Ablation Study on TD3

(Addressing Function Approximation Error in Actor-Critic Methods)

RL Team Implementation Project
Team 11 施奕成、高嘉豪、楊宗穎、謝宏笙

Content

- **Introduction (TD3)**
- **Ablation Study**
  - TD3 modules tuning
  - Network architecture
  - Exploration
  - Sampling
  - Training Procedure
- **Conclusion**

Introduction

- DDPG has some issues (e.g. overestimation)
- **TD3 (Twin Delayed Deep Deterministic policy gradient)** is proposed
  - Clipped double Q-learning

$$y_1 = r + \gamma \min_{i=1,2} Q_{\theta'_i}(s', \pi_{\phi_1}(s')).$$

  - Target policy smoothing

$$y = r + \gamma Q_{\theta'}(s', \pi_{\phi'}(s') + \epsilon), \quad \epsilon \sim \operatorname{clip}(\mathcal{N}(0, \sigma), -c, c),$$

  - Delayed policy update

Introduction

- Experiments are done with the **official released code** from the authors
    - (Authors said it **doesn’t match** the code used in the paper anymore)
- Our ablations are done in the following 4 MuJoCo environments
    - Ant-v3
    - HalofCheetah-v3
    - Hopper-v3
    - Walker2d-v3
- The final results are average of 5 runs with different random seed

TD3 Modules Tuning

- Delayed policy update
- Target action noise
- Double Q

TD3 Modules Ablation

AHE = (TD3 architecture, hyper-parameters and exploration, no DP/TPS/CDQ)

- With the code, environment, and hyper-parameters changed, the modules ablation results from paper don't exactly hold true
- Most agents trained **without CDQ** suffer more significantly

Paper

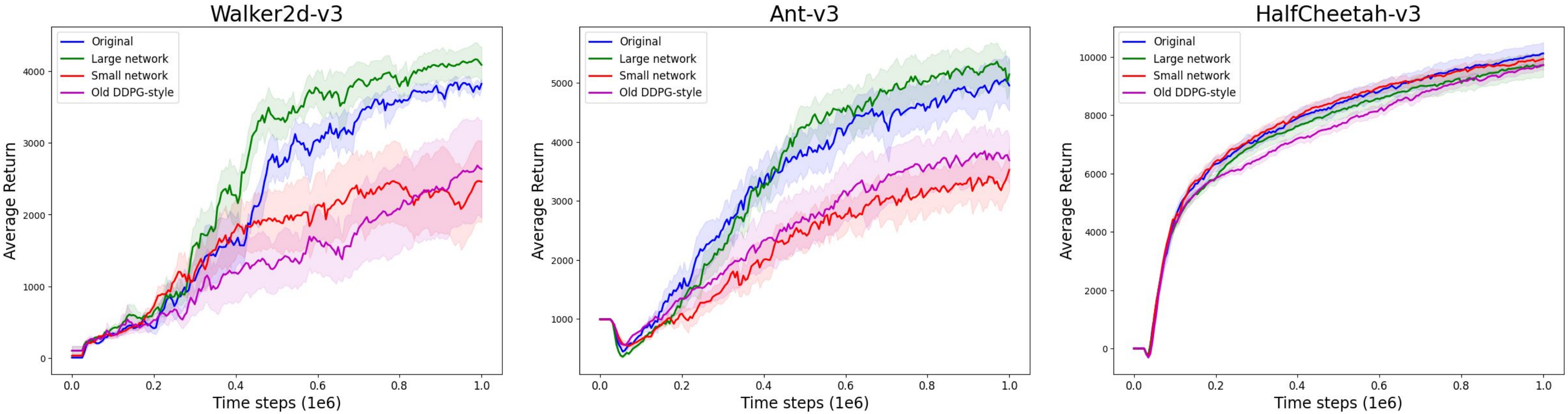
Method	HCheetah	Hopper	Walker2d	Ant
TD3	9532.99	**3304.75**	**4565.24**	**4185.06**
AHE	8401.02	1061.77	2362.13	564.07
AHE + DP	7588.64	1465.11	2459.53	896.13
AHE + TPS	9023.40	907.56	2961.36	872.17
AHE + CDQ	6470.20	1134.14	3979.21	3818.71
TD3 - DP	9590.65	2407.42	**4695.50**	3754.26
TD3 - TPS	8987.69	2392.59	4033.67	**4155.24**
TD3 - CDQ	9792.80	1837.32	2579.39	849.75

Ours

Method	HCheetah	Hopper	Walker2d	Ant
TD3	10118.32	**3272.81**	**4956.59**	3821.05
AHE	11129.36	1785.28	511.73	1291.70
AHE+DP	9857.86	1723.75	753.06	1694.79
AHE+TPS	10805.13	1444.09	1082.89	1588.60
AHE+CDQ	9681.62	3049.95	4179.50	**4185.18**
TD3-DP	9425.92	**3322.63**	**5290.79**	3977.29
TD3-TPS	10287.62	3039.83	3806.50	**4138.42**
TD3-CDQ	**11571.69**	2347.96	1547.19	2147.88

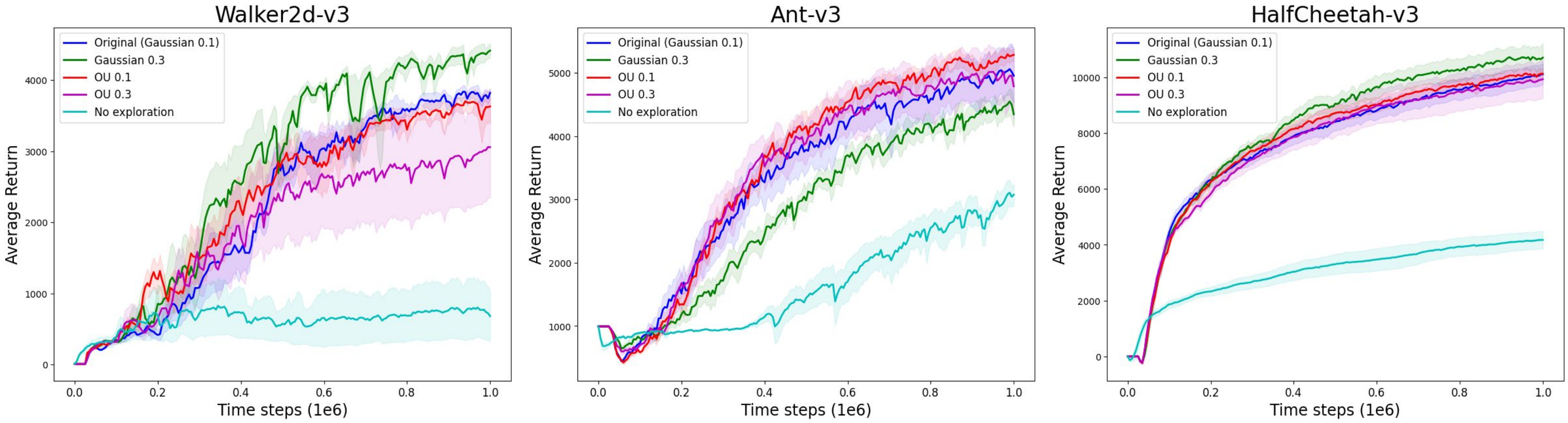
Network Architecture

- Channel size of two FC layers:
  - Original [256, 256]
  - Large [400, 300]
  - Small [128, 128]
- Old DDPG-style:
  - The action in critic network only go through 1 layer (instead of 2)



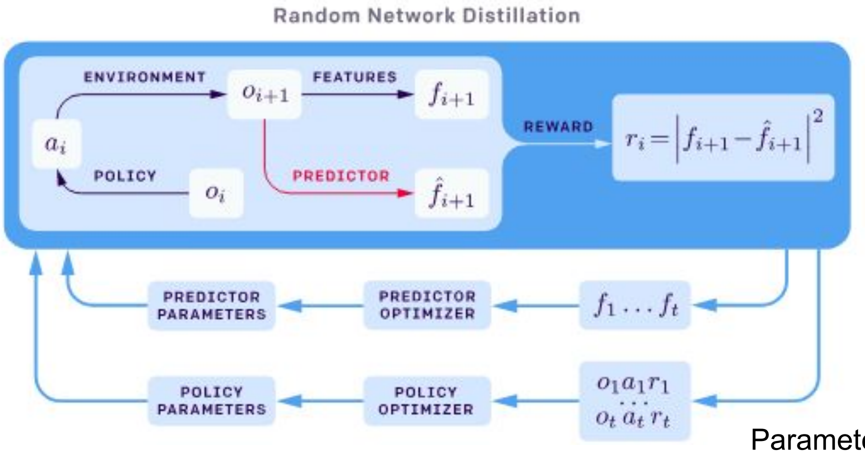
Exploration - Policy Noise

- Gaussian noise vs. Ornstein–Uhlenbeck noise
- Noise scale (variance for Gaussian): 0.1 vs. 0.3



Exploration - Exploration Module

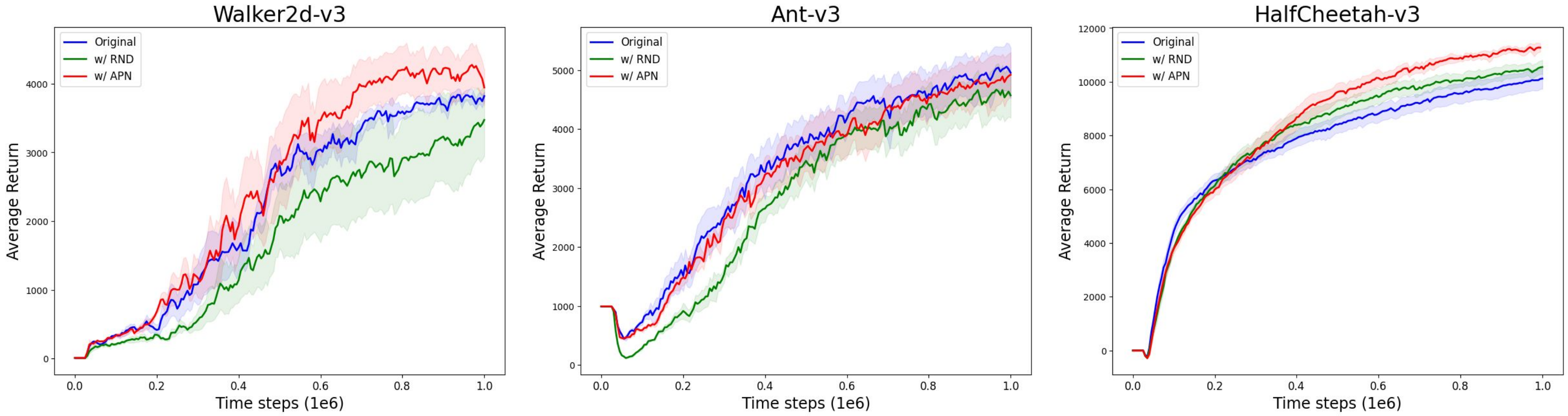
- **APN** (Adaptive Parameter Noise)
  - Add adaptive noise to the **parameters** of the neural network policy (rather than to its action space)
- **RND** (Random Network Distillation)



Parameter Space Noise for Exploration, Plappert et al., ICLR 2018
Exploration by Random Network Distillation, Burda et al., ICLR 2019

Exploration - Exploration Module

- Agents trained with **APN** show better overall results compared to original and RND



Sampling

- **PER** (Prioritized Experience Replay)
  - Batch sampling by prioritization in memory.
  - If the TD-error is bigger, that means there still be room for prediction accuracy to rise, then the prioritization is higher.

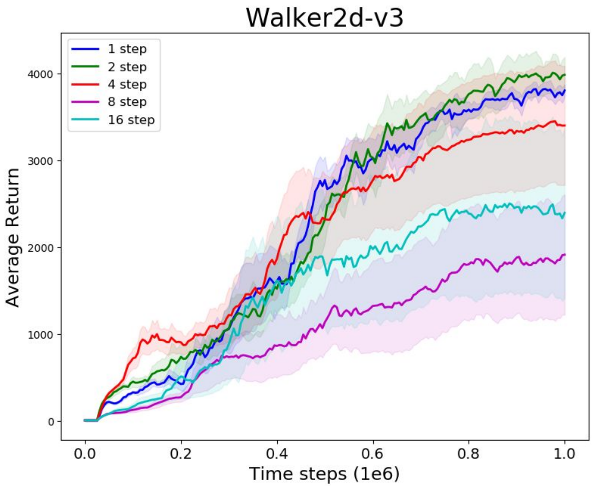
Prioritized Experience Replay, Schaul et al., ICLR 2016

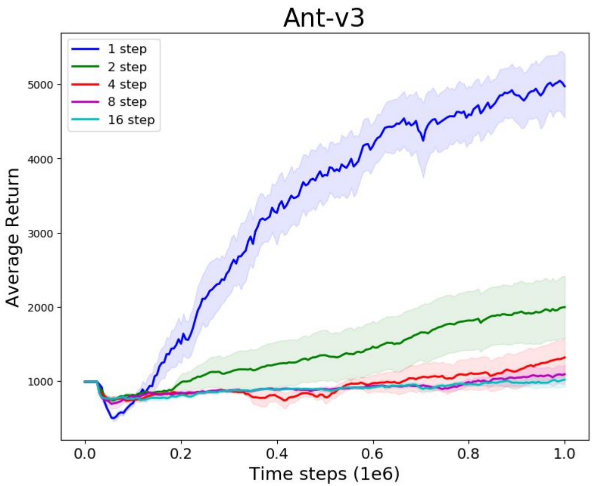
Sampling

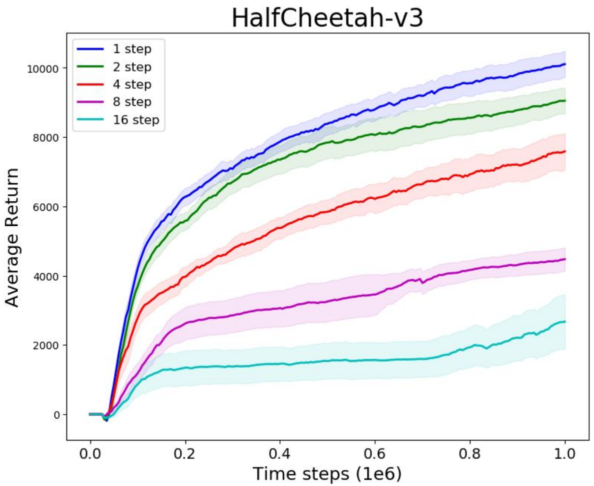
- results w/ PER

Training Procedure

- N-step return
  - Bias-variance tradeoff







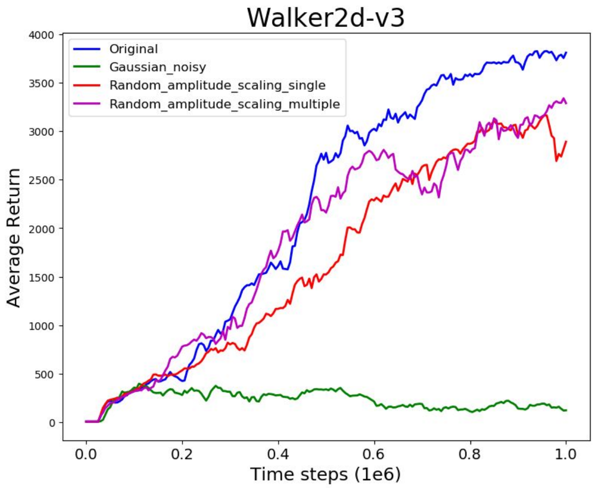
Training Procedure

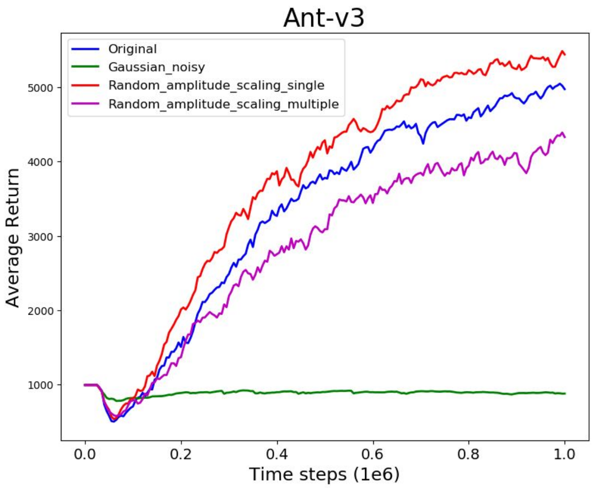
- Data augmentation
  - Gaussian noisy: add Gaussian noisy
  - Random ampilitude scaling (single): multiplies the uniform noisy (scalar)
  - Random ampilitude scaling (multiple): multiplies the uniform noisy (vector)

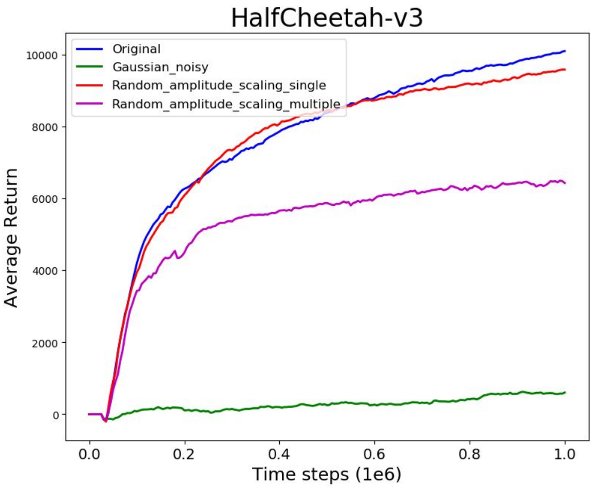
Reinforcement Learning with Augmented Data, Laskin et al., NIPS 2020

Training Procedure

- Data argumentation may not really helpful







Conclusion

- We tested TD3 algorithm with various hyper-parameters setting and removed/added several modules to see how it perform.
- For certain environments, some modification we made lead to better results while other might not.
- Due to limited time and the number of settings, we didn’t do the mix-and-match of different modifications, which may be something to explore later.

Exploration - Warm-up Steps

- At timestep < warm-up steps, a random action is chosen and no update
- Warm-up steps: [25000 (original), 10000, 50000]