Presentation Key Point

2023/11/14

Paper Presentation Schedule and Rules

- Presentation Time:
  - Week 1: 12/5
  - Week 2: 12/12
  - Week 3: 12/19
- Presentation Rules:
  1. Each team will have a total of 20 minutes for their presentation.
  2. The presentation will consist of a 12-minute talk with no more than 30 slides in PDF or PPT format, followed by an 8-minute Q&A session.
  3. During the presentation, each member of the group must present at least a part.

Notifications

- The main points of the paper below must be addressed during the presentation, but that doesn’t mean you only need to say the main points below.

Prioritized Experience Replay, PER

- How to define the priority of samples?
- How to correct the bias of PER by Importance sampling?

Bootstrapped DQN

- How can multi-head architecture help exploring?

Noisy DQN

- Where to add noise?
- How to add noise?
- Compare with original exploration methods.
  - Epsilon greedy.
  - Entropy bonus.

Multi-labelled Value Networks

- BV-ML value network.

AlphaZero with PBT

- How to do population based training in:
  - Self play
  - Optimization

Path Consistency AlphaZero

- What is path consistency?
- What is feature consistency?

EfficientZero

- Introduce the improvements to MuZero
  - Self-Supervised Consistency Loss
  - Prediction of the Value Prefix
  - Off-Policy Correction

Cumulative Regret: UCB

- Choose one of the papers
  - Explain UCT
  - Explain UCB

Sutton Proof

- The derivation of the surrogate function in TRPO.
  a. $\eta \rightarrow L$
- Show that surrogate function in TRPO satisfies the two properties:
  a. $$L_{\pi_{\theta_{old}}}(\pi_{\theta_{old}}) = \eta(\pi_{\theta_{old}}),$$
  b. $$\nabla_\theta L_{\pi_{\theta_{old}}}(\pi_\theta)\Big|_{\theta=\theta_{old}} = \nabla_\theta \eta(\pi_\theta)\Big|_{\theta=\theta_{old}}$$

KL/TV distance

- Derive the Policy Improvement Bound (PIB) in both TV and KL terms.

Let $\alpha = D_{TV}^{max}(\pi, \tilde{\pi})$, then the following bound holds:

$$\eta(\tilde{\pi}) \geq L_{\pi_{old}}(\tilde{\pi}) - \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2$$

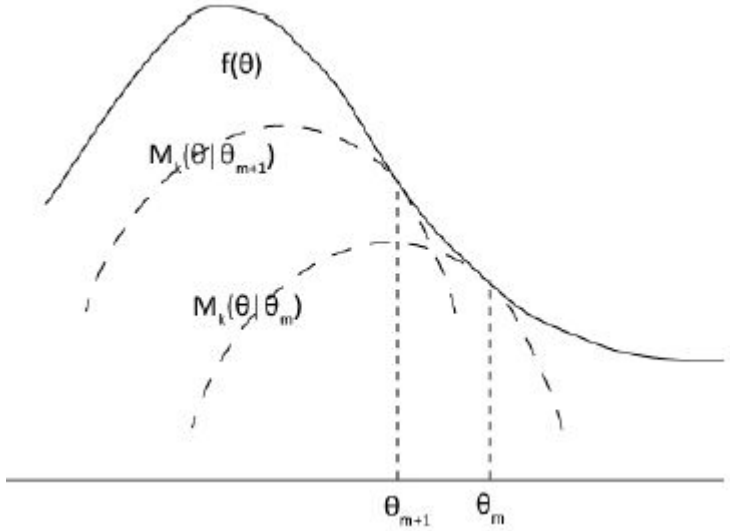
$$\text{where } \epsilon = \max_{s,a}|A_{\pi}(s,a)|$$

$$\eta(\tilde{\pi}) \geq L_{\pi}(\tilde{\pi}) - C \cdot D_{KL}^{max}(\pi, \tilde{\pi})$$

$$\text{where } C = \frac{4\epsilon\gamma}{(1-\gamma)^2}$$

Policy Improvement based on MM

- Explain the Minorization-Maximization (MM) algorithm.
- MM with TRPO.



Stochastic MuZero

- What is afterstate dynamics and afterstate prediction?
- What is chance outcome?
- How to do stochastic search?

Gumbel MuZero

- What is Gumbel-Max trick and Gumbel-Top-k trick?
- How to plan with Gumbel at the root node?
- How to learn an improved policy?

Never Give Up (NGU)

- Life-long novelty module.
- Episodic novelty module.
- Calculating intrinsic reward.
- A family of policies.

Agent57

- Add some improvements to NGU.

Decision Transformer

- How does the Decision Transformer cast the problem of RL as conditional sequence modeling?
- What is its trajectory representation?

IQN

- Compare the difference with QR-DQN.
- What is distortion risk measure?

FQF

- Introduce its networks.
  - Fraction proposal network
  - Quantile value network
- What is the side effect comparing with IQN?

GAIL

- Generator
- Discriminator
- How does they work together?

RLPD

- What’s the strategy it use to incorporate offline data?
- Layer normalization mitigates catastrophic overestimation.
- How does it deal with update-to-data ratio? What is random ensemble distillation?

Q-Mix

- The mixing network
  - Neural network structure
  - Non-linear mixing
- Why is the CTDE method?
- Compare with VDN

MAT

- How does the Multi-Agent Transformer cast the problem of Coop MARL as conditional sequence modeling?
- How does it compute value? (Critic)
- What does it do when training and testing, respectively?

AlphaStar

- Prioritized Fictitous Self-Play (PFSP)

OpenAI Five

- Horizon effect
- OpenAI Five Model Architecture
- System Overview: controller, rollout worker, optimizer
- Reward
- Team spirit

Groups

Paper Lists

- Week 1 (8 papers)
  - Prioritized Experience Replay, PER
    - 312554030 王之炫　311551157 簡昕益　312553016 周原慶
  - Bootstrapped DQN
    - 109705001 陳以瑄　110705013 沈昱宏　109705003 吳振豪
  - Noisy DQN
    - 312581020 許瀚丰　312551108 林書緯　312551095 張鈞奕
  - Multi-labelled Value Networks
  - AlphaZero with PBT
  - Path Consistency AlphaZero
  - EfficientZero
  - Cumulative Regret: UCB

Paper Lists

- Week 2 (8 papers)
  - Sutton Proof (gradient is the same)
    - 312551091 陳子安　312554032 陳騰睿　312554035 姜柔嘉
  - KL/TV distance (total variation divergence and the KL divergence) in TRPO
  - Policy Improvement based on MM (minorization-maximization) for TRPO
  - Stochastic MuZero
    - 312551033 邱恆毅　312553051 陳建樺　312551113 蔡昀叡
  - Gumbel MuZero
    - 葉佳翰　　陳允關　　吳柏憲
  - Never Give Up (NGU)
    - 311706007 楊雅喬　311706009 王廣和　311706013 林韋臻
  - Agent57
    - 312554012 王偉誠　11110101俞丞訓　111101018 孫揚喆
  - Decision Transformer
    - 311581019 何立平　311553043 陳弘軽　311551142 江孟修

Paper Lists

- Week 3 (8 papers)
  - IQN
    - 311352004 童政瑜　311356003 陳澤昕　311554060 張偉誠
  - FQF
  - GAIL
    - 109612019 林伯偉　312551056 許瀚宇　311511056 游翔竣
  - RLPD
    - 312552022 田詠恩　312554041 謝博舟　412551017 王廣達
  - Q-Mix
  - MAT
    - 311552052 張壬豪　312552026 蔡濟謙　312551124 馮信華
  - AlphaStar
    - 312551045 施泰俊　312551161 張宸愷　312551047 黃玟綾
  - OpenAI Five
    - 312551131 李建緯　B121016 許仁覺　a121502 孫利東

Not been chosen yet (9 papers)

- Week 1 (5 papers)
  - Multi-labelled Value Networks
  - AlphaZero with PBT
  - Path Consistency AlphaZero
  - EfficientZero
  - Cumulative Regret: UCB
- Week 2 (2 papers)
  - KL/TV distance (total variation divergence and the KL divergence) in TRPO
  - Policy Improvement based on MM (minorization-maximization) for TRPO
- Week 3 (2 papers)
  - FQF
  - Q-Mix