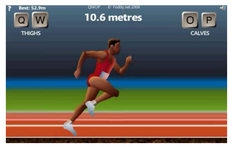
DL4CV Project

Haoran Chen[1], Lixin Xue[1], Kai Wu[1], Pengyuan Wang[1], and Yingqiang Gao[1]

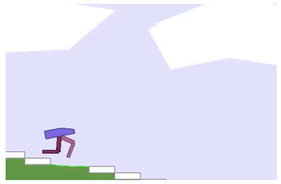
[1]Technical University of Munich

Introduction

- Achieve better performance on various environments such as QWOP and OpenAI Gyms' BipedalWalker
- Apply Computer Vision approaches to represent states of Reinforcement Learning



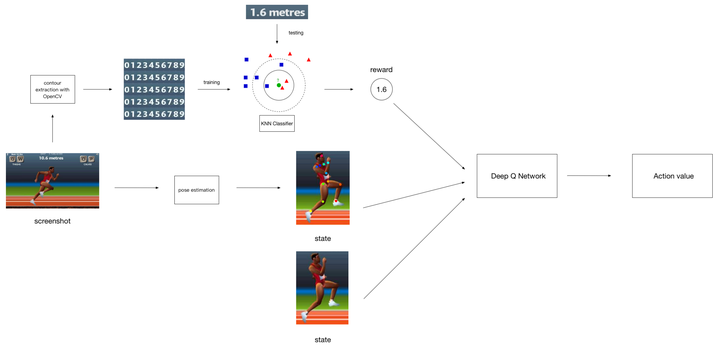
QWOP



BipedalWalker

Deep Q Network plays QWOP

- Train a KNN classifier to recognize numbers and use CNN to perform pose estimation from screenshot



- Use CNN + FC to Approximate the optimal action-value function

$$Q^*(s,a) = \max_a \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ...|s_t = s, a_t = a, \pi]$$

- Use Experience Play to Learn From the Past and Q-learning Update

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim U(D)}[(r + \gamma \max_{a'} Q(s', a'; \theta_i^-) - Q(s,a;\theta_i))^2]$$

- Use uinput to simulate the process of press button since QWOP is a closed-sourced

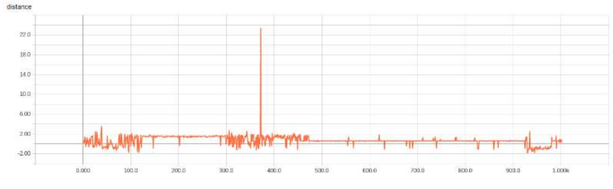


Fig: raw pixel as input of Deep Q Network

Deep Q Network plays QWOP

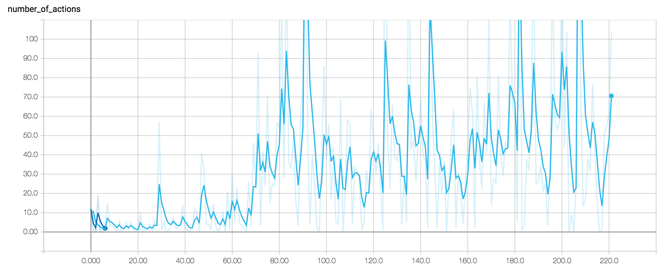


Fig: pose as input of Deep Q Network

DDPG and PPO plays BipedalWalker

DDPG

- Update critic by minimizing the loss

$$L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$$

- Update actor policy using the sampled policy gradient

$$\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_i \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|s_i$$

DDPG and PPO plays BipedalWalker

PPO

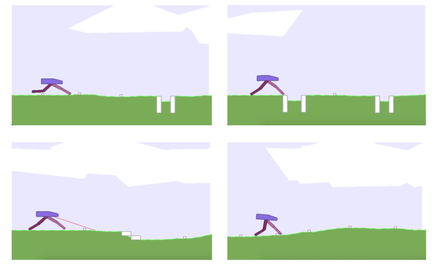


Fig: states after 50, 150, 250 and 450 episodes

4. Reference

References

[1] N. Heess, S. Sriram, J. Lemmon, J. Merel, G.Wayne, Y. Tassa, T. Erez, Z. Wang, A. Eslami, M. Riedmiller, et al. Emergence of locomotion behaviours in rich environments. arXiv preprint arXiv:1707.02286, 2017.

[2] T. Insider. Googles deepmind ai just taught itself to walk. July 2017.

[3] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971, 2015.

[4] J. Merel, Y. Tassa, S. Srinivasan, J. Lemmon, Z. Wang, G. Wayne, and N. Heess. Learning human behaviors from motion capture by adversarial imitation. arXiv preprint arXiv:1707.02201, 2017.

[5] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013.

[6] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529533, 2015.

[7] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for cnn.

[8] Z.Wang, J. S. Merel, S. E. Reed, N. de Freitas, G.Wayne, and N. Heess. Robust imitation of diverse behaviors. In Advances in Neural Information Processing Systems, pages 53245333, 2017.