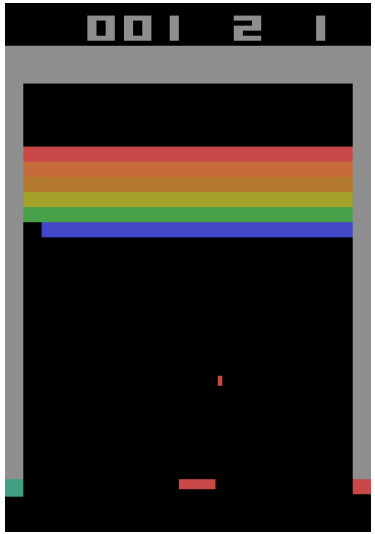
Benchmarking Deep Reinforcement Learning for Continuous Control

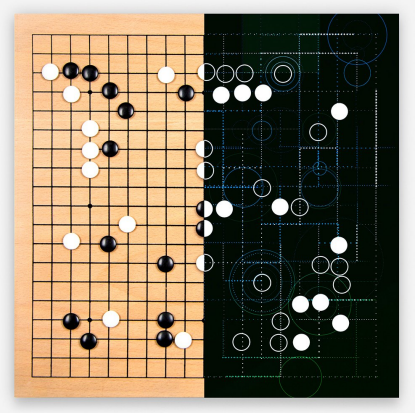
Yan (Rocky) Duan[†], Xi Chen[†], Rein Houthooft[†‡], John Schulman[†§], Pieter Abbeel[†]

[†]Berkeley Artificial Intelligence Research (BAIR) laboratory, [‡]Ghent University, [§]OpenAI

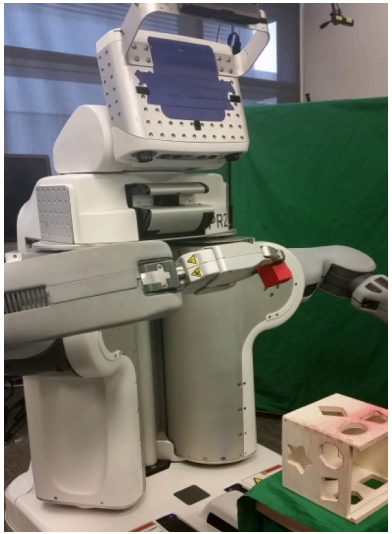
Deep Reinforcement Learning



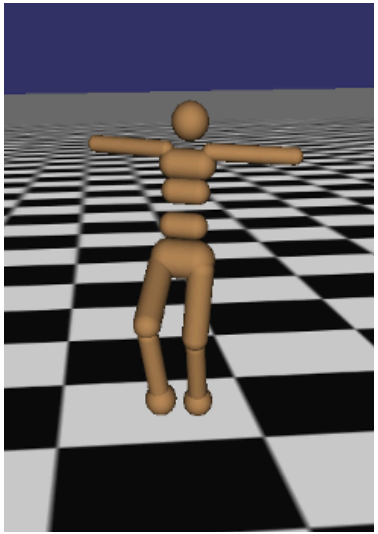
[Mnih et al. 2013; Guo et al. 2014; Mnih et al. 2015; Oh et al. 2015; Nair et al. 2015; Van Hasselt et al. 2015; Schulman et al. 2015; Parisotto et al. 2015; Wang et al. 2015; Kulkarni et al. 2016]



[Silver et al. 2016; Tian et al. 2016; Maddison et al. 2014; Clark et al. 2015]



[Levine et al. 2015; Finn et al. 2016; Watter et al. 2015; Tzeng et al. 2015]



[Schulman et al. 2015; Lillicrap et al. 2015; Heess et al. 2015; Gu et al. 2016]

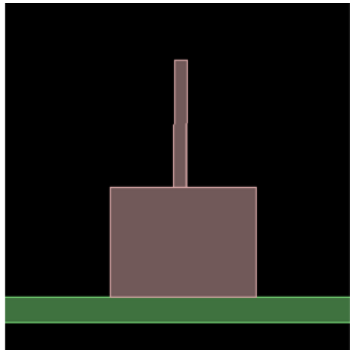
Deep Reinforcement Learning

- High-dimensional state space + discrete actions:
  - Arcade Learning Environment (ALE) [Bellemare et al., 2012], Minecraft [Tessler et al., 2016; Abel et al., 2016; Oh et al., 2016]…
- High-dimensional continuous actions
  - ?

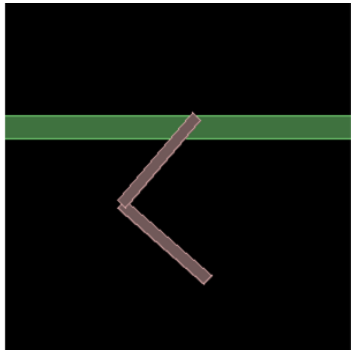
Proposed Benchmark

- 5 classic tasks
- 7 locomotion tasks
- 15 partially observable tasks
- 4 hierarchical tasks
- Also: reference implementations of 8 algorithms
- ***Everything open source***

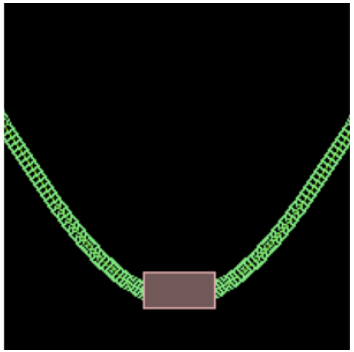
Classic Tasks



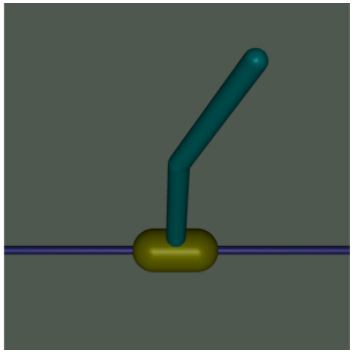
Cart Pole Balancing
+
Inverted Pendulum



Acrobot

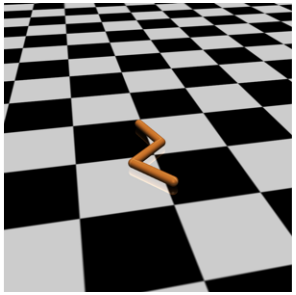


Mountain Car

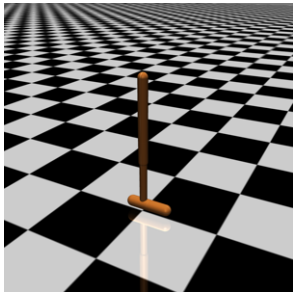


Double Inverted Pendulum

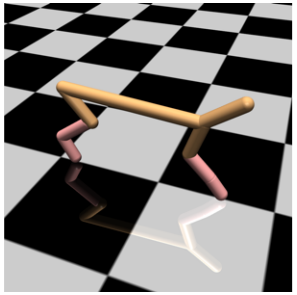
Locomotion Tasks



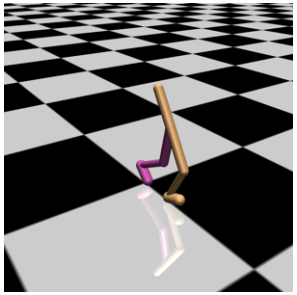
Swimmer



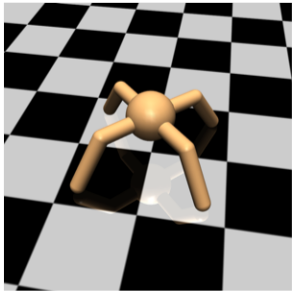
Hopper



Half Cheetah



Walker



Ant

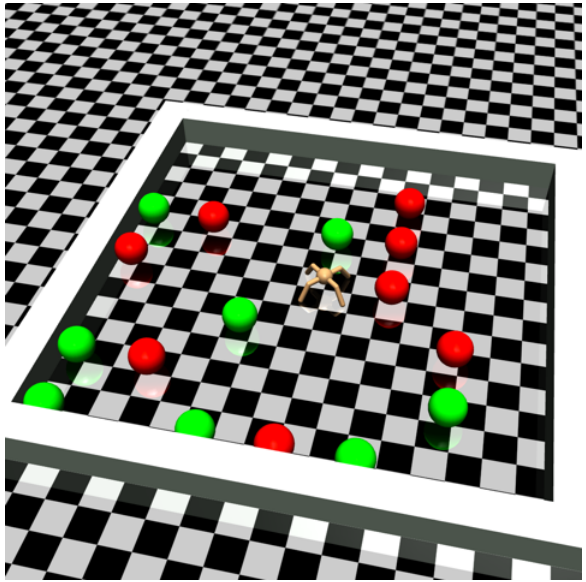
Simplified Humanoid

Full Humanoid

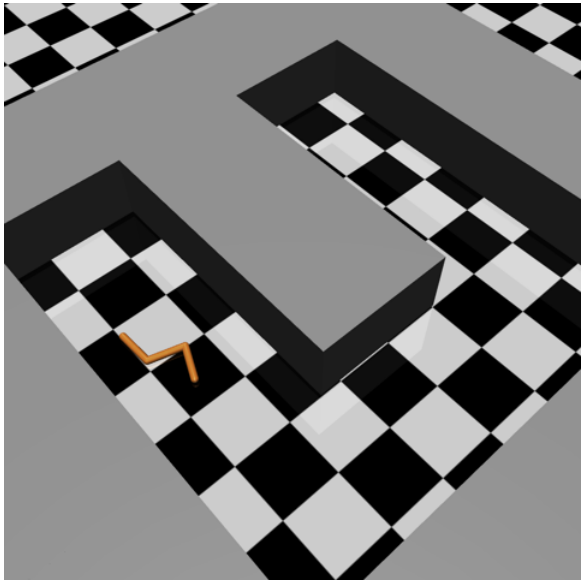
Partially Observable Tasks

- Limited Sensors (LS)
- Noisy Observations and Delayed Actions (NO)
- System Identification (SI)

Hierarchical Tasks



Ant + Gather
(Also: Swimmer + Gather)



Swimmer + Maze
(Also: Ant + Maze)

Algorithms

- REINFORCE [Williams, 1992] + ADAM [Kingma & Ba, 2014]
- TNPG [Kakade, 2002; Peters et al., 2003; Bagnell & Schneider, 2003; Schulman et al., 2015]
- RWR [Peters & Schaal, 2007; Kober & Peters, 2009]
- REPS [Peters et al., 2010]
- TRPO [Schulman et al. 2015]
- CEM [Rubinstein, 1999; Szita & Lorincz, 2006]
- CMA-ES [Hansen & Ostermeier, 2001]
- DDPG [Silver et al., 2014; Lillicrap et al., 2015]

Batch, Gradient-Based Algorithms

- REINFORCE [Williams, 1992] + ADAM [Kingma & Ba, 2014]

$$\hat{g} \leftarrow \text{PolicyGradient}$$
$$\Delta\theta \leftarrow \text{AdamUpdate}(\theta, \hat{g})$$

- NPG [Kakade, 2002; Peters et al., 2003; Bagnell & Schneider, 2003]

$$\Delta\theta \propto \hat{F}(\theta)^{-1}\hat{g}$$

- TNPG [Schulman et al., 2015] Compute descent direction via truncated conjugate gradient.

- TRPO [Schulman et al., 2015] Solve constrained optimization problem

$$\Delta\theta = \arg\max_{\Delta\theta} \eta(\theta + \Delta\theta)$$
$$\text{s.t. } D_{KL}(\pi_\theta \| \pi_{\theta+\Delta\theta}) \leq \delta$$

Batch, Gradient-Based Algorithms

- RWR [Peters & Schaal, 2007; Kober & Peters, 2009]

Formulated as Expectation-Maximization problem.

- REPS [Peters et al., 2010]

Limits the loss of information in the state-action distribution.
Formulated as dual optimization problem.

Batch, Gradient-Free Algorithms

- CEM [Rubinstein, 1999; Szita & Lorincz, 2006]

Black-box algorithm which maintains and updates a factored Gaussian distribution over the parameters.

- CMA-ES [Hansen & Ostermeier, 2001]

Similar to CEM, but maintains a full covariance matrix.

Online Algorithms

- DDPG [Silver et al., 2014; Lillicrap et al., 2015]

$$\Delta\theta \propto \left.\frac{\partial}{\partial a}\hat{Q}^{\pi}(s,a)\right|_{a=\pi_\theta(s)} \frac{\partial}{\partial\theta}\pi_\theta(s)$$

Policy Representation

- All Markov environments
  - 3-layer MLP, 100-50-25 units with relu nonlinearity
- Partially observable tasks
  - Single-layer LSTM with 32 hidden units
- DDPG: same architecture as in [Lillicrap et al. 2015]

Results

Task	Random	REINFORCE	TNPG	RWR	REPS	TRPO	CEM	CMA-ES	DDPG
Cart-Pole Balancing	77.1 ± 0.0	4693.7 ± 14.0	**3986.4 ± 748.9**	**4861.5 ± 12.3**	565.6 ± 137.6	**4869.8 ± 37.6**	4815.4 ± 4.8	2440.4 ± 568.3	4634.4 ± 87.8
Inverted Pendulum*	−153.4 ± 0.2	13.4 ± 18.0	**209.7 ± 55.5**	84.7 ± 13.8	−113.3 ± 4.6	**247.2 ± 76.1**	38.2 ± 25.7	−40.1 ± 5.7	40.0 ± 244.6
Mountain Car	−415.4 ± 0.0	−67.1 ± 1.0	**-66.5 ± 4.5**	−79.4 ± 1.1	−275.6 ± 166.3	**-61.7 ± 0.9**	−66.0 ± 2.4	−85.0 ± 7.7	−288.4 ± 170.3
Acrobot	−1904.5 ± 1.0	−508.1 ± 91.0	−395.8 ± 121.2	−352.7 ± 35.9	−1001.5 ± 10.8	−326.0 ± 24.4	−436.8 ± 14.7	−785.6 ± 13.1	**-223.6 ± 5.8**
Double Inverted Pendulum*	149.7 ± 0.1	4116.5 ± 65.2	**4455.4 ± 37.6**	3614.8 ± 368.1	446.7 ± 114.8	**4412.4 ± 50.4**	2566.2 ± 178.9	1576.1 ± 51.3	2863.4 ± 154.0
Swimmer*	−1.7 ± 0.1	92.3 ± 0.1	**96.0 ± 0.2**	60.7 ± 5.5	3.8 ± 3.3	**96.0 ± 0.2**	68.8 ± 2.4	64.9 ± 1.4	85.8 ± 1.8
Hopper	8.4 ± 0.0	714.0 ± 29.3	**1155.1 ± 57.9**	553.2 ± 71.0	86.7 ± 17.6	**1183.3 ± 150.0**	63.1 ± 7.8	20.3 ± 14.3	267.1 ± 43.5
2D Walker	−1.7 ± 0.0	506.5 ± 78.8	**1382.6 ± 108.2**	136.0 ± 15.9	−37.0 ± 38.1	**1353.8 ± 85.0**	84.5 ± 19.2	77.1 ± 24.3	318.4 ± 181.6
Half-Cheetah	−90.8 ± 0.3	1183.1 ± 69.2	**1729.5 ± 184.6**	376.1 ± 28.2	34.5 ± 38.0	**1914.0 ± 120.1**	330.4 ± 274.8	441.3 ± 107.6	**2148.6 ± 702.7**
Ant*	13.4 ± 0.7	548.3 ± 55.5	**706.0 ± 127.7**	37.6 ± 3.1	39.0 ± 9.8	**730.2 ± 61.3**	49.2 ± 5.9	17.8 ± 15.5	326.2 ± 20.8
Simple Humanoid	41.5 ± 0.2	128.1 ± 34.0	**255.0 ± 24.5**	93.3 ± 17.4	28.3 ± 4.7	**269.7 ± 40.3**	60.6 ± 12.9	28.7 ± 3.9	99.4 ± 28.1
Full Humanoid	13.2 ± 0.1	262.2 ± 10.5	**288.4 ± 25.2**	46.7 ± 5.6	41.7 ± 6.1	**287.0 ± 23.4**	36.9 ± 2.9	N/A ± N/A	119.0 ± 31.2
Cart-Pole Balancing (LS)*	77.1 ± 0.0	420.9 ± 265.5	**945.1 ± 27.8**	68.9 ± 1.5	898.1 ± 22.1	**960.2 ± 46.0**	227.0 ± 223.0	68.0 ± 1.6	
Inverted Pendulum (LS)	−122.1 ± 0.1	−13.4 ± 3.2	**0.7 ± 6.1**	−107.4 ± 0.2	−87.2 ± 8.0	**4.5 ± 4.1**	−81.2 ± 33.2	−62.4 ± 3.4	
Mountain Car (LS)	−83.0 ± 0.0	−81.2 ± 0.6	**-65.7 ± 9.0**	−81.7 ± 0.1	−82.6 ± 0.4	**-64.2 ± 9.5**	**-68.9 ± 1.3**	**-73.2 ± 0.6**	
Acrobot (LS)*	−393.2 ± 0.0	−128.9 ± 11.6	**-84.6 ± 2.9**	−235.9 ± 5.3	−379.5 ± 1.4	**-83.3 ± 9.9**	−149.5 ± 15.3	−159.9 ± 7.5	
Cart-Pole Balancing (NO)*	101.4 ± 0.1	616.0 ± 210.8	**916.3 ± 23.0**	93.8 ± 1.2	99.6 ± 7.2	606.2 ± 122.2	181.4 ± 32.1	104.4 ± 16.0	
Inverted Pendulum (NO)	−122.2 ± 0.1	6.5 ± 1.1	**11.5 ± 0.5**	−110.0 ± 1.4	−119.3 ± 4.2	**10.4 ± 2.2**	−55.6 ± 16.7	−80.3 ± 2.8	
Mountain Car (NO)	−83.0 ± 0.0	−74.7 ± 7.8	**-64.5 ± 8.6**	−81.7 ± 0.1	−82.9 ± 0.1	**-60.2 ± 2.0**	−67.4 ± 1.4	−73.5 ± 0.5	
Acrobot (NO)*	−393.5 ± 0.0	**-186.7 ± 31.3**	**-164.5 ± 13.4**	−233.1 ± 0.4	−258.5 ± 14.0	**-149.6 ± 8.6**	−213.4 ± 6.3	−236.6 ± 6.2	
Cart-Pole Balancing (SI)*	76.3 ± 0.1	431.7 ± 274.1	**980.5 ± 7.3**	69.0 ± 2.8	702.4 ± 196.4	**980.3 ± 5.1**	746.6 ± 93.2	71.6 ± 2.9	
Inverted Pendulum (SI)	−121.8 ± 0.2	−5.3 ± 5.6	**14.8 ± 1.7**	−108.7 ± 4.7	−92.8 ± 23.9	**14.1 ± 0.9**	−51.8 ± 10.6	−63.1 ± 4.8	
Mountain Car (SI)	−82.7 ± 0.0	−63.9 ± 0.2	**-61.8 ± 0.4**	−81.4 ± 0.1	−80.7 ± 2.3	**-61.6 ± 0.4**	−63.9 ± 1.0	−66.9 ± 0.6	
Acrobot (SI)*	−387.8 ± 1.0	**-169.1 ± 32.3**	**-156.6 ± 38.9**	−233.2 ± 2.6	−216.1 ± 7.7	**-170.9 ± 40.3**	−250.2 ± 13.7	−245.0 ± 5.5	
Swimmer + Gathering	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Ant + Gathering	−5.8 ± 5.0	−0.1 ± 0.1	−0.4 ± 0.1	−5.5 ± 0.5	−6.7 ± 0.7	−0.4 ± 0.0	−4.7 ± 0.7	N/A ± N/A	−0.3 ± 0.3
Swimmer + Maze	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Ant + Maze	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	N/A ± N/A	0.0 ± 0.0

Results

	TNPG	TRPO
Inverted Pendulum	**209.7±55.5**	**247.2±76.1**
Mountain Car	**-66.5±4.5**	**-61.7±0.9**
Swimmer	**96±0.2**	**96±0.2**
Half-Cheetah	**1729.5±184.6**	**1914±120.1**
Cart-Pole Balancing (LS)	**945.1±27.8**	**960.2±46**
Mountain Car (NO)	**-64.5±8.6**	**-60.2±2**
Inverted Pendulum (SI)	**14.8±1.7**	**14.1±0.9**

Results

	REINFORCE	TRPO
Cart-Pole Balancing	4693.7±14	**4869.8±37.6**
Inverted Pendulum	**13.4±18**	**247.2±76.1**
Mountain Car	-67.1±1	**-61.7±0.9**
Double Inverted Pendulum	4116.5±65.2	**4412.4±50.4**
Swimmer	92.3±0.1	**96±0.2**
Hopper	**714±29.3**	**1183.3±150**
Half-Cheetah	**1183.1±69.2**	**1914±120.1**

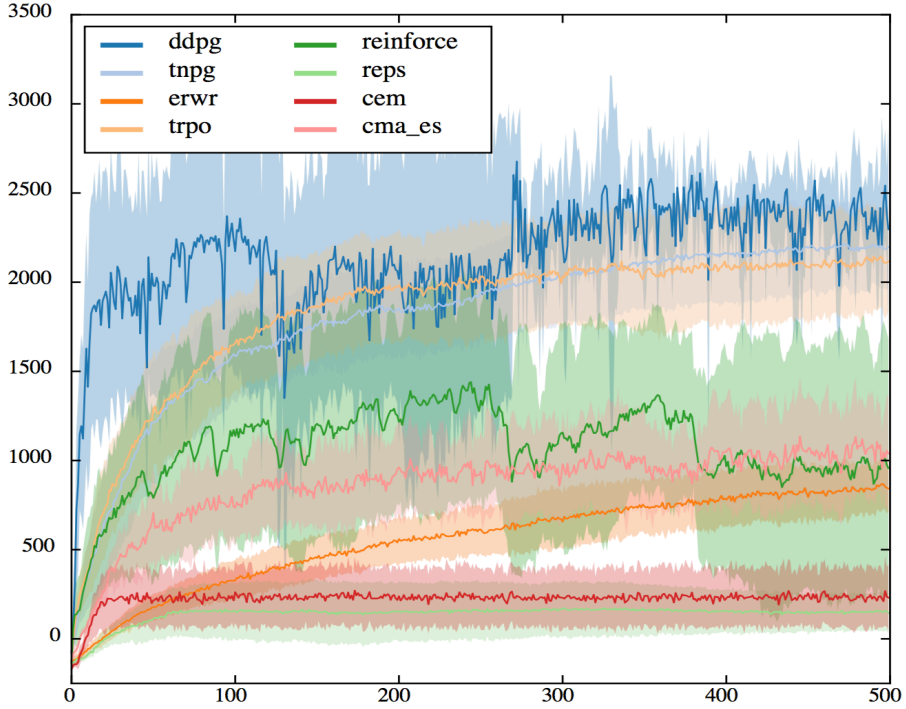
Results

	CEM	TRPO
Cart-Pole Balancing	4815.4±4.8	**4869.8±37.6**
Mountain Car	-66±2.4	**-61.7±0.9**
Swimmer	68.8±2.4	**96±0.2**
Half-Cheetah	**330.4±274.8**	**1914±120.1**

Results

	DDPG	TRPO
Cart-Pole Balancing	4634.4±87.8	**4869.8±37.6**
Inverted Pendulum	**40±244.6**	**247.2±76.1**
Acrobot	**-223.6±5.8**	**-326±24.4**
Hopper	**267.1±43.5**	**1183.3±150**
Half-Cheetah	**2148.6±702.7**	1914±120.1

Results


ddpg
tnpg
erwr
trpo
reinforce
reps
cem
cma_es
3500
3000
2500
2000
1500
1000
500
0
0
100
200
300
400
500

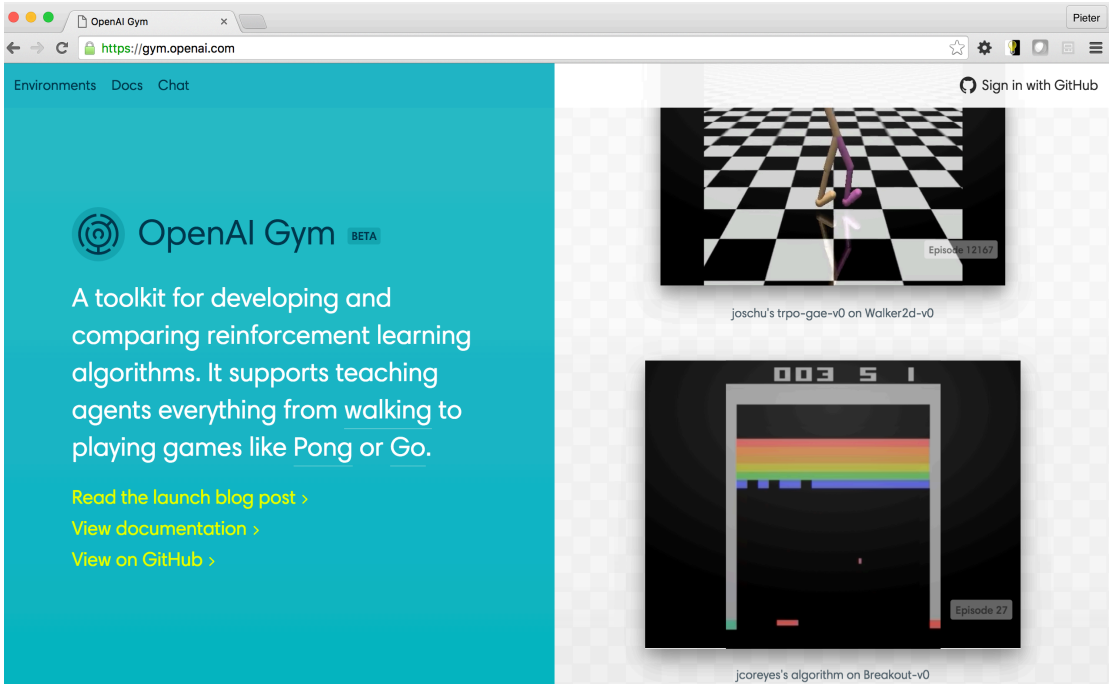
Results

	TNPG	TRPO	DDPG	Others
Best Performance	40%	52%	8%	0%

Results

	Random	REINFORCE	TNPG	RWR	REPS	TRPO	CEM	CMA-ES	DDPG
Swimmer + Gathering	0±0	0±0	0±0	0±0	0±0	0±0	0±0	0±0	0± 0.0
Ant + Gathering	-5.8±5	-0.1±0.1	-0.4±0.1	-5.5±0.5	-6.7±0.7	-0.4±0	-4.7±0.7	N/A	-0.3± 0.3
Swimmer + Maze	0±0	0±0	0±0	0±0	0±0	0±0	0±0	0±0	0± 0.0
Ant + Maze	0±0	0±0	0±0	0±0	0±0	0±0	0±0	N/A	0±0

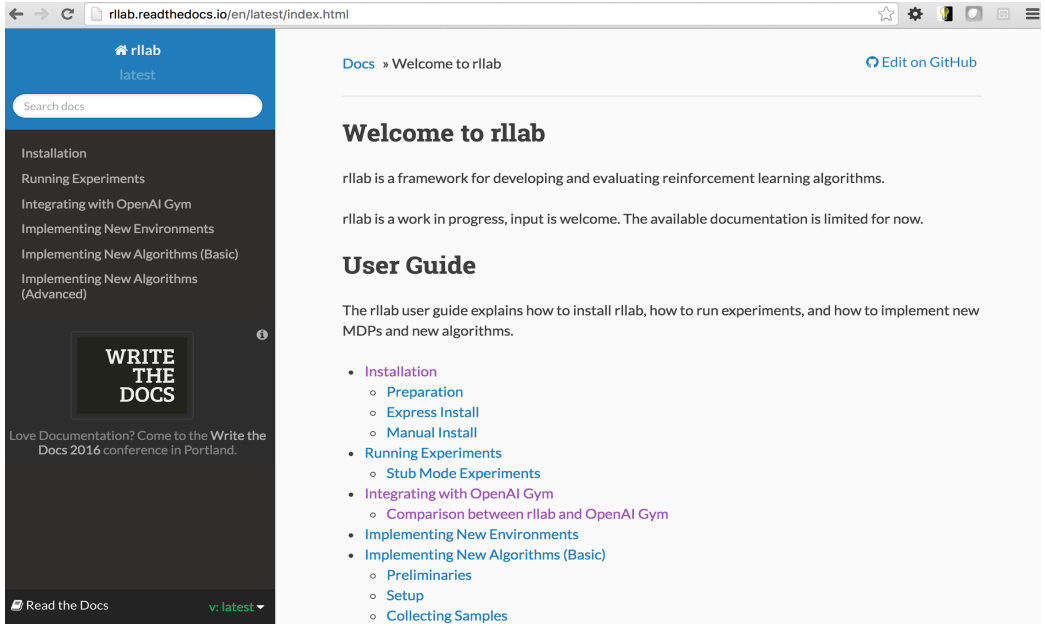
OpenAI Gym



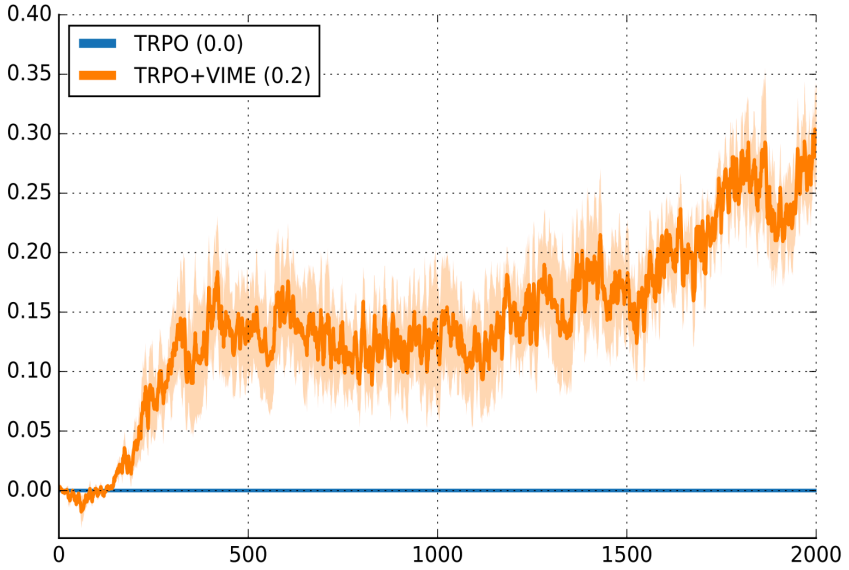
[Brockman et al. 2016]

rllab

https://github.com/rllab/rllab



VIME



[Houthooft et al. 2016]

Thank you

Poster session: Monday 3pm-7pm #11