Semester Project for DL4CV WS17/18

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

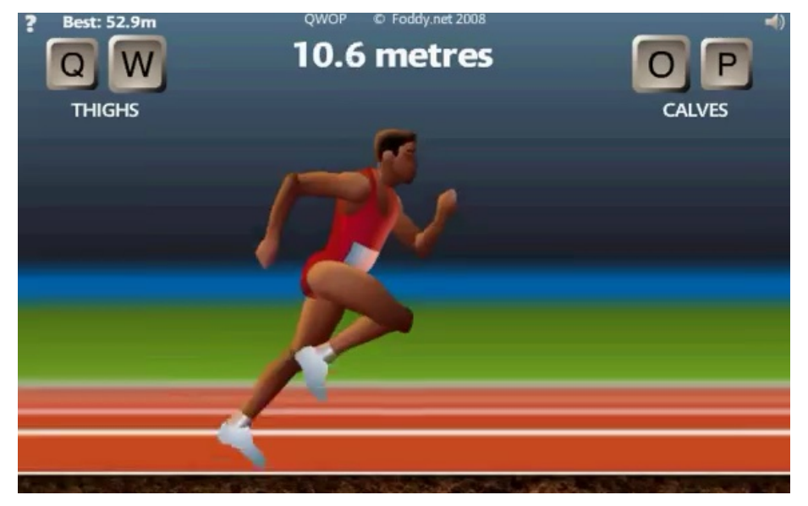
[1]Technical University of Munich

Introduction

- In this project, we studied several reinforcement learning algorithm, implemented them into two "experiment" environment, with respect to the implementation, raw pixel and modeled state was used as out input.
- The goal of this project is to use neural network to make a bipedal agent teaching himself to walk like human in test environment.

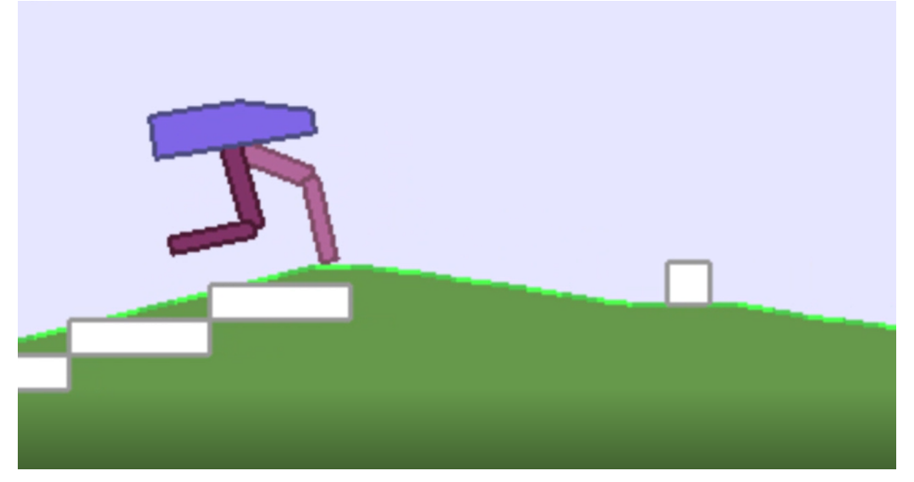
Motivation

- End to end training, network directly outputs action commands from high dimensional raw pixel inputs.
- However walking simulation as a hard control problem, is located in continuous action space. DQN can only handle with low dimensional action space.
- Heuristically, walking like human, multiple sources of "knowledge" could be a inspiration, an emergence of behaviours from rich environments.

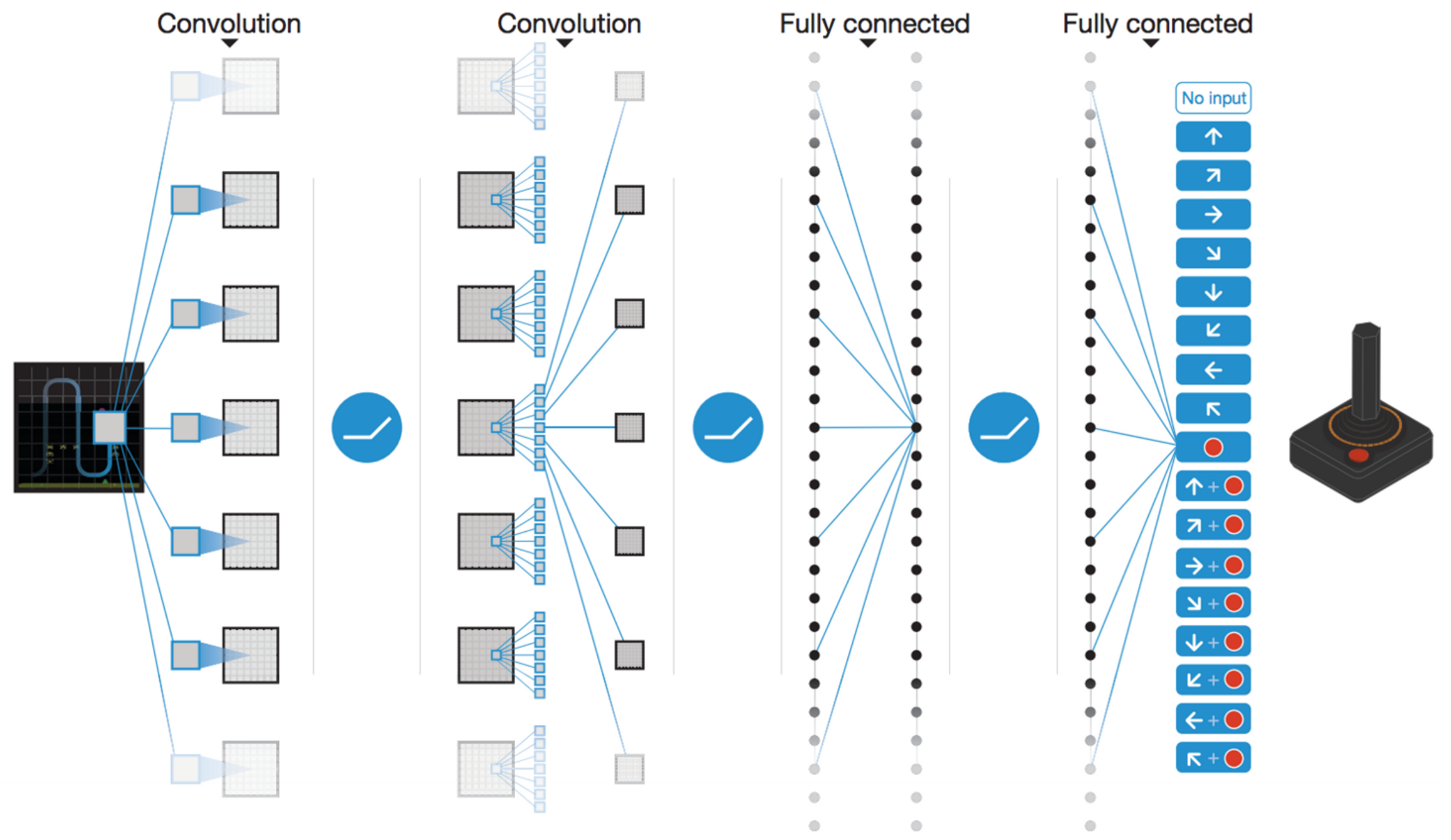


QWOP

DeepMind


BipedalWalker

Using DQN to play QWOP

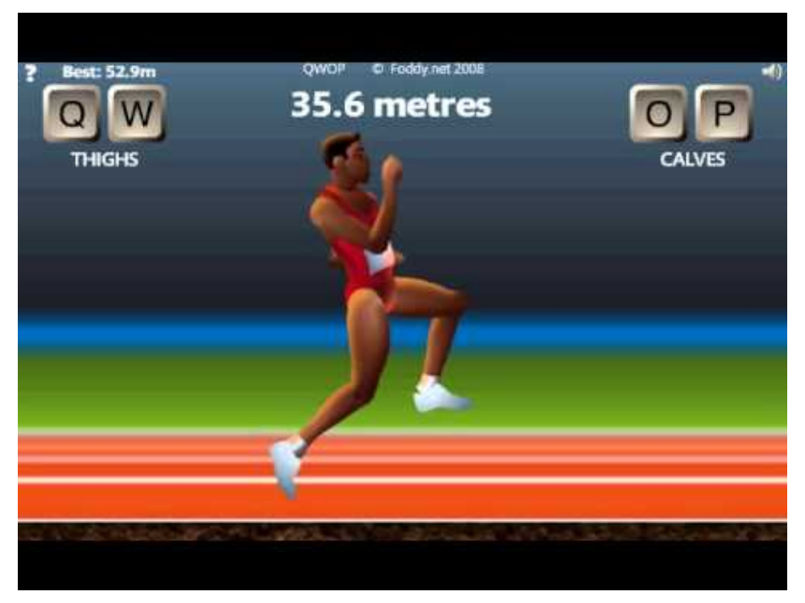


- The goal of Q learning is to maximize long-term cumulative reward with Q value(action state value)

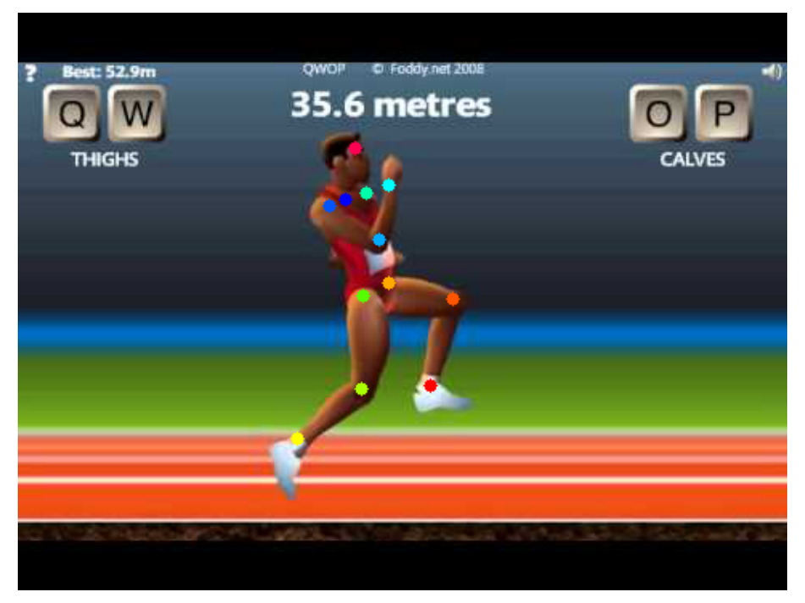
$$\boxed{\mathrm{R}_{t_0} = \sum_{t=t_0}^{\inf} \gamma^{t-t_0} r_t, Q^*(s,a) = \max_a \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ...|s_t = s, a_t = a, \pi]}$$

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

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

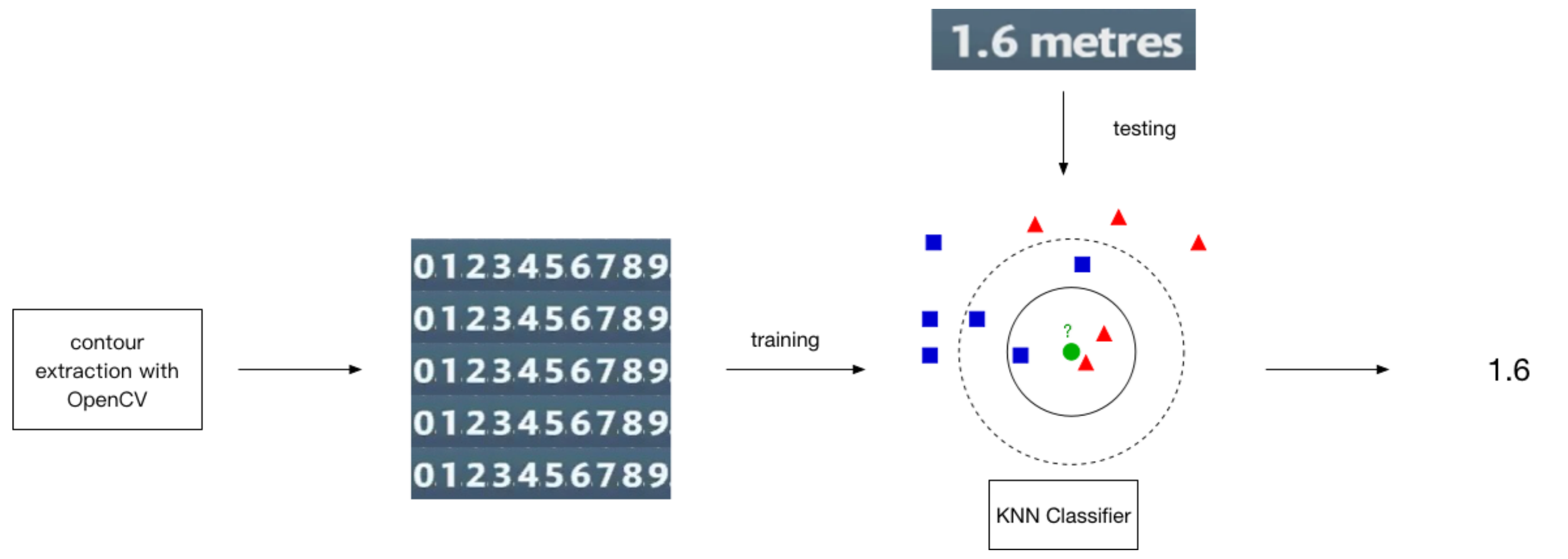


QWOP



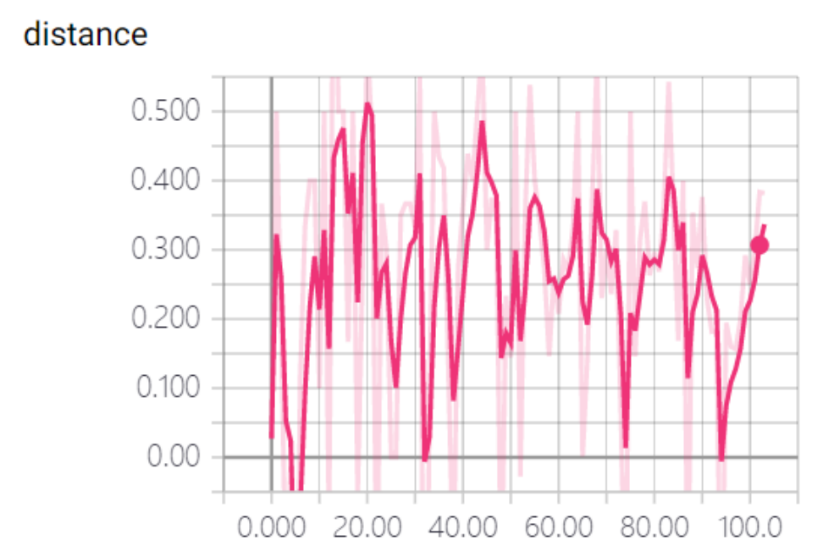
Pose Extraction

- Pretrained pose estimation model was used to predict moving joint from the screenshot of QWOP.

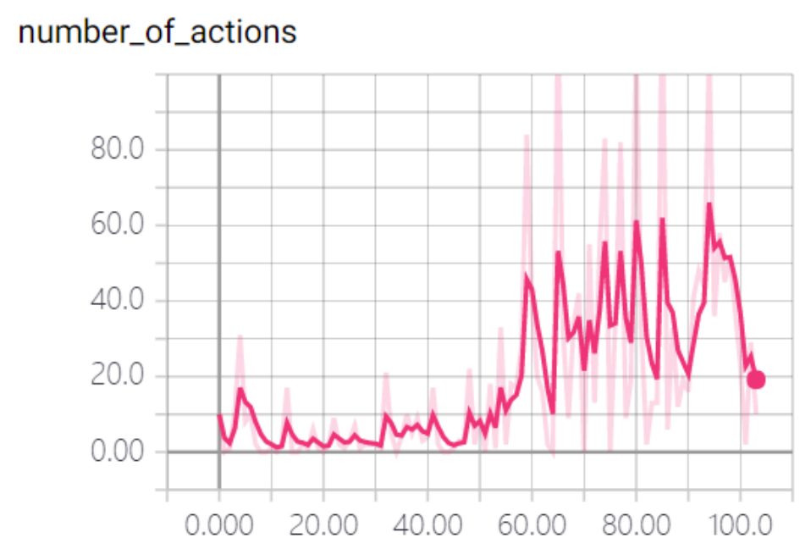


- There is a time delay when pose estimation outputs flows into Q network, which greatly increase the trainging time. - another problem is
- Another problem is the training result is very unsastisfactory, possible reason for this could be there are only 16 combinations of inputs(considering to play QWOP only four keys are needed), discrete action space may can not well represent the complex behaviour s likewalking. - The last problem is QWOP is total blackboxed, whcih could indicatef that envrioment is not fully observable.

Problems when using DQN

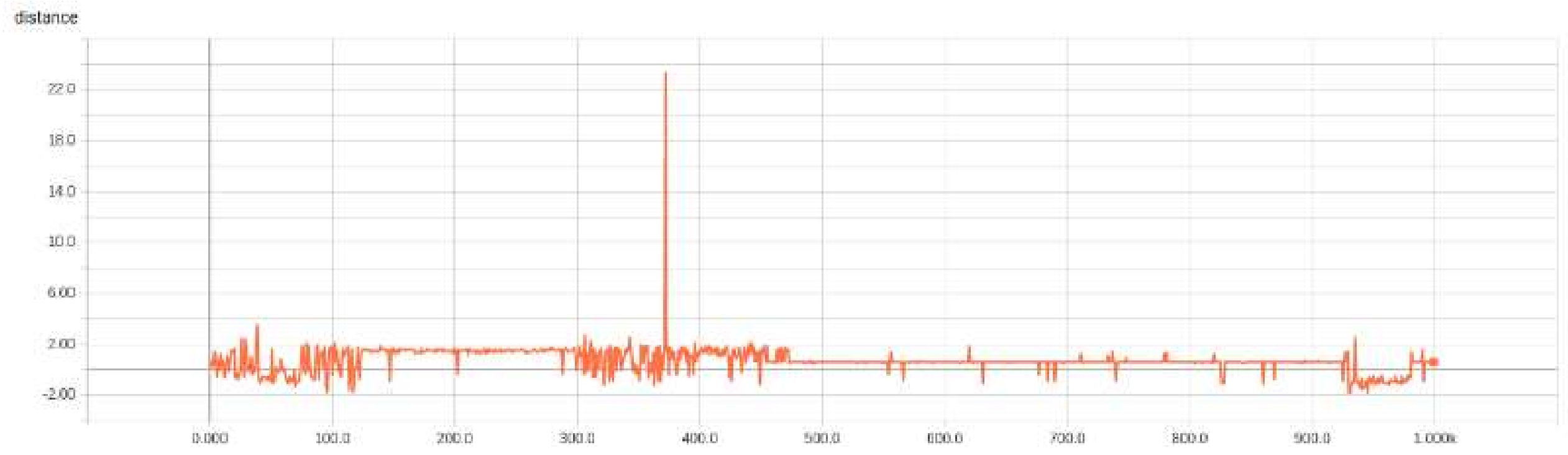


running distance wrt. episodes



actions wrt. episodes

- QWOP is also too blackboxed, only partial observations could be correctly transfered to states.



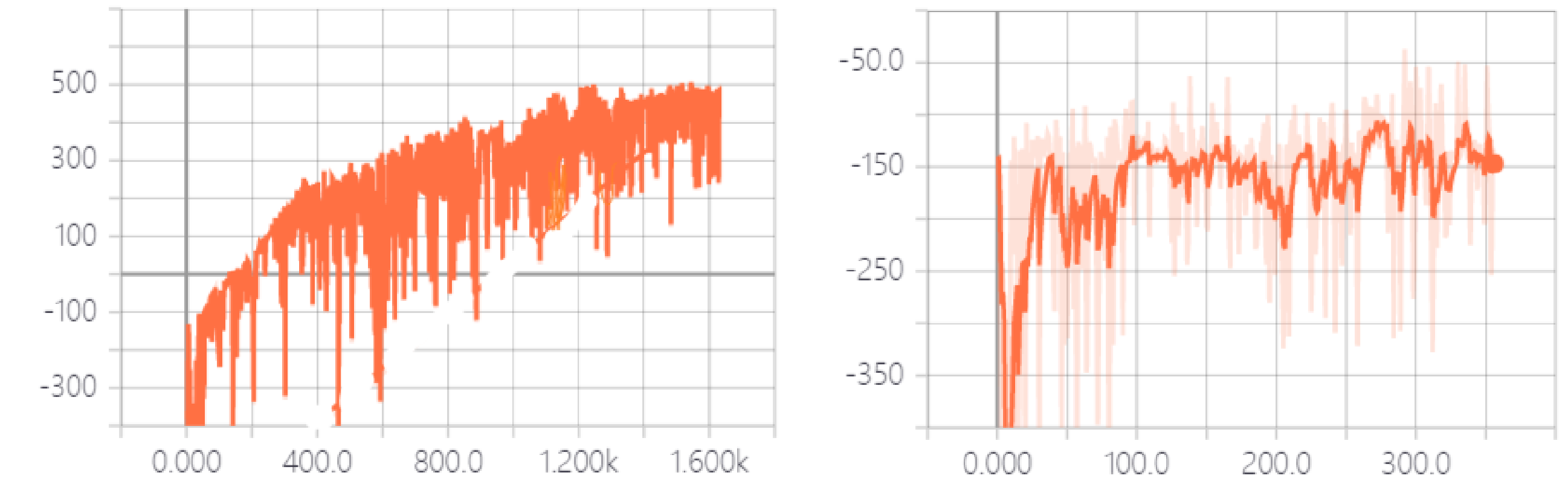
- Possible reason could be the discrete low-dimensional action space, since in QWOP there are only 16 combinations of inputs.

DDPG and PPO plays BipedalWalker

DDPG

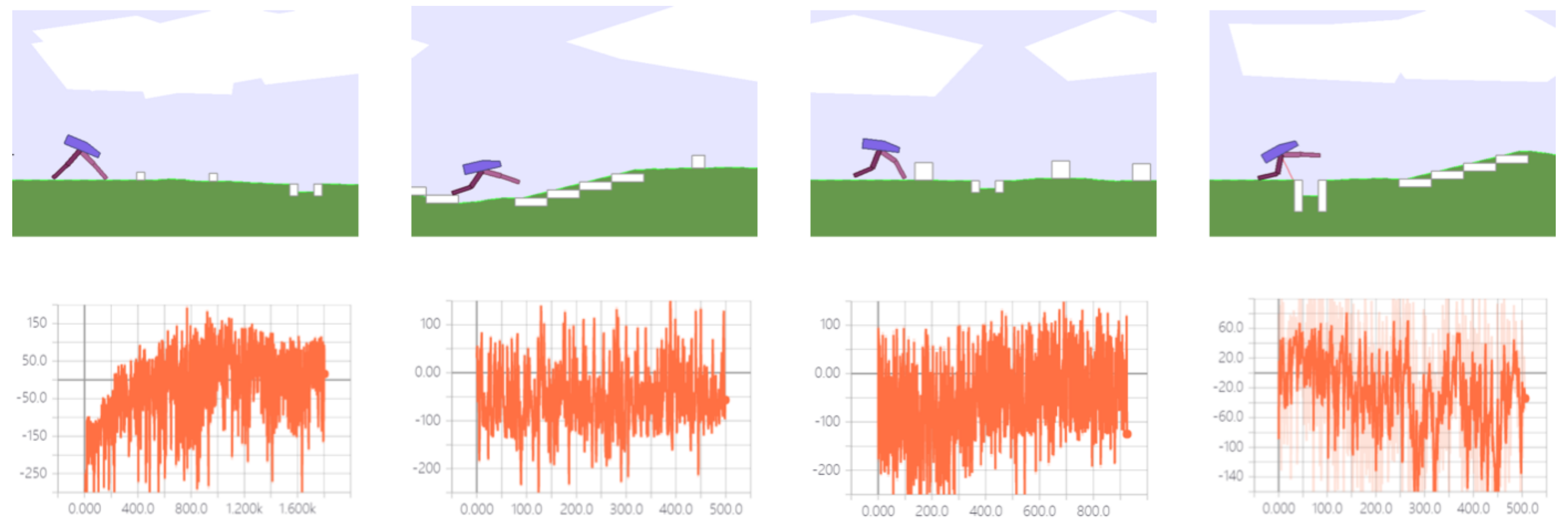
- DDPG introduce continuous action control into reinforcement learning, where actor part can be implemented with many different policy optimizations, PPO is one of the state of the art algorithm.

- Since QWOP is too bloackbox, we test our PPO implementation in Open AI gym's BipedalWalker environment.



PPO

- Agent could finish BipedalWalker-v environment within 1000 episodes. but even can't make any sound advances in hardcore environment, which includes 3 kinds of obstacles.



- We then tweaked the parameters of environment to procedurally generated tasks with increasing difficulties.

- With increasing difficulties, agent's rewards remains in certain range, which could indicate that agent is able to overcome challenges with increasing abilities, just like human.

Conclusion

- Modifying DQN with continuous output achieves better control result than discrete output
- The time cost of pose extraction has great influence on the performance
- Diversity of environment could help promoting the learning of agent