Contrastive Offline Reinforcement Learning

https://www.yuewu.ml/projects/contrast_RL/

Yue Wu
Advisors: Paul Liang, Dr. Louis-Philippe Morency
Carnegie Mellon University

Abstract

Offline Reinforcement Learning promises to learn effective policies from previously-collected, static datasets without the need for exploration. However, existing Q-learning and actor-critic based off-policy RL algorithms fail when bootstrapping from out-of-distribution (OOD) actions or states. We hypothesize that a training curriculum motivated through contrastive learning can improve the models' robustness against OOD backups. The goal of this project is to invent a contrastive training curriculum that is robust against OOD samples for actor-critic algorithms in the context of offline reinforcement learning. Achieving this goal will enable RL agents to better learn from demonstrations and improve the sample efficiency of online RL in general.

1 Introduction

Deep reinforcement learning (RL) has seen a surge of interest over the recent years. It has achieved remarkable success in simulated tasks [15, 14, 5], where the cost of data collection is low. However, one of the drawbacks of RL is its difficulty of learning from prior experiences. Therefore, the application of RL to unstructured real-world tasks is still in its primal stages, due to the high cost of active data collection. It is thus crucial to make full use of previously collected datasets whenever large scale online RL is infeasible.

Offline batch RL algorithms offer a promising direction to leveraging prior experience [9]. However, most prior off-policy RL algorithms [5, 11, 6, 1, 13] fail on offline datasets, even on expert demonstrations [2]. The sensitivity to the training data distribution is a well known issue in practical offline RL algorithms [4, 7, 8, 13, 16]. A large portion of this problem is attributed to actions or states not being covered within the training set distribution. Since the value estimate on out-of-distribution (OOD) actions or states can be arbitrary, OOD value or reward estimates can incur destructive estimation errors that propagates through the Bellman loss and destabilizes training. Prior attempts try to avoid OOD actions or states by imposing strong constraints or penalties that force the actor distribution to stay within the training data [7, 8, 4, 10]. While such approaches achieve some degree of experimental success, they suffer from the loss of generalization ability of the Q function. For example, a state-action pair that does not appear in the training set can still lie within the training set distribution, but policies trained with strong penalties will avoid the unseen states regardless of whether the Q function can produce an accurate estimate of the state-action value. Therefore, strong penalty based solutions often promote a pessimistic and sub-optimal policy. In the extreme case, e.g., in certain benchmarking environments with human demonstrations, the best performing offline algorithms only achieve the same performance as a random agent [2], which demonstrates the need of robust offline RL algorithms.

Since an online actor-critic loss alone cannot handle OOD samples, prior works [3, 4, 7, 16] make use of conditional generative models which can cause more instabilities and do not scale well with input dimension. Moreover, these models are often computationally intense, and waste capacity at modeling the complex relationships in the data x, often ignoring the context c. For example, images may contain thousands of bits of information while the high-level latent variables such as the class label contain much less information (10 bits for 1,024 categories). On the other hand, contrastive learning objective [12] maximizes the mutual information between the encoded representations, and pushes away dissimilar representations.

We hypothesize that a contrastive training curriculum can help the critic distinguish OOD samples from samples that lies in the training set distribution. This should enable the critic to reliably identify and avoid OOD training states or actions.

We aim to answer the following specific research questions from both empirical and theoretical perspectives:

1. Can we design an efficient model and training curriculum that makes best use of contrastive learning to help reduce the effect of OOD samples in offline RL?
2. Can we mathematically show better convergence properties with our contrastive framework?

The outcome of this research will present fundamental theoretical and practical insights in multimodal learning, allowing researchers to design models that capture the benefits of multimodal data sources while accurately considering and mitigating the potential risks involving robustness in the presence of imperfect data and fairness in the presence of biased data.

2 Research Plan

2.1 Project Goals

Yiwei will look into several areas of existing work:

75% Design a contrastive offline RL framework with performance comparable performance to the state-of-the-art [8].

100% (a) Performance surpasses the state-of-the-art [8].

(b) Mathematically show better convergence guarantee than the state-of-the-art [8].

125% Show better performance on off-policy RL (online RL with experience replay buffer)

2.2 Milestones

Yiwei will look into several areas of existing work:

1. **1st Technical Milestone for 15-300:** Finish initial literature review
2. **February 15th:** Formulate and design an initial algorithm
3. **March 1st:** Evaluate the algorithm
4. **March 15th:** Reformulate/evaluate the algorithm accordingly
5. **March 29th:** Reformulate/evaluate the algorithm accordingly
6. **April 12th:** Conduct literature review for convergence guarantees for contrastive learning.
7. **April 26th:** Prove convergence for proposed algorithm using the guarantees for contrastive learning.
8. **May 10th:** Prepare for the presentation of results.

2.3 Resources Needed:

1. Pytorch (opensource)
2. D4RL offline RL benchmarks [2] (opensource)
3. GPU machines (partially covered by SCS clusters)

Advisor Endorsement:

Paul Liang

Signature:_____

Date:_____

Dr. Louis-Philippe Morency

Signature: _____

Date:_____

References

[1] Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *ICML*, 2018.

[2] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[3] S Fujimoto, H van Hoof, D Meger, et al. Addressing function approximation error in actor-critic methods. *Proceedings of Machine Learning Research*, 80, 2018.

[4] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062, 2019.

[5] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870, 2018.

[6] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, pages 651–673, 2018.

[7] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pages 11784–11794, 2019.

[8] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

[9] Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

[10] Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661. PMLR, 2019.

[11] Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1054–1062, 2016.

[12] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[13] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

[14] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[15] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550 (7676):354–359, 2017.

[16] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.