Augmenting datasets for Offline Reinforcement Learning

Project proposal

Gianluca Galletti
Technische Universität München
Department of Informatics
g.galletti@tum.de

Hamza Haddaoui
Technische Universität München
Department of Informatics
hamza.haddaoui@tum.de

I. Objective

Offline reinforcement learning has become a very popular topic in the last couple of years, due to its competitive results when compared to online algorithms and versatility in situations where live interaction with the environment would not be possible. Although the great success, state of the art offline reinforcement learning algorithms seem to tend to overfit to the data used.

The objective of our project is to evaluate the impact of data on generalization and performance of offline RL algorithms. We will use of popular datasets [1] [2] and tackle a wide range of tasks on different environments (dexterous manipulation tasks or autonomous driving - e.g. CARLA).

State of the art DL methods rely on a vast amounts of data to train. This is not yet possible with RL, mainly due to costly data collection (labeling datasets with rewards usually requires human supervision). We want to address this problem by implementing augmentation techniques [3] on the already collected datasets and by generating new data [4] [5]. Finally we want to be able to select the best trajectories among the artificial data using heuristics or value-based approaches.

Further work on the topic could include investigating more advanced generative techniques, and training the model fully on synthetic roll-outs using a world model architecture [6] [7].

II. Related work

`D4RL` [1] provided a great approach on benchmarking and comparing offline reinforcement learning algorithms with special regard towards the type of data used in the training: the dataset includes rollouts of different quality (i.e. poor or good performance) and of different sources (i.e. from non Markovian agents) to provide a more realistical setting, where not all data is ideal and comes from different places.

The idea of data augmentation as a self supervision has already been explored by S4RL [3], although not with the purpose of specifically evaluating the generalization.

`d3rlpy` [8] provides high quality implementations of state of the art offline and online RL algorithms.

`ExORL` [4] presented a way to manufacture rollouts artificially and assign a reward downstream.

III. Technical Outline

Augmentation. Develop a set of augmentation techniques: uniform and Gaussian noise, mixup on states [9], adversarial state training. A more advanced objective would be to come up with problem dependent algorithms for data augmentation.

Algorithms. Develop a simple behavioural cloning agent to act as control. Use BEAR [10], BCQ [11] and CQL [12]. We chose d3rlpy [8] as code base for those algorithms due to its simplicity.

Environments. We will test and benchmark the algorithms on environments for which we have datasets available. We will start with the two famous environments in Mujoco environment: Cheetah and humanoid. Then we will extend the analysis to Atari games suite and finally to manipulator tasks. To see how well the policy generalizes, we should modify environments to evaluate on new situations never seen before (for example, different track in a racing game, one less bunker in space invaders).

Data. Divide some D4RL datasets in a number of subsets, by increasing percentage of original data included. Copy those subsets and perform our augmentation techniques (separate or together). Afterwards, implement generative models to create new trajectories. Finally assign to the trajectories a reward using trained neural networks.

Evaluations. Our idea is to get at the same time the impact of growing available data due to augmentation on performance and generalization, especially when training on ”changed” environments.

Milestones. The key steps for the project are the following:

- Implement data augmentation techniques on the available datasets
- Test and benchmark the selected offline algorithms in the Mujoco environments (half-cheetah, humanoid) using first the standard dataset and then comparing the results with augmented data.
- Extend the analysis to more complex environments by assessing the impact of augmented data on generalization and performance of the models.
- Implement and experiment with heuristic methods for data quality evaluation and benchmark
- Consider training using purely artificial rollouts. This requires a generator to manufacture plausible trajectories either unsupervised or from the available offline rollouts, then label those trajectories with a reward.

REFERENCES

[1] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, Sergey Levine: D4RL: Datasets for Deep Data-Driven Reinforcement Learning, 2020; [http://arxiv.org/abs/2004.07219 arXiv:2004.07219].

[2] Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Tom Le Paine, Sergio Gomez Colmenarejo, Konrad Zolna, Rishabh Agarwal, Josh Merel, Daniel Mankowitz, Cosmin Paduraru, Gabriel Dulac-Arnold, Jerry Li, Mohammad Norouzi, Matt Hoffman, Ofir Nachum, George Tucker, Nicolas Heess, Nando de Freitas: RL Unplugged: A Suite of Benchmarks for Offline Reinforcement Learning, 2020; [http://arxiv.org/abs/2006.13888 arXiv:2006.13888].

[3] Samarth Sinha, Ajay Mandlekar, Animesh Garg: S4RL: Surprisingly Simple Self-Supervision for Offline Reinforcement Learning, 2021; [http://arxiv.org/abs/2103.06326 arXiv:2103.06326].

[4] Denis Yarats, David Brandfonbrener, Hao Liu, Michael Laskin, Pieter Abbeel, Alessandro Lazaric, Lerrel Pinto: Don't Change the Algorithm, Change the Data: Exploratory Data for Offline Reinforcement Learning, 2022; [http://arxiv.org/abs/2201.13425 arXiv:2201.13425].

[5] Tianhe Yu, Aviral Kumar, Yevgen Chebotar, Karol Hausman, Chelsea Finn, Sergey Levine: How to Leverage Unlabeled Data in Offline Reinforcement Learning, 2022; [http://arxiv.org/abs/2202.01741 arXiv:2202.01741].

[6] David Ha, Jrgen Schmidhuber: World Models, 2018; [http://arxiv.org/abs/1803.10122 arXiv:1803.10122]. DOI: [https://dx.doi.org/10.5281/zenodo.1207631 10.5281/zenodo.1207631].

[7] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, Jimmy Ba: Mastering Atari with Discrete World Models, 2020; [http://arxiv.org/abs/2010.02193 arXiv:2010.02193].

[8] Takuma Seno, Michita Imai: d3rlpy: An Offline Deep Reinforcement Library, 2021; [https://github.com/takuseno/d3rlpy].

[9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization, 2017; [https://arxiv.org/abs/1710.09412 arXiv:1710.09412].

[10] Aviral Kumar, Justin Fu, George Tucker, Sergey Levine: Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction, 2019; [http://arxiv.org/abs/1906.00949 arXiv:1906.00949].

[11] Scott Fujimoto, David Meger, Doina Precup: Off-Policy Deep Reinforcement Learning without Exploration, 2018; [http://arxiv.org/abs/1812.02900 arXiv:1812.02900].

[12] Kumar, Aviral and Zhou, Aurick and Tucker, George and Levine, Sergey: Conservative Q-Learning for Offline Reinforcement Learning, 2020; [https://arxiv.org/abs/2006.04779 arXiv:2006.04779].