深度强化学习的对抗攻防算法研究

毕业设计开题报告

翁家翌

清华大学计算机科学与技术系

2020 年 1 月 2 日

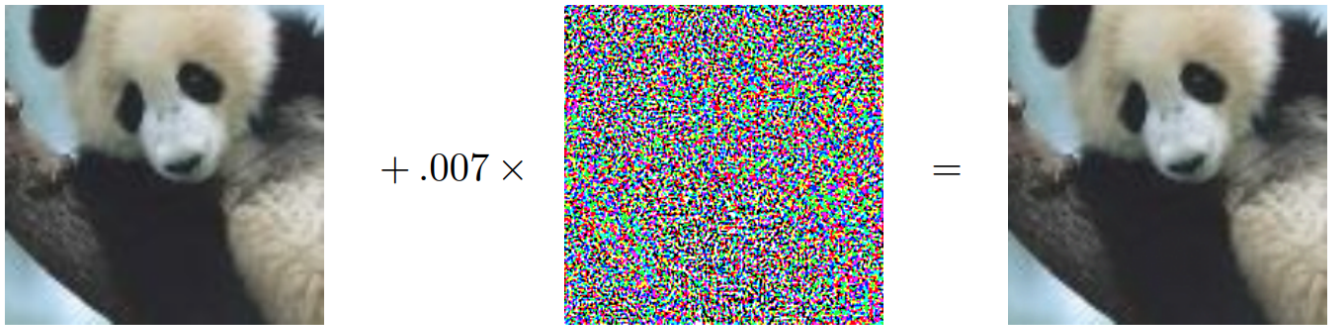

图: 对抗样本在图像分类上的应用 [GSS15]

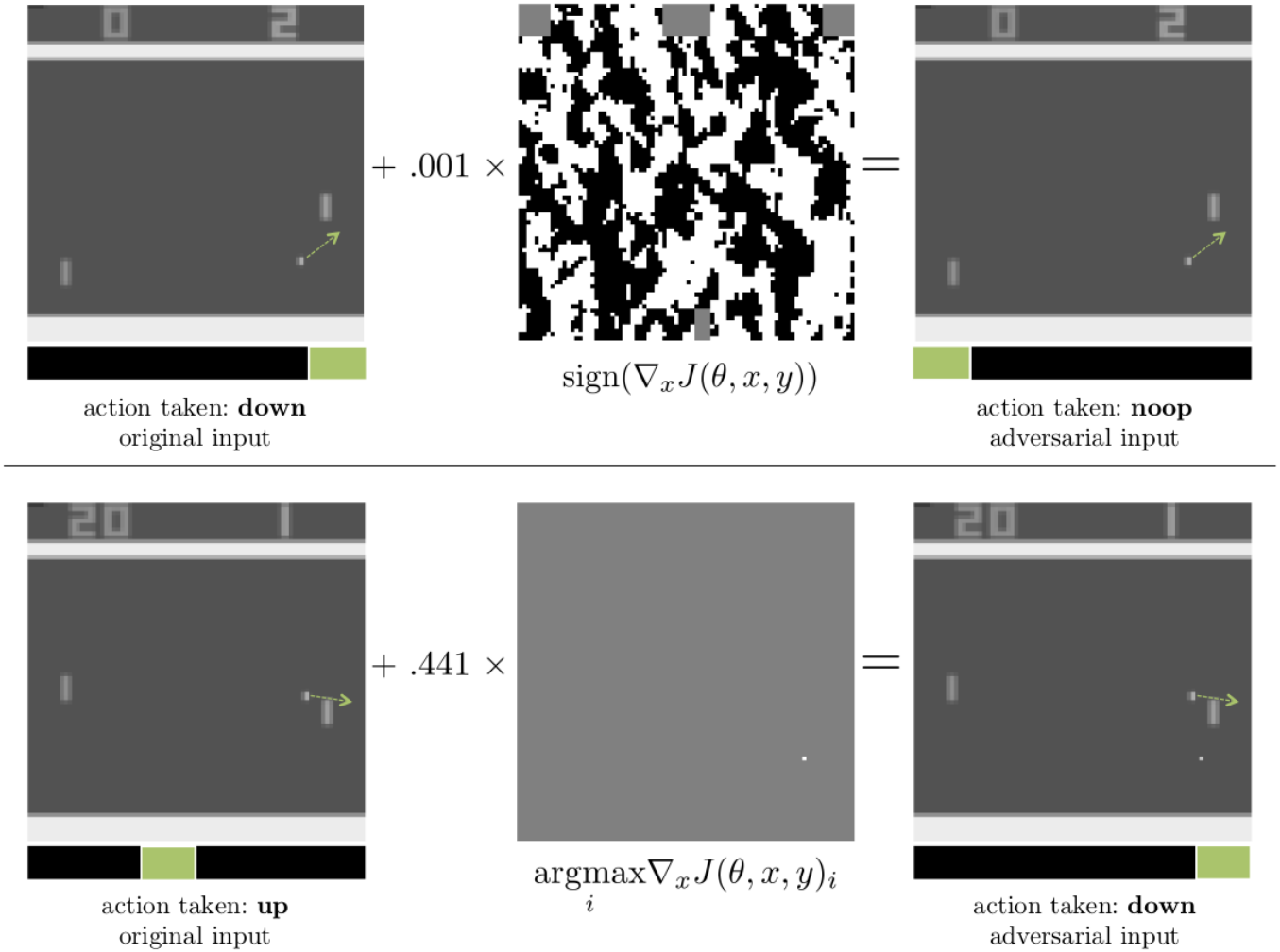

图: 对抗攻防在强化学习场景 Atari Pong 中的应用 [HPG+17]

- 逐帧攻击 / 挑帧攻击
- 有目标攻击 / 无目标攻击
- 攻击观测状态 / 攻击抽象状态 / 攻击 MDP 转移函数
- 基于梯度的方法 / 基于值函数的方法 / 启发式攻击方法 / 迁移攻击方法 / 基于强化学习的攻击方法

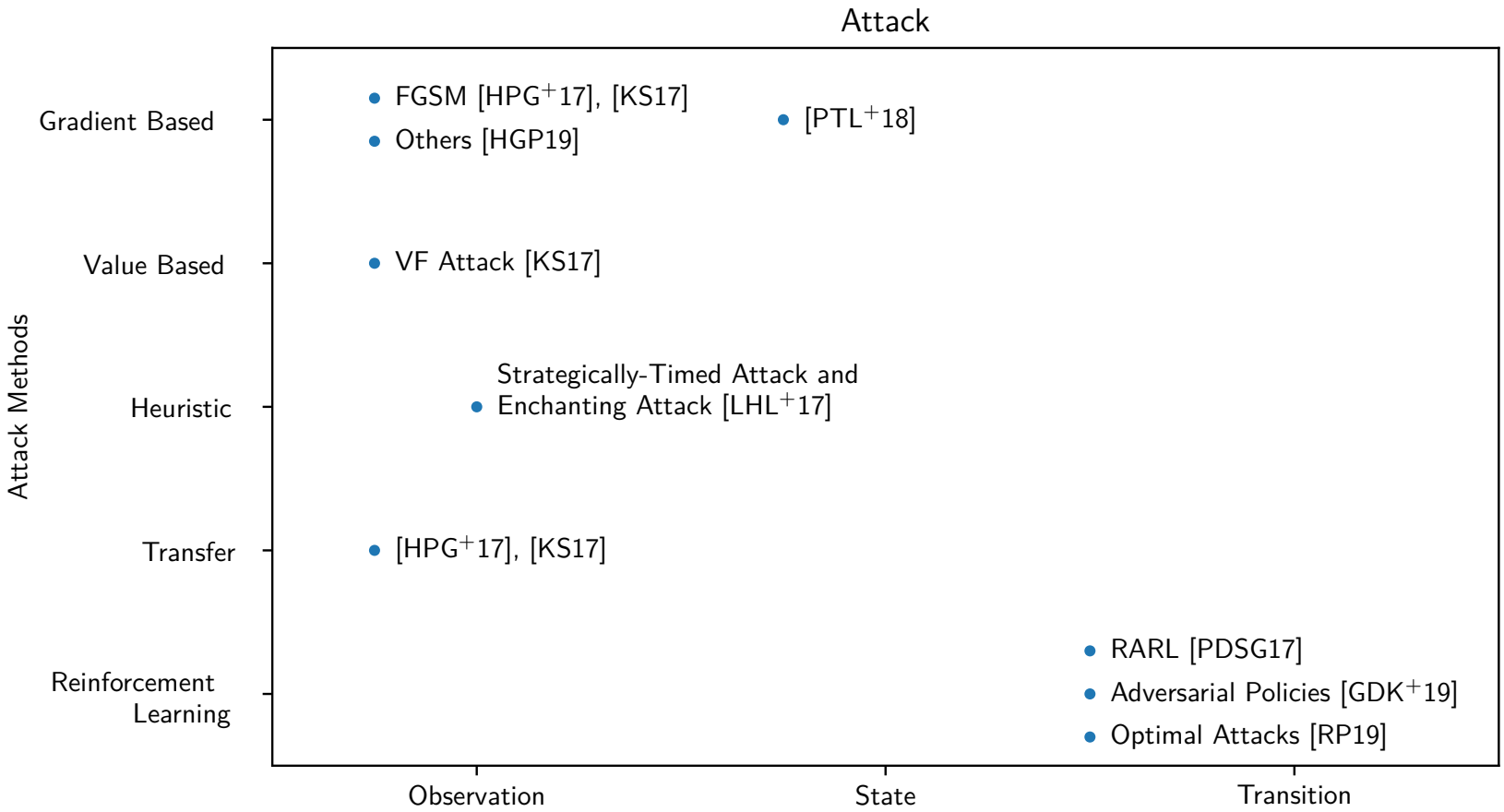

图: 强化学习攻击算法分类

- Optimal adversarial return
  攻击之后的智能体所能拿到的最多奖励
- Adversarial regret
  攻击前与攻击后，智能体所能拿到的奖励的差值
- Per-episode mean cost of attacker
  攻击者实施攻击需要花费的代价

- 之前的一些攻击算法大部分基于 Model-free 的方法，诸如 DQN [MKS+15] / DDPG [LHP+16] / A3C [MBM+16] / TRPO [SLA+15] / PPO [SWD+17]
- 基于 [BH19] 中的评测指标，使用开源框架 Ray [MNW+18] 和 RLlib [LLN+18] 实现对这些方法的鲁棒性评测，并同时评测最新的 Model-free 算法 TD3 [FvHM18] 和 SAC [HZAL18]
- 如果条件允许，可以评测 Model-based 的一些经典算法

- 基于 [PDSG17] 和 [GDK+19] 等工作，将其拓展到更难场景（比如 Atari）中，需要解决攻击者决策空间过大以及收敛性等问题
- 结合 Game Theory 的一些现有结论给出定性分析以及理论证明
- 期望能够开发出在对抗训练框架下能够比之前的工作有显著提升的算法

- 一月：完成文献调研
- 二月：复现具体攻防算法及性能测评
- 三、四月：基于现有方法的一些改进
- 五月：论文撰写


Vahid Behzadan and William Hsu.
RI-based method for benchmarking the adversarial resilience and robustness of deep reinforcement learning policies.
In *Computer Safety, Reliability, and Security - SAFECOMP 2019 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Turku, Finland, September 10, 2019, Proceedings*, pages 314–325, 2019.


Scott Fujimoto, Herke van Hoof, and David Meger.
Addressing function approximation error in actor-critic methods.
In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1582–1591, 2018.


Adam Gleave, Michael Dennis, Neel Kant, Cody Wild, Sergey Levine, and Stuart Russell.
Adversarial policies: Attacking deep reinforcement learning.
CoRR, abs/1905.10615, 2019.


Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy.
Explaining and harnessing adversarial examples.
In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.


Léonard Hussenot, Matthieu Geist, and Olivier Pietquin.
Targeted attacks on deep reinforcement learning agents through adversarial observations.
CoRR, abs/1905.12282, 2019.


Sandy H. Huang, Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and Pieter Abbeel.
Adversarial attacks on neural network policies.
In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.


Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine.
Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor.
In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1856–1865, 2018.


Jernej Kos and Dawn Song.
Delving into adversarial attacks on deep policies.
In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.


Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun.
Tactics of adversarial attack on deep reinforcement learning agents.
In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3756–3762, 2017.


Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra.
Continuous control with deep reinforcement learning.
In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.


Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldberg, Joseph Gonzalez, Michael I. Jordan, and Ion Stoica.
Rllib: Abstractions for distributed reinforcement learning.
In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 3059–3068, 2018.


Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu.
Asynchronous methods for deep reinforcement learning.
In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1928–1937, 2016.


Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis.
Human-level control through deep reinforcement learning.
Nature, 518(7540):529–533, 2015.

Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I. Jordan, and Ion Stoica.
Ray: A distributed framework for emerging AI applications.
In *13th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2018, Carlsbad, CA, USA, October 8-10, 2018*, pages 561–577, 2018.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta.
Robust adversarial reinforcement learning.
In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 2817–2826, 2017.


Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary.
Robust deep reinforcement learning with adversarial attacks.
In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, pages 2040–2042, 2018.


Alessio Russo and Alexandre Proutière.
Optimal attacks on reinforcement learning policies.
CoRR, abs/1907.13548, 2019.


John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz.
Trust region policy optimization.
In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1889–1897, 2015.


John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov.
Proximal policy optimization algorithms.
CoRR, abs/1707.06347, 2017.