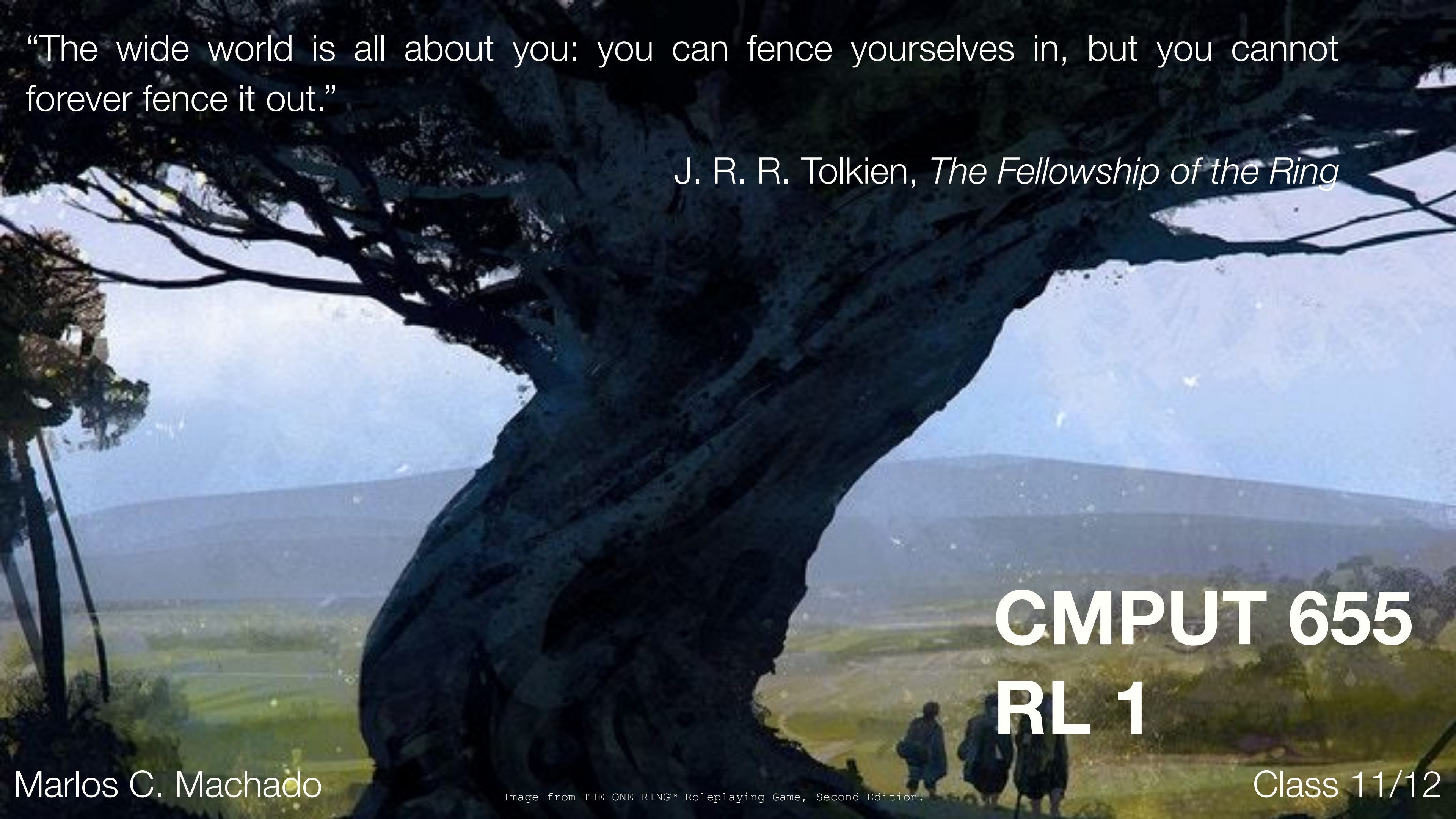


"The wide world is all about you: you can fence yourselves in, but you cannot forever fence it out."

J. R. R. Tolkien, *The Fellowship of the Ring*

CMPUT 655
RL 1

Marlos C. Machado

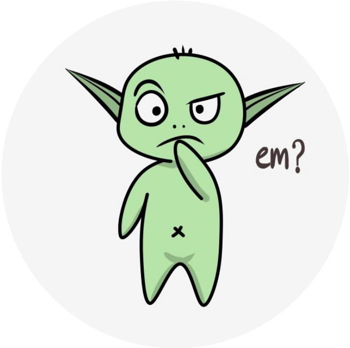
Image from THE ONE RING™ Roleplaying Game, Second Edition.

Class 11/12

Reminders I

- We'll have Rich Sutton as guest lecturer today at 15:30.
- The final Project Report is due on December 15th.
    - The link on eClass is already open.
    - I cannot accept late submissions.
- The final grades (best 9) for the Coursera activities are *almost* available on eClass.
- The Student Perspectives of Teaching (SPOT) Survey is now available.

Please, interrupt me at any time!



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Last Class (Recorded)

Chapter 13

Policy Gradient Methods

Deep Reinforcement Learning

Chapter 16

Applications and Case Studies

-ish

Many High-profile Stories in RL are due to Deep RL

Human-level control through deep reinforcement learning

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare,

Many High-profile Stories in RL are due to Deep RL


Startups

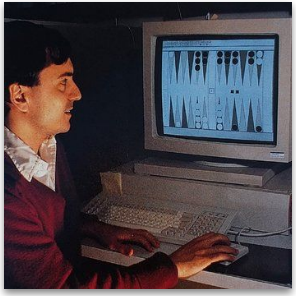
Google Acquires Artificial Intelligence Startup DeepMind For More Than $500M

Catherine Shu @catherineshu / 6:20 PM MST • January 26, 2014

Join TechCrunch+
Login
Search
Comment

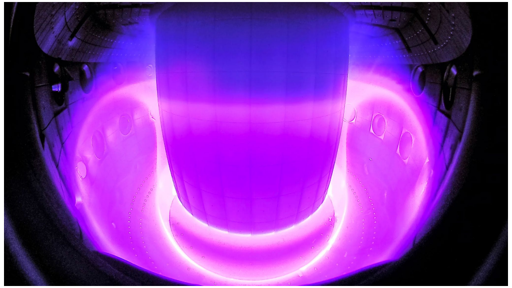
Many High-profile Stories in RL are due to Deep RL

*[Tesauro, 1994]

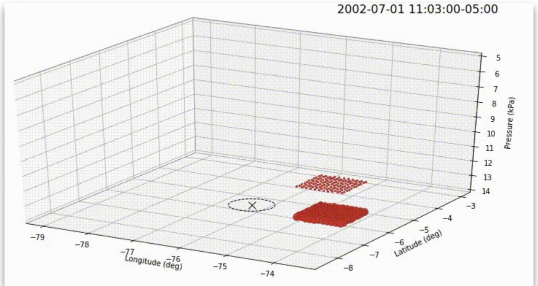


[Mnih et al., 2015]

[Degrave et al., 2022]



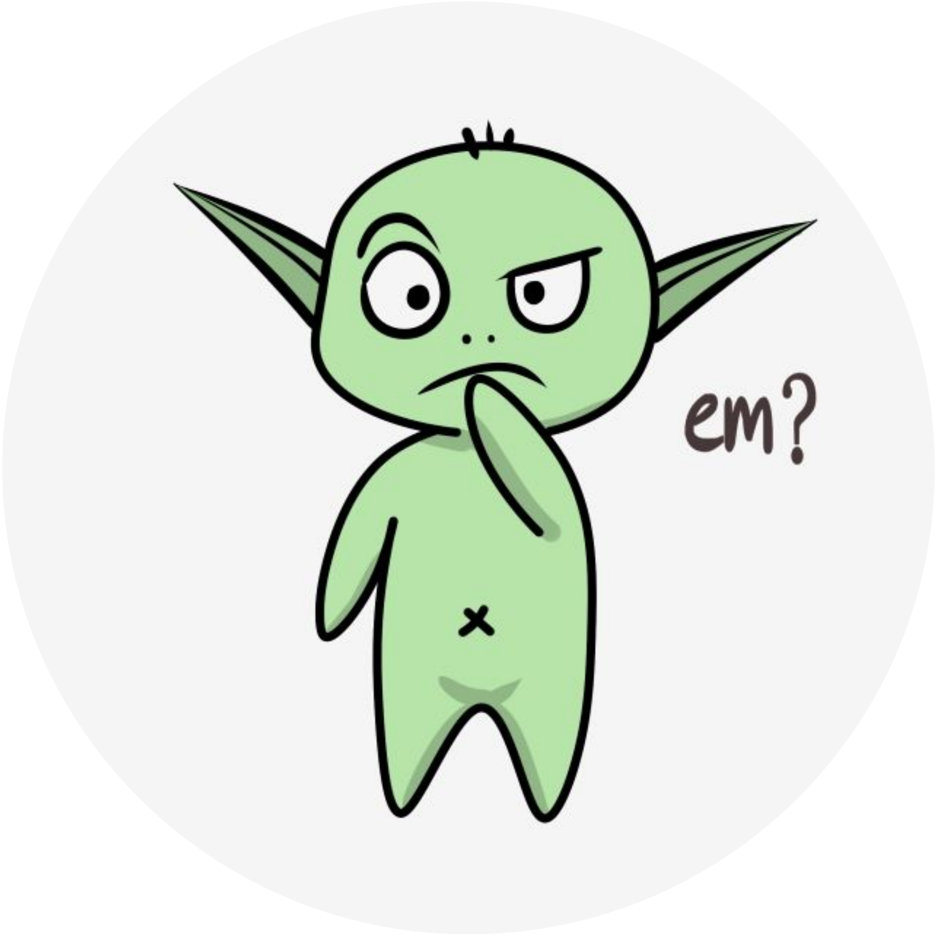
[Vinyals et al., 2019]



[Bellemare et al., 2020]

Deep Reinforcement Learning

- Deep RL is more than neural networks + RL.
  It is actually designing algorithms *to use* deep learning.
- Similar to RL, deep RL is simultaneously a problem, a class of solution methods, and the field that studies this problem and its solution methods.
  - Deep RL studies control (and prediction) problems in which a computational agent learns to make decisions. It focuses on problems with high-dimensional observations.
  - The solution methods in deep RL consist of using (at least) a neural network for either value function or policy approximation. Ideally, these solution methods are general in sense of being applicable to a wide range of problems.
  - Given how general reinforcement learning is, it can definitely be seen as a specific subset of RL.
    - We still need to deal with the same problems, generalization, credit assignment, and exploration.



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Some history

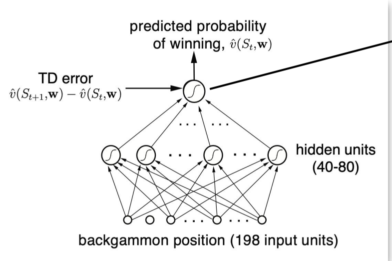
RL + Neural Networks: TD Gammon

[Tesauro, 1992, 1994, 1995]

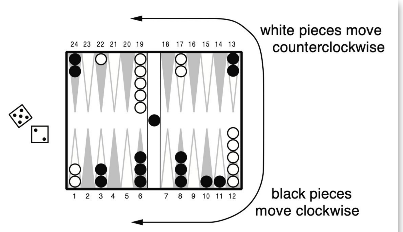
- A straightforward combination of the TD(λ) algorithm and nonlinear function approximation using a multilayer artificial neural network to play backgammon.

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\Big[R_{t+1} + \gamma\hat{v}(S_{t+1},\mathbf{w}_t) - \hat{v}(S_t,\mathbf{w}_t)\Big]\mathbf{z}_t \qquad \mathbf{z}_t \doteq \gamma\lambda\mathbf{z}_{t-1} + \nabla\hat{v}(S_t,\mathbf{w}_t)$$

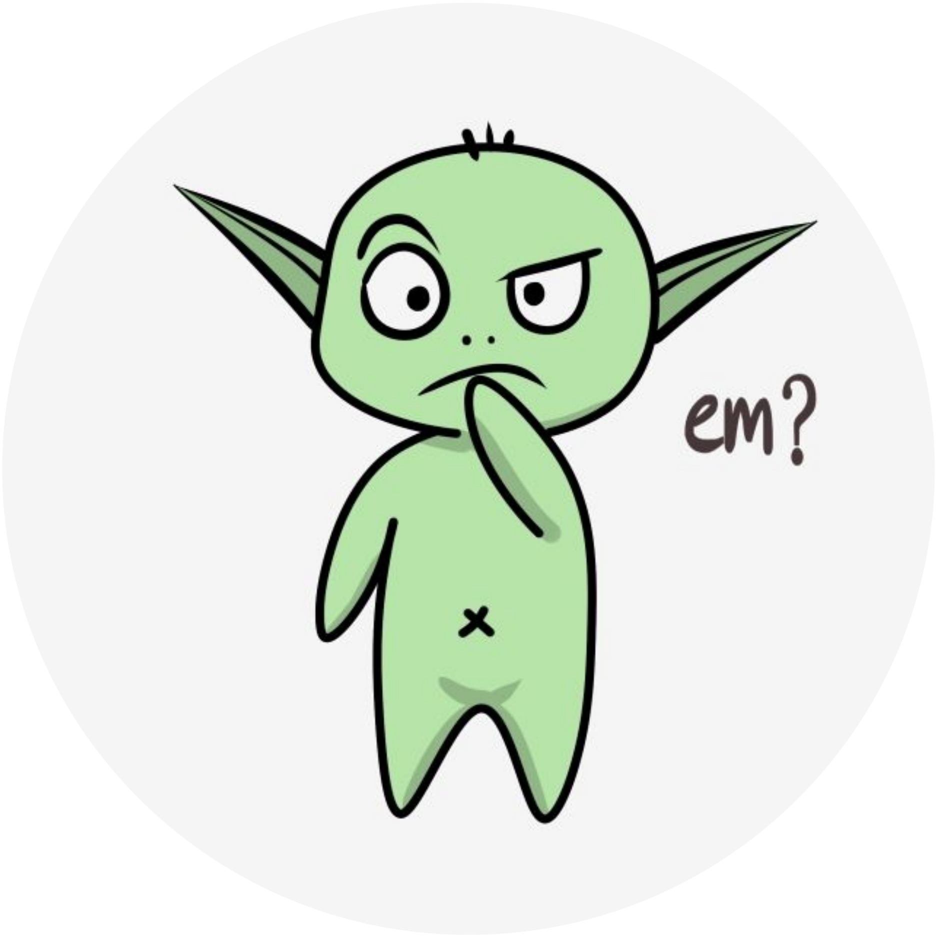
- Backgammon has too many positions and an effective branch. factor of about 400.
- Self-play for data generation.
- It had many versions:
  - v0: straightforward input with little domain knowledge.
  - v1: added specialized backgammon features.
  - v2 / v2.1: bigger network (40 and then 80) and selective two-ply search.
  - v3 / v3.1: bigger network (160) and selective three-ply search.



Sigmoid: $h(j) = \sigma\left(\sum_i w_{ij}x_i\right) = \frac{1}{1+e^{-\sum_i w_{ij}x_i}}$



A backgammon position



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Journal of Artificial Intelligence Research 47 (2013) 253–279 Submitted 02/13; published 06/13

The Arcade Learning Environment: An Evaluation Platform for General Agents

Marc G. Bellemare MG17@CS.UALBERTA.CA
University of Alberta, Edmonton, Alberta, Canada

Yavar Naddaf YAVAR@EMPIRICALRESULTS.CA
Empirical Results Inc., Vancouver,
British Columbia, Canada

Joel Veness VENESS@CS.UALBERTA.CA
Michael Bowling BOWLING@CS.UALBERTA.CA
University of Alberta, Edmonton, Alberta, Canada

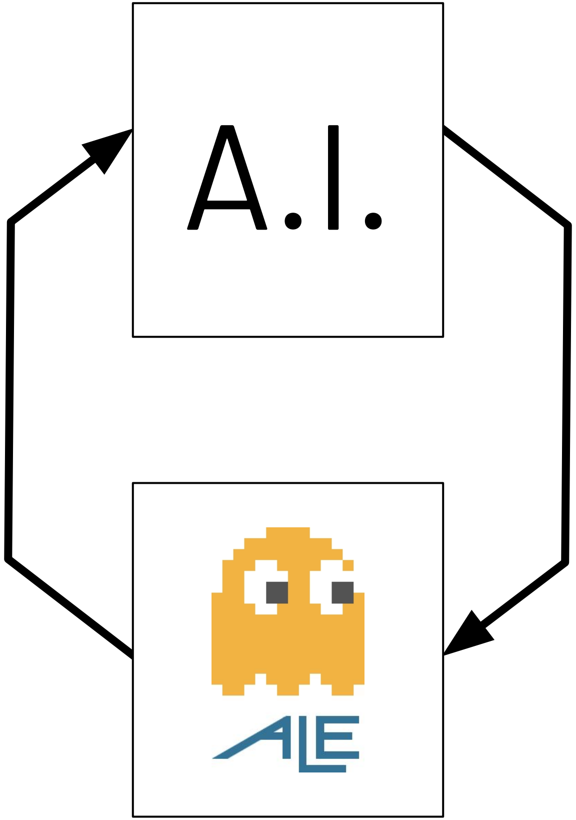
Abstract

In this article we introduce the Arcade Learning Environment (ALE): both a challenge problem and a platform and methodology for evaluating the development of general, domain-independent AI technology. ALE provides an interface to hundreds of Atari 2600 game environments, each one different, interesting, and designed to be a challenge for human players. ALE presents significant research challenges for reinforcement learning, model learning, model-based planning, imitation learning, transfer learning, and intrinsic motivation. Most importantly, it provides a rigorous testbed for evaluating and comparing approaches to these problems. We illustrate the promise of ALE by developing and benchmarking domain-independent agents designed using well-established AI techniques for both reinforcement learning and planning. In doing so, we also propose an evaluation

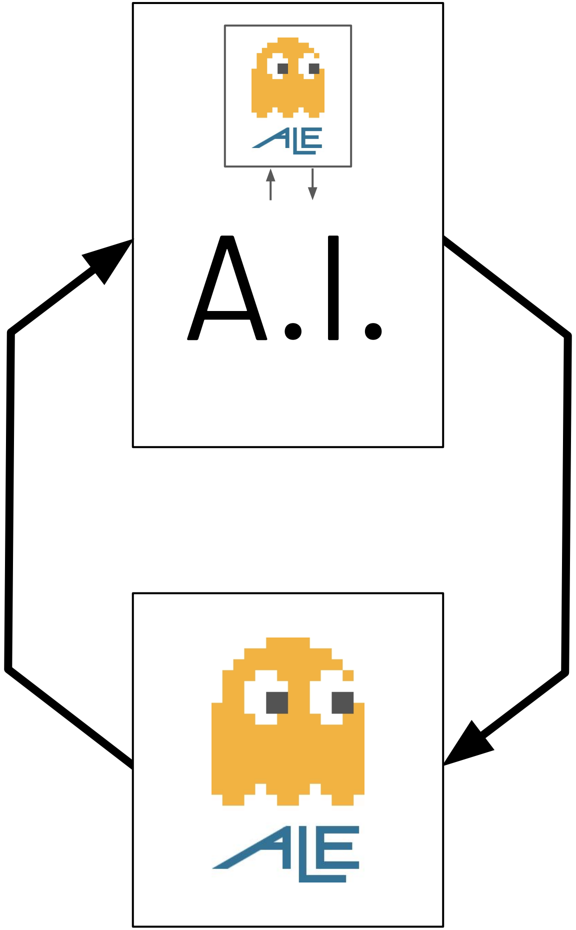
Arcade Learning Environment

Over 50 Domains in 8 Minutes 23 Seconds

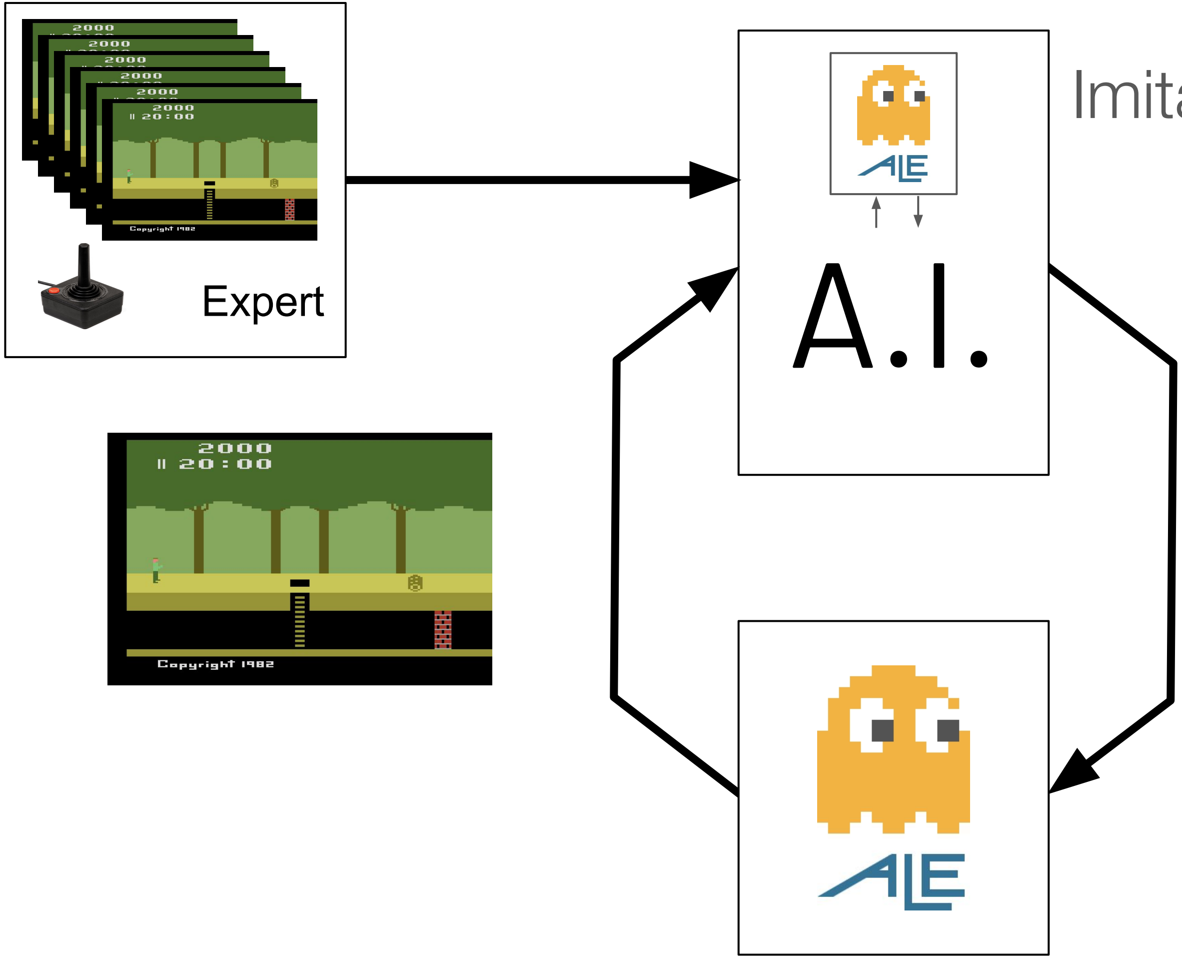
Reinforcement Learning



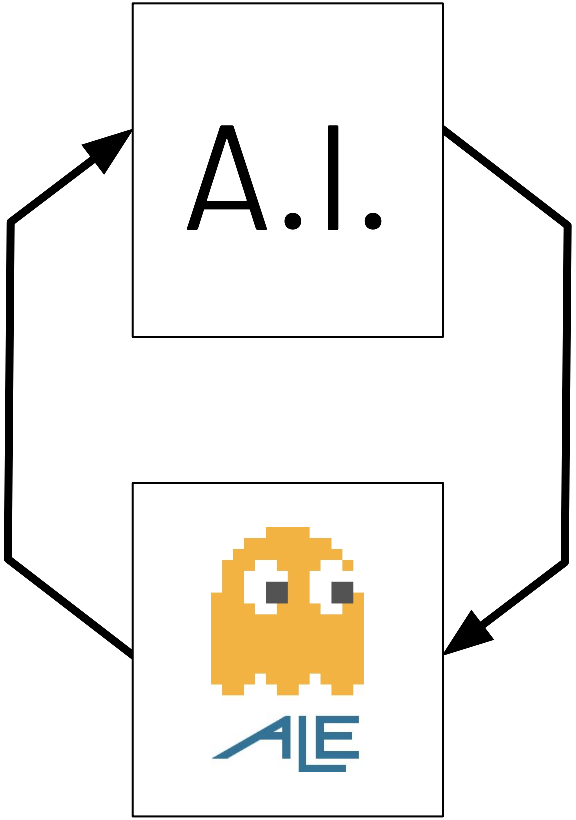
Model Learning



Imitation/Apprenticeship Learning

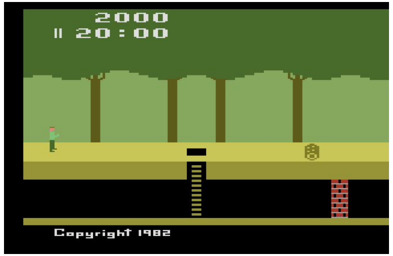


Exploration

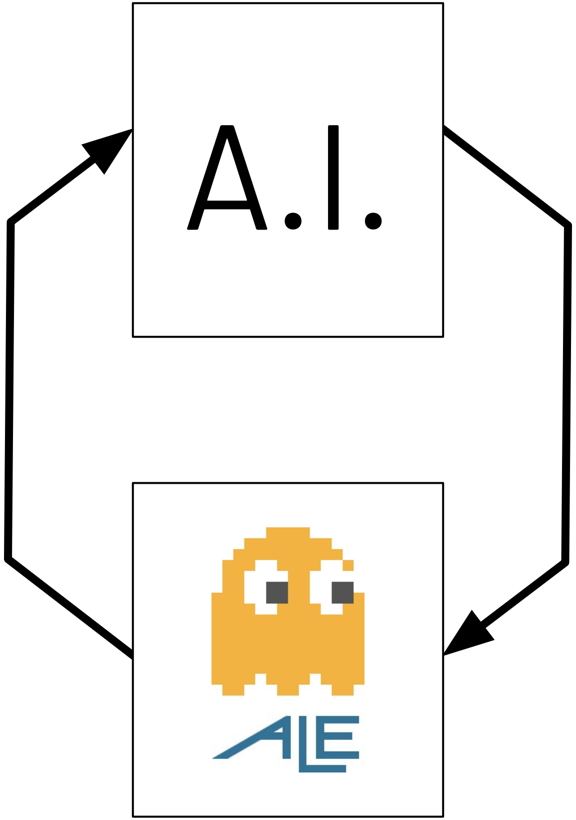


Transfer Learning

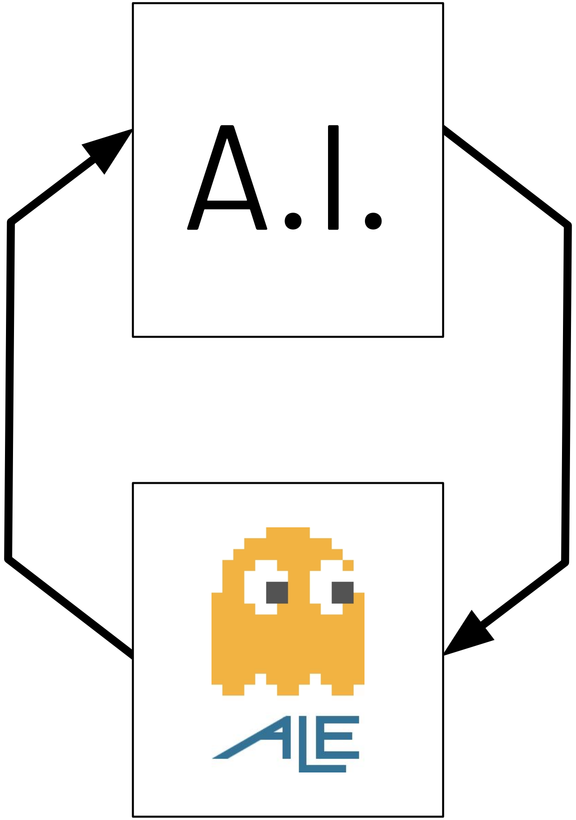
Pitfall!



Pitfall II



Intrinsic Motivation



Deep Q-Network (and Deep RL)
[Mnih et al., 2013, 2015]

Playing Atari with Deep Reinforcement Learning

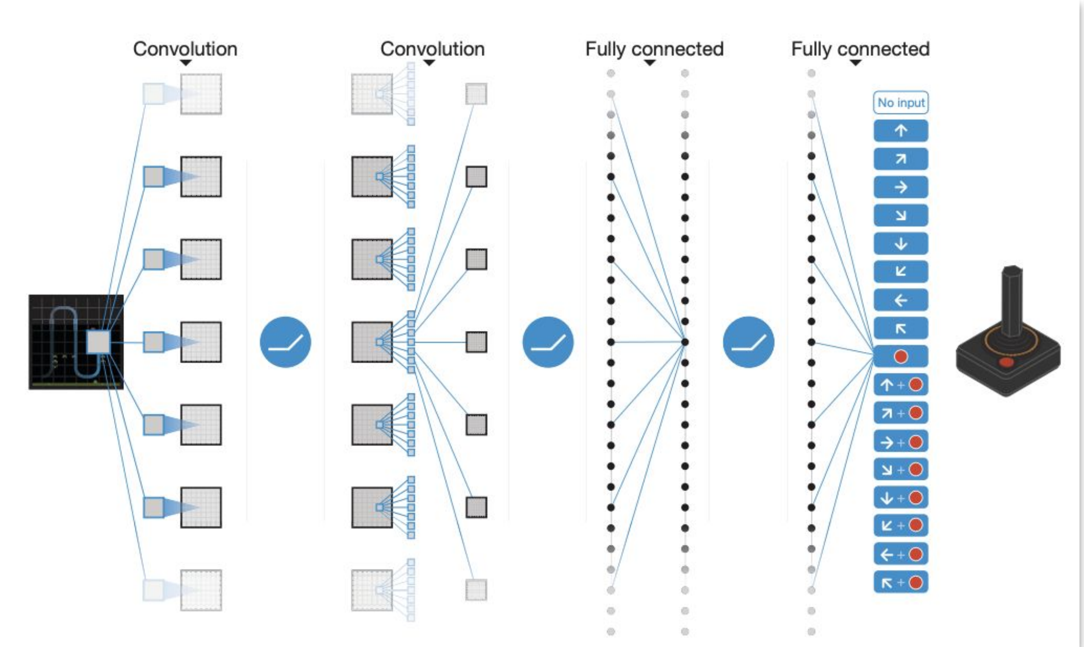
Volodymyr Mnih Koray Kavukcuoglu David Silver Alex Graves Ioannis Antonoglou
Daan Wierstra Martin Riedmiller
DeepMind Technologies
{vlad,koray,david,alex.graves,ioannis,daan,martin.riedmiller} @ deepmind.com

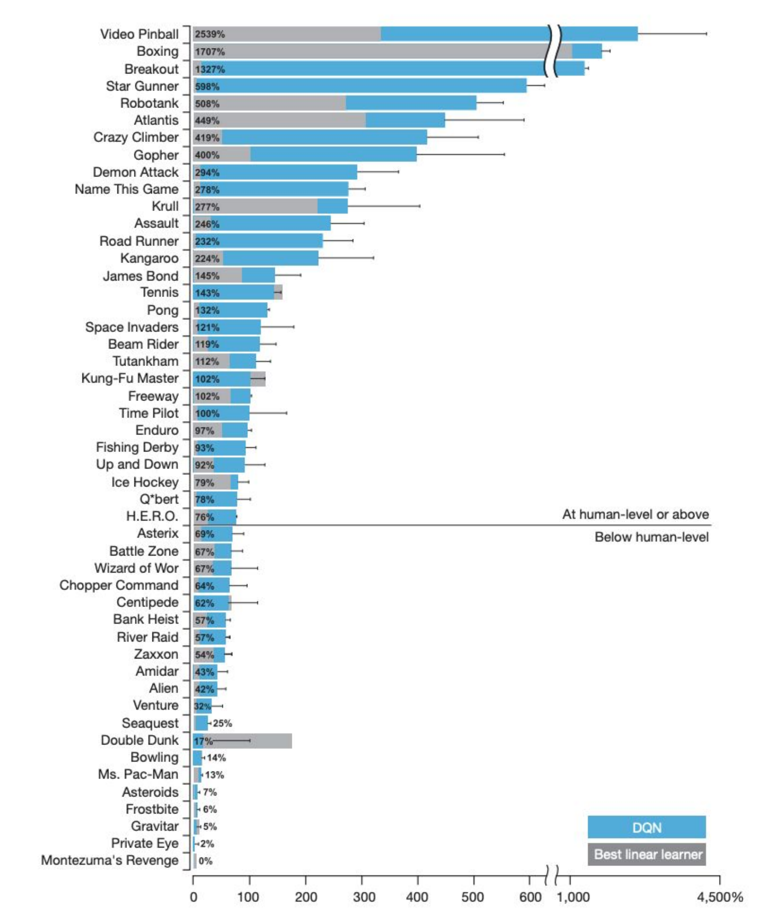
Dec 2013

Abstract

Deep Q-Network (and Deep RL)

[Mnih et al., 2013, 2015]





Deep Q-Network (DQN)

[Mnih et al., 2013, 2015]

Stacked frames

-1, 0, +1 rewards

Clipped error term

$$\mathcal{L}^{\mathrm{DQN}} = \mathbb{E}_{(s,a,r,s')\sim U(\mathcal{D})}\left[\left(R_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q(S_{t+1}, a'; \boldsymbol{\theta}^{-}) - Q(S_t, A_t, \boldsymbol{\theta}_t)\right)^2\right]$$

Experience replay buffer (Lin, 1993)
Original size: 1M frames

Target network
Original update frequency: 10k

ε decay
Originally, from 1.0 to 0.1 over 1M frames

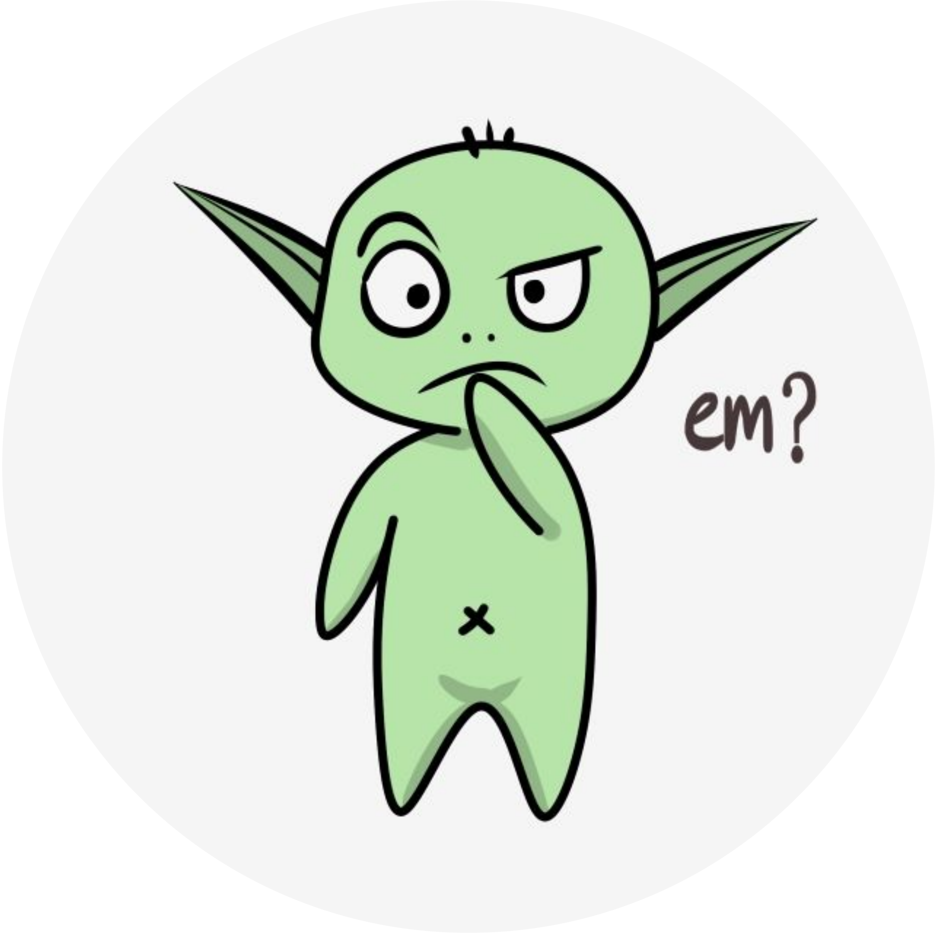
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha\left[R_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q(S_{t+1}, a'; \boldsymbol{\theta}^{-}) - Q(S_t, A_t, \boldsymbol{\theta}_t)\right]\nabla_{\boldsymbol{\theta}_t} Q(S_t, A_t; \boldsymbol{\theta}_t)$$

RMSProp

Deep Q-Network (DQN)

[Mnih et al., 2013, 2015]

Game	Random Play	Best Linear Learner	Contingency (SARSA)	Human	DQN (± std)	Normalized DQN (% Human)
Alien	227.8	939.2	103.2	6875	3069 (±1093)	42.7%
Amidar	5.8	103.4	183.6	1676	739.5 (±3024)	43.9%
Assault	222.4	628	537	1496	3359(±775)	246.2%
Asterix	210	987.3	1332	8503	6012 (±1744)	70.0%
Asteroids	719.1	907.3	89	13157	1629 (±542)	7.3%
Atlantis	12850	62687	852.9	29028	85641(±17600)	449.9%
Bank Heist	14.2	190.8	67.4	734.4	429.7 (±650)	57.7%
Battle Zone	2360	15820	16.2	37800	26300 (±7725)	67.6%
Beam Rider	363.9	929.4	1743	5775	6846 (±1619)	119.8%
Bowling	23.1	43.9	36.4	154.8	42.4 (±88)	14.7%



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Deep Q-Network (DQN)

[Mnih et al., 2013, 2015]

Game	Random Play	Best Linear Learner	Contingency (SARSA)	Human	DQN (± std)	Normalized DQN (% Human)
Alien	227.8	939.2	103.2	6875	3069 (±1093)	42.7%
Amidar	5.8	103.4	183.6	1676	739.5 (±3024)	43.9%
Assault	222.4	628	537	1496	3359(±775)	246.2%
Asterix	210	987.3	1332	8503	6012 (±1744)	70.0%
Asteroids	719.1	907.3	89	13157	1629 (±542)	7.3%
Atlantis	12850	62687	852.9	29028	85641(±17600)	449.9%
Bank Heist	14.2	190.8	67.4	734.4	429.7 (±650)	57.7%
Battle Zone	2360	15820	16.2	37800	26300 (±7725)	67.6%
Beam Rider	363.9	929.4	1743	5775	6846 (±1619)	119.8%
Bowling	23.1	43.9	36.4	154.8	42.4 (±88)	14.7%

Tables can be misleading

[Machado et al., 2018]

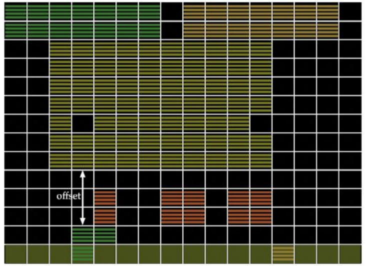
- Tables imply an apples-to-apples comparison, even when they are not:
- DQN saw much more data than the baselines.
- DQN measured its performance differently than the baselines.
- DQN used domain knowledge other baselines didn't:
    - Lives signal
    - Action set

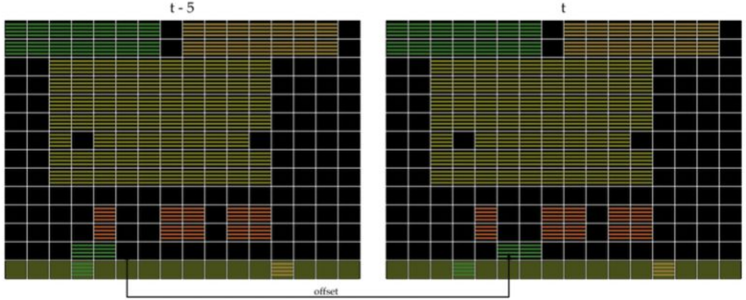
This can be a big deal

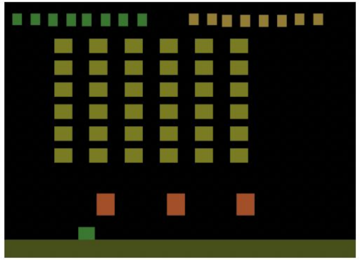
[Liang et al., 2016]

State of the Art Control of Atari Games Using Shallow Reinforcement Learning

Yitao Liang†, Marlos C. Machado‡, Erik Talvitie†, and Michael Bowling‡
†Franklin & Marshall College
Lancaster, PA, USA
{yliang, erik.talvitie}@fandm.edu
‡University of Alberta
Edmonton, AB, Canada
{machado, mbowling}@ualberta.ca



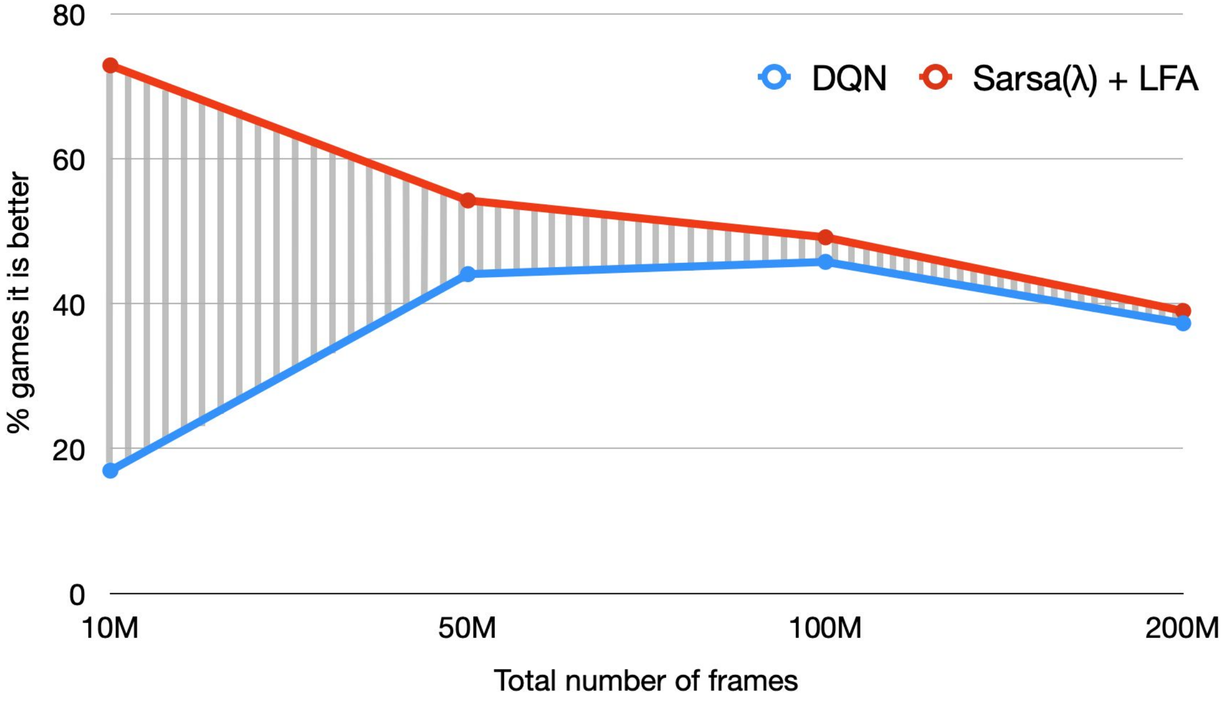


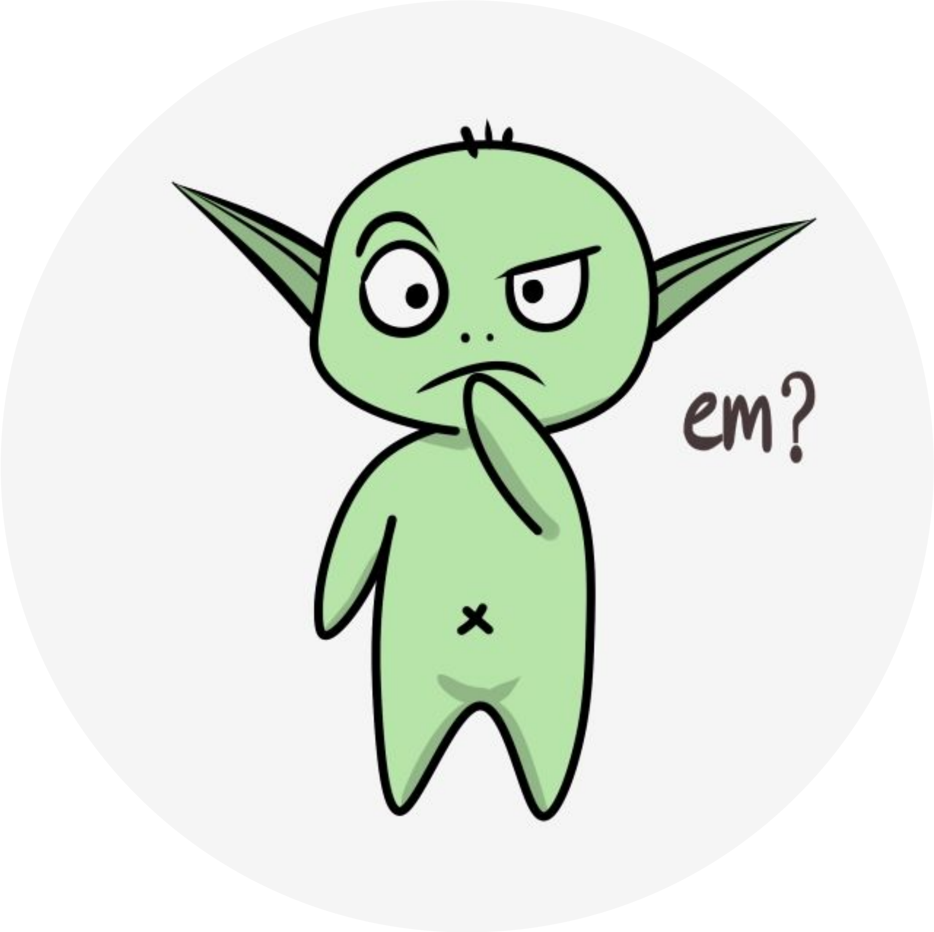


It is not that we should be using linear function approxim.

[Liang et al., 2016; Machado et al. 2018]

but we should understand and be careful with our claims





https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Deep Q-Network (DQN)

[Mnih et al., 2013, 2015]

Stacked frames

-1, 0, +1 rewards

Clipped error term

$$\mathcal{L}^{\text{DQN}} = \mathbb{E}_{(s,a,r,s')\sim U(\mathcal{D})}\left[\left(R_{t+1} + \gamma \max_{a'\in\mathcal{A}} Q(S_{t+1}, a'; \boldsymbol{\theta}^{-}) - Q(S_t, A_t, \boldsymbol{\theta}_t)\right)^2\right]$$

Experience replay buffer (Lin, 1993)
Original size: 1M frames

Target network
Original update frequency: 10k

ε decay
Originally, from 1.0 to 0.1 over 1M frames

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha\left[R_{t+1} + \gamma \max_{a'\in\mathcal{A}} Q(S_{t+1}, a'; \boldsymbol{\theta}^{-}) - Q(S_t, A_t, \boldsymbol{\theta}_t)\right]\nabla_{\boldsymbol{\theta}_t} Q(S_t, A_t; \boldsymbol{\theta}_t)$$

RMSProp

To be continued...