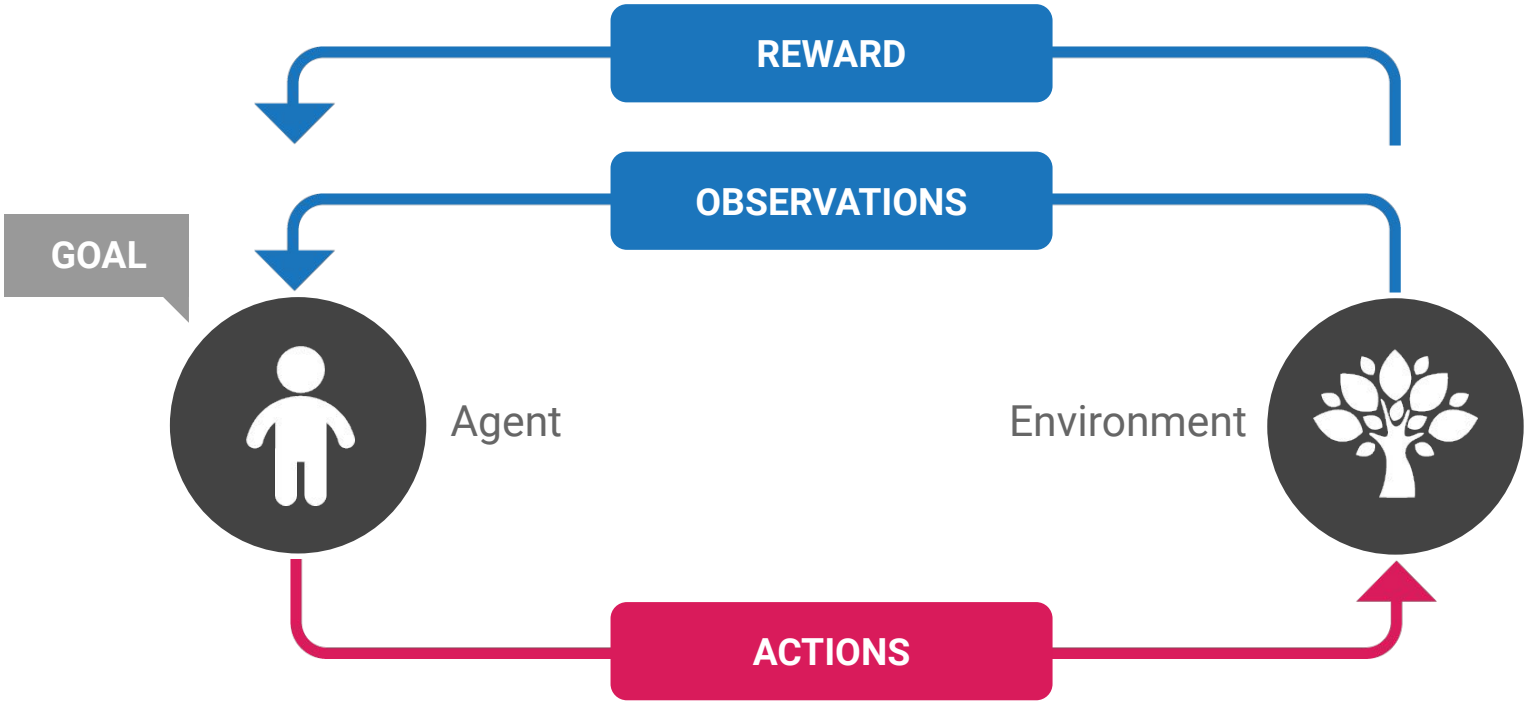
Promising Research directions

What's coming

Image Credit - Wilson Joseph

Reinforcement Learning

The overall setup



Reinforcement Learning

A framework for sequential decision making processes

The goal is to find a **policy** that maximizes the **return**:

Policy $\pi(a|s)$

a function that picks the action to take in a given state

Return $G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

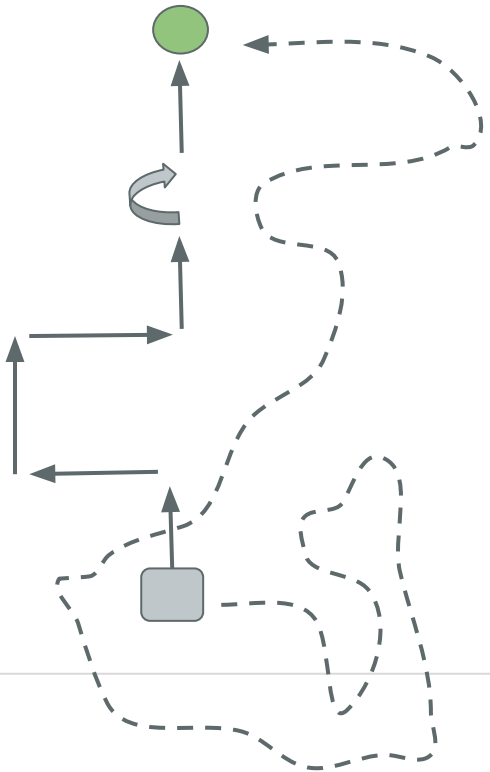
sum of (discounted) rewards into the future.

Supervised Learning vs Reinforcement Learning

Difference in the detail of feedback

Do exactly this:

1. Go forward
2. Turn left
3. Go forward
4. Turn right
5. Go forward
6. Jump
7. Go forward
8. Done, you arrived



Ocasional feedback as you move:

1. Nope… getting cold…
2. Cold…
3. A little warmer…
4. Cold again… Keep trying
5. Warmer…
6. …

…

999. Getting there...

Exploration vs Exploitation

A central problem in RL

EXPLORATION

- Should I try something new?

EXPLOITATION

- Should I stick to what I know?

The RL promise

No need to specify how to solve a problem

1. Specify what you care about

2. Let the system learn and improve itself

   possibly to super-human level!

RL successes in games and simulators

Beyond human-level performance

Atari

Go

DM Lab

Starcraft

DOTA 2

Applications (partial use of RL)

What we really want: solve real world problems!

Recommender Systems

Data Centers cooling

Real World Reinforcement Learning

Challenges ahead

Organisational and People

Engineering and Infrastructure

Open Research problems

Organisational and People factors

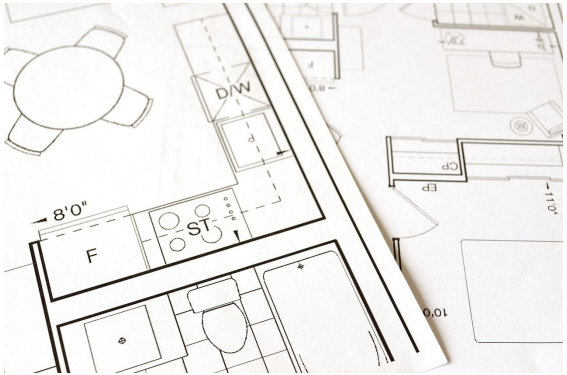
Not all challenges are technical

- Unawareness of the RL potential and opportunity
- Shortage of RL experts
- Risk aversion in experimentation

Engineering and Infrastructure Limitations

The missing basics

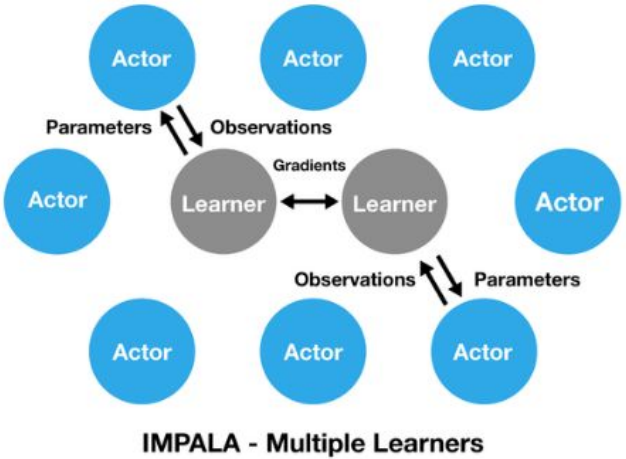
- No standard data logging formats
- No well-established RL libraries / frameworks
- No support for online exploration (e.g. in online recommender systems)



Scalability (not a problem!)

Massively Parallel Methods for Deep Reinforcement Learning

Distributed Training: Parameter Servers, Actors and Learners



IMPALA - Multiple Learners

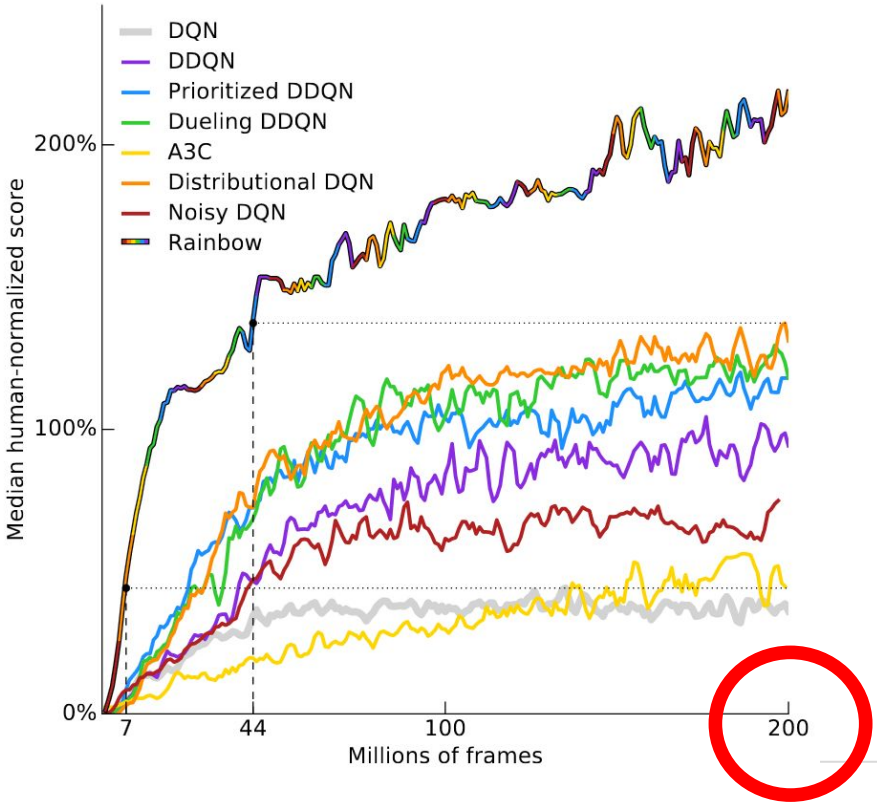
L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, *et al.*, "Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures," *arXiv preprint arXiv:1802.01561*, 2018.

Real World Reinforcement Learning

Challenges ahead

Limited samples

When we can't afford hundreds of millions of steps

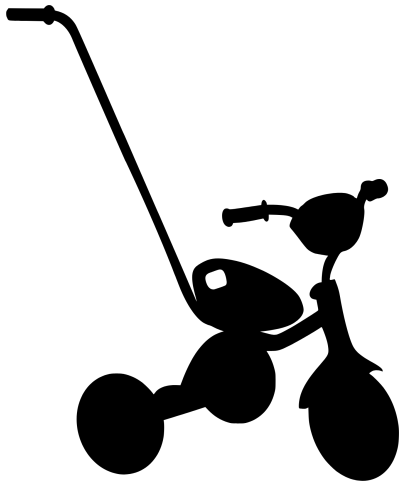


RL algorithms often do millions of interactions with the environment before performing well.

Safe Exploration

Risk awareness

- Epsilon-greedy is a very basic exploration strategy



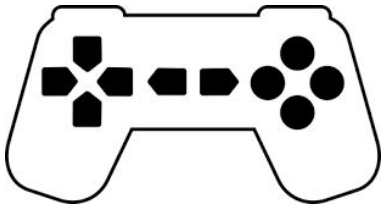
- In real systems, safety or user experience could be compromised during learning

Learning from logs

When no simulator is available

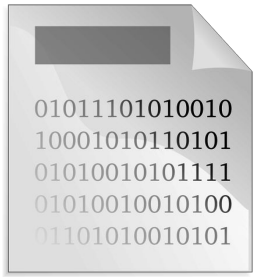
Learning with a simulator

- Online learning
- On-policy data
- Fast feedback loop
- Can explore the state space



Learning from logs

- Offline/Batch learning
- Off-policy data
- Can not explore



Large Action Spaces

Orders of magnitude away

Atari games have up to 18 actions

Go has up to 361 actions

How about a recommender system?

- It can have **millions or billions** of actions!

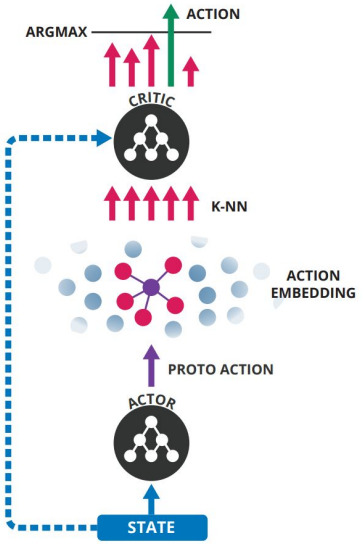


Figure 1. Wolpertinger Architecture

G. Dulac-Arnold, R. Evans, H. van Hasselt, P. Sunehag, T. Lillicrap, J. Hunt, T. Mann, T. Weber, T. Degris, and B. Coppin, "Deep reinforcement learning in large discrete action spaces," *arXiv preprint arXiv:1512.07679*, 2015.

Partial Observability

When we don't see everything

- When the state of the world is not fully observable
- Markovian assumption might not hold

- Need to create a state representation by summarizing many past observations
  (e.g. using recurrent neural net)

Robustness

Policies that work well under perturbations

Real systems suffer from perturbations

Conditions are never the same

RL policies deployed in the real world should be robust

Modeling Uncertainty and Explainability

Known unknowns

Most of our predictions encompass some level of uncertainty

What if we knew how confident we are in our predictions?

Can the system decisions be understood by humans?

Promising Research Directions

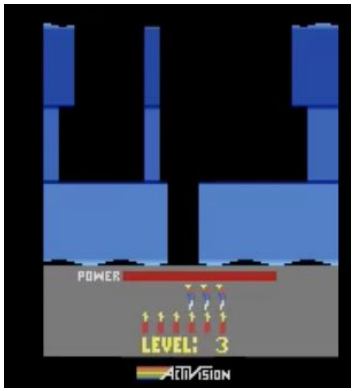
Learning from Expert Demonstrations

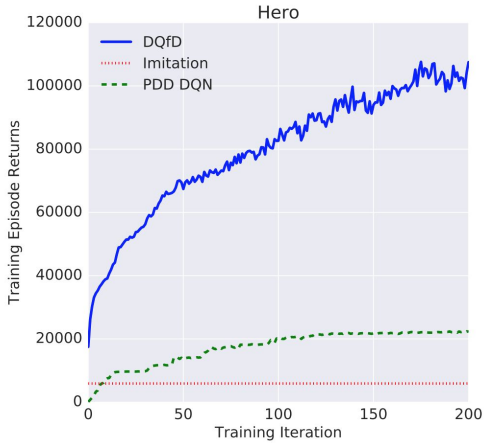
How not to start from scratch

A human expert plays each Atari game and the actions are recorded

Adding a supervised term in the standard RL loss function

You can learn from an expert, and **then** surpass it



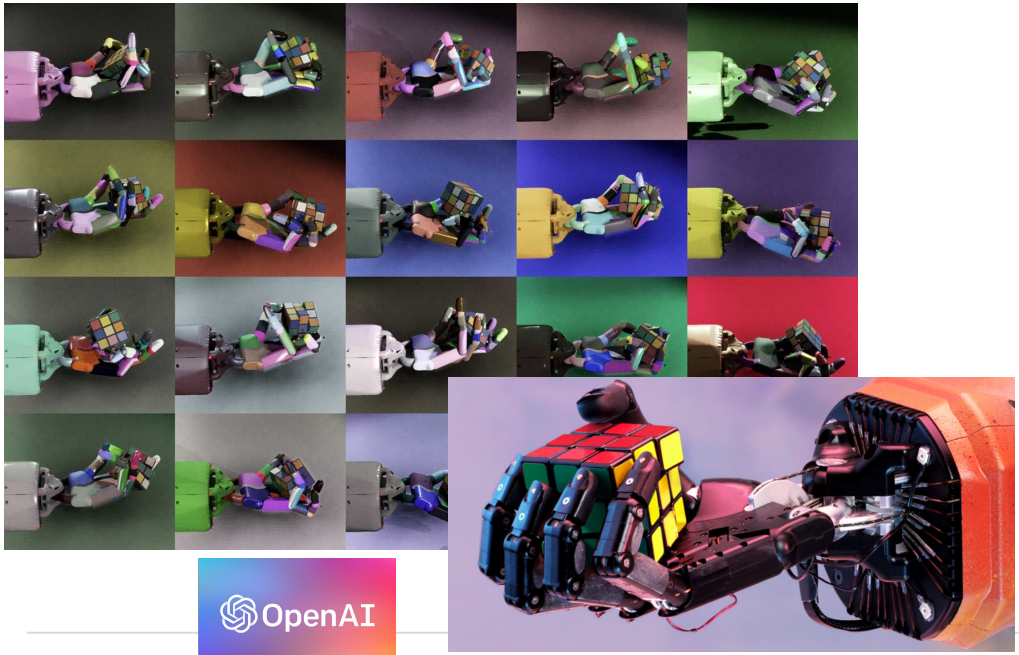


T. Hester, M. Vecerik, O. Pietquin, M. Lanctot, T. Schaul, B. Piot, D. Horgan, J. Quan, A. Sendonaris, G. Dulac-Arnold, *et al.*, "Learning from demonstrations for real world reinforcement learning. arxiv," 2017.

Sim2Real

Bridging the reality gap

Domain Randomization



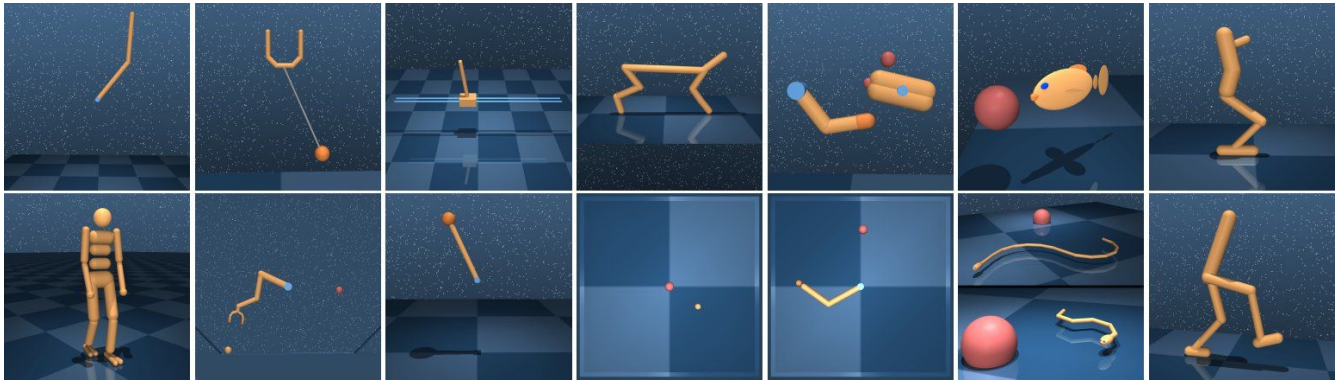
Model Fine tuning

Continual Learning

(e.g Progressive Networks)

Model-based RL

Learning the environment dynamics



The “Holy Grail” of reinforcement learning

Planning, Sample Efficiency, Task adaptation, etc. become “easy”

Research challenge: how to use inaccurate models?

Real-World Reinforcement Learning is coming

Reinforcement Learning can go beyond human level performance

Until now, successes have been mostly restricted to games

Multiple promising directions to address open research problems

THANK YOU

Before I came here, I was confused about this subject. Having listened to your lecture, I am still confused -- but on a higher level.

Enrico Fermi