Research proposal

Pedram Agand

August 2022

1 Autonomous navigation

Mobile robot exploration in a new environment is a common issue in robotic applications. Robots often need complete knowledge of all obstacles and a topological model of the environment in order to use sensor data [14]. To solve the issue of information constraint, machine learning has recently gained a lot of attention. The autonomous agent relies on the data it collects in an effort to extrapolate useful information from the environment that may be applied to decision-making. The truth is that some of the current methods for controlling and guiding the robot to achieve full coverage require a lot of processing and are subject to noise and inconsistent sensor data [5].

One of the most crucial capabilities of autonomous mobile robots is obstacle avoidance. Because it allows the robot to learn from data and does not necessitate a complete comprehension of the world, reinforcement learning (RL) is a powerful technique for reaching this objective. In many safe-critical robotics activities where exploration is expensive and unsafe, offline RL has proven effective. It also has the potential to allow sample-efficient online RL for downstream applications. Studies on the shift from offline to online often discover that likelihood-based strategies, such as sequence modelling objectives, are more effective [15].

Robots that successfully manoeuvre in their space use a variety of sensors such as GPS, LiDARs, ultrasonic sensors, and sometimes a camera, as the primary means of tracking and localization. Ranging sensors only collect a small amount of data, and some of them are too costly or heavy/power-consuming. Monocular cameras, offer details about the robot's working environment, are inexpensive, lightweight, and can be used on a variety of platforms. However, visionary input suffers from uncertainty due to occlusion and precision. Probabilistic methods have shown promise in an uncertain and dynamic world.

Transformers are able to assign credit directly through self-attention, as opposed to Bellman backups which must wait for rewards to spread slowly and are vulnerable to "distractor" signals, as cited by [10]. Because of this, transformers may continue to function efficiently in the midst of limited or distracting rewards. Last but not least, empirical data indicate that a transformer modelling technique may simulate a wide range of behaviours, improving generalisation and transfer.[18].

Transformer-based RL differs significantly from previous offline RL algorithms in that it does not require conservatism or policy regularisation to perform well. Because TD-learning algorithms learn an approximative value function, they can optimise this value function to make the policy better. The process of optimising a learnt function has the potential to aggravate and take advantage of any approximation errors in the value function, leading to ineffective policy improvement. Transformer-based RL does not need regularisation or conservatism because it does not required explicit optimization with learnt functions as the targets [6].

2 Literature

State-of-the-art methods (e.g. [12, 27, 9]) often use re-planning techniques in the presence of dynamic obstacles, which re-call a planning algorithm to look for an alternate direction if the robot finds a new obstacle. However, such approaches sometimes result in unneeded detours [24]. Meanwhile, some approaches (e.g. [26, 1, 20]) consider a deterministic constraint around dynamic obstacles in a robust planning scheme which mainly suffers from performance due to being too conservative. There are a few probabilistic approaches in the literature (e.g. [23, 19]) that require a complete knowledge of the obstacles to be ready at the beginning of the simulation. Not only are not available these level of knowledge in many cases, but also in practice, they could only be available when the robot approaches them (not in advance). To address these problems, one can propose a learning-based approach that uses environmental spatio-temporal knowledge in the form of consecutive local maps.

Human behaviours vary greatly, can take many different forms, and frequently lack reward labels. Due to these characteristics, existing offline RL and behavior cloning (BC) techniques cannot effectively learn from large, precollected datasets [21].

A path planning was used to find a global guidance map considering only static obstacles and planned the immediate steps considering both the static and dynamic obstacles in the agent's field of vision (local observation) along with the use of global guidance map [25]. They defined the reward function such that the agent is not required to follow the fixed guidance map strictly. It Has a reward function independent of the environment, thus enabling scalability as well as generalizability across environments. The decentralized approach cannot handle the local crowded space and generalizes poorly to various environments.

The issue of offline RL for temporally extended tasks is investigated by the authors in [13]. They provide a hierarchical planning system made up of a high-level goal planner and a low level goal-conditioned RL policy. Offline RL is used to train the low-level policies. By using a perturbed goal sampling technique, they enhance offline training to deal with out-of-distribution goals. The high-level planner makes use of model-based planning techniques to choose intermediate sub-goals. It makes preparations for future sub-goal sequences based on the low-level policy's learnt value function. To tackle the high-level long-

term strategy optimization problem and sample meaningful high-dimensional sub-goal candidates, they use a Conditional Variational Autoencoder.

Decision Transformer [6] is a return-conditioned model-free approach that learns a Transformer-based policy that inputs previous states and actions as well as a target return and outputs the action most likely to result in a trajectory sequence that achieves the target return. Choosing an appropriate target return in stochastic situations can be challenging without having a solid understanding of the testing area beforehand. The distribution of potential returns significantly depends on the stochastic changes in the environment, which makes this issue particularly difficult. Using the Decision Transformer architecture, conditional sequence modelling is used to represent the RL problem. Decision Transformer generates the best course of action by utilising a causally masked Transformer, in contrast to earlier approaches to RL that fit value functions or compute policy gradients. They want to investigate whether generative trajectory modelling, which models the joint distribution of the sequence of states, actions, and rewards, can function as a replacement for traditional RL algorithms, in contrast to prior work using transformers as an architectural choice for components within traditional RL algorithms [16].

A model-based technique called Trajectory Transformer [11] trains a trajectory model based on Transformer that may simulate alternative paths in the environment. The primary issue in stochastic environments is how to effectively use this model to look for an appropriate next action or trajectory sequence given the uncontrollable stochastic changes. Without taking into consideration these uncontrollable stochastic transitions, naively applying NLP-style beam search as suggested in the original Trajectory Transformer work will bias the exploration and selection of trajectories to those where the environment just so happens to unfold favourably.

Current results have been attained in a number of the highly deterministic offline Atari and D4RL benchmarks by recent efforts based on this paradigm. These methods, however, find it difficult to separate the impact of the policy and the dynamics of the outside world on the return since they simultaneously describe the states and actions as a single sequencing problem. Therefore, these techniques produce excessively optimistic behaviour in adversarial or stochastic contexts, which might be risky in safety-critical systems like autonomous vehicles. The authors in [22] establish a technique called SeParated Latent Trajectory Transformer (SPLT Transformer), which develops separate generative Transformer models for the policy and dynamics, in order to address this optimistic bias in past techniques. The reason they express each of these models as discrete latent variable Variational Auto-Encoders is that they concentrate on the autonomous driving domain (VAE). However, learning complex generative models necessitates extensive training and precise hyper parameter tweaking due to the modelling of high-dimensional, continuous multi-modal action distributions [21].

3 Problem statement

The goal is to output an action at each time step that enables the robot to move from any start location to the goal cell with the least expected number of steps while avoiding conflicts with static and dynamic obstacles. The inputs are the global map (from which we compute the preferred set of way points), the occupancy map (position of static obstacles), and the local observation of dynamic obstacles.

4 Proposed algorithm

We provide a hierarchical path planning system made up of a high-level goal planner and a low-level goal-conditioned RL policy. We make use of the GPT design described in [17] Improving, which replaces the summation/softmax over the n tokens with just the prior tokens in the sequence in order to enable autoregressive generation. We simplify policy sampling to autoregressive generative modelling by training an autoregressive model on sets of states, actions, and returns. By choosing the desired return tokens, which serve as a prompt for creation, we can describe the expertise of the policy—which "skill" to query—by doing so.

Different from existing hierarchical learning-based methods, we introduce **T**ransformer **I**ntegrated **L**ocal **M**ap **A**ssisted **P**ath-planning (TILMAP), which incorporates a novel state representation, and sequence problem to arbitrary environments. Global guidance map is the preferred set of way points throughout the run that is obtained via the low-level offline policy. Generally, choice of global guidance map is not restrictive; it can be either obtained from an optimizer or handcrafted to satisfies the task. Local probabilitic occupancy map are the likelihood that an obstacle is in that particular cell, which is obtained via the CNN.

As the robot's global guidance does not necessarily have to be precise; inaccuracy can be caught in a form of uncertainty. This allows the robot to compromise between exploitation and exploration that uplift *flexibility*. TILMAP models the distribution of returns, making it a generative model as opposed to imitation learning. By employing all of the dataset's trajectories to improve generalisation, even if they differ from the return conditioning aim, it is anticipated to surpass imitation learning (such as BC). We use k-means to cluster continuous activities into discrete bins, drawing inspiration from Behavior Transformers [21]. This is due to the intrinsic applicability of transformer-based sequence models for predicting discrete classes. We learn a residual action corrector to generate continuous actions for a sampled action bin in order to guarantee that the sampled actions are useful for online rollouts.

5 Previous contributions

Using a Bayesian paradigm, we proposed adaptive recursive MCMC (ARMCMC) algorithm for online estimation of the complete probability distribution of model parameters [3]. Conventional methods work poorly when applied to systems involving sudden shifts in model parameters, which may occur in hybrid systems. The proposed approach adapts easily to sudden shifts, allowing the method to be applicable to a wider range of systems.

We proposed HNIPO, a probabilistic model and updating framework to predict human future states [4]. It is used to calculate the distribution of model parameters and individual states over time. Adding velocity and heading to the dynamics is supposed to make a difference between close and far targets appearing in the same direction. An importance sampling method is proposed to estimate the likelihood of the state, and the goal to handle the algorithm's complexity given the 4D model and three tuning parameters.

Although monocular distance estimation has been studied in the literature, previous methods mostly rely on knowing an object's class in some way. This can result in deteriorated performance for dataset with multi-class objects and objects with an undefined class. We aim to overcome the potential downsides of class-specific approaches, and provide an alternative technique called DMODE that does not require any information relating to its class [2]. Using differential approaches, we combine the changes in an object's size over time together with the camera's motion to estimate the object's distance. Since DMODE is class agnostic method, it is easily adaptable to new environments.

6 More details

The environment is represented as a discrete 2-dimensional grid that contains both known static and unknown dynamic obstacles. We assume that the environment is represented on a map that includes the car's current location, its desired location (destination), and any stationary obstacles. The occupancy map also explicitly depicts static obstructions, with a "1" indicating an obstruction at the associated grid point and a "0" indicating free space. We assume that the dynamic obstacles are *a priori* unknown, and become probabilistically known only if robot is within some distance threshold to them. In practice, the moving obstacles are either deterministic but not fully known like blocks in the roud, or stochastic like other cars or pedestrians. For the first case, one can use *ARMCMC* to render full probability distribution of the other agents position in realtime. As in the second case, *HNIPO* can be employed to enable an inference framework to estimate the possible future states of the pedestrians. Both of these knowledge are only obtained if the robot get close to the target and gather some information from it. These make TILMAP to be a more *realistic* approach.

The Unreal Engine is used by CARLA [7] to deliver a simulated driving scenario in a realistic environment. While the observation space is a (224,224,3)-

dimensional RGB image of the car, the agent action space is 2D (accelerate/brake and left/right steer). A total of 100 examples operate in two different modes as they circle a building block. The difficulty of inferring behaviour from high dimension observations is highlighted by this setting. We utilise a frozen ResNet-18 encoder, pretrained on ImageNet, for visual observations with VILMAP, and the offline dataset obtained from D4RL [8].

just some idea: https://arxiv.org/pdf/2203.10638.pdf

References

[1] Baqir Nasser AbdulSamed, Ammar A Aldair, and Auday Al-Mayyahi. Robust trajectory tracking control and obstacles avoidance algorithm for quadrotor unmanned aerial vehicle. *Journal of Electrical Engineering & Technology*, 15(2):855–868, 2020.

[2] Pedram Agand, Michael Chang, and Mo Chen. Dmode: Differential monocular object distance estimation module without class specific information. *arXiv preprint arXiv:2210.12596*, 2022.

[3] Pedram Agand, Mo Chen, and Hamid D Taghirad. Online probabilistic model identification using adaptive recursive mcmc. *arXiv preprint arXiv:2210.12595*, 2022.

[4] Pedram Agand, Mahdi Taherahmadi, Angelica Lim, and Mo Chen. Human navigational intent inference with probabilistic and optimal approaches. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8562–8568. IEEE, 2022.

[5] Omar Boufous. Deep reinforcement learning for complete coverage path planning in unknown environments, 2020.

[6] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

[7] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[8] 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.

[9] Mehmet Hasanzade and Emre Koyuncu. A dynamically feasible fast replanning strategy with deep reinforcement learning. *Journal of Intelligent & Robotic Systems*, 101(1):1–17, 2021.

[10] Chia-Chun Hung, Timothy Lillicrap, Josh Abramson, Yan Wu, Mehdi Mirza, Federico Carnevale, Arun Ahuja, and Greg Wayne. Optimizing agent behavior over long time scales by transporting value. *Nature communications*, 10(1):1–12, 2019.

[11] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286, 2021.

[12] Geesara Kulathunga, Roman Fedorenko, Sergey Kopylov, and Alexandr Klimehik. Real-time long range trajectory replanning for mavs in the presence of dynamic obstacles. In *2020 5th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*, pages 145–153. IEEE, 2020.

[13] Jinning Li, Chen Tang, Masayoshi Tomizuka, and Wei Zhan. Hierarchical planning through goal-conditioned offline reinforcement learning. *arXiv preprint arXiv:2205.11790*, 2022.

[14] Ming Liu, Francis Colas, Luc Oth, and Roland Siegwart. Incremental topological segmentation for semi-structured environments using discretized gvg. *Autonomous Robots*, 38(2):143–160, 2015.

[15] Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

[16] Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, et al. Stabilizing transformers for reinforcement learning. In *International conference on machine learning*, pages 7487–7498. PMLR, 2020.

[17] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[18] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[19] Ankit A Ravankar, Abhijeet Ravankar, Takanori Emaru, and Yukinori Kobayashi. Hpprm: Hybrid potential based probabilistic roadmap algorithm for improved dynamic path planning of mobile robots. *IEEE Access*, 8:221743–221766, 2020.

[20] Baskın Şenbaşlar, Wolfgang Hönig, and Nora Ayanian. Robust trajectory execution for multi-robot teams using distributed real-time replanning. In *Distributed Autonomous Robotic Systems*, pages 167–181. Springer, 2019.

[21] Nur Muhammad Mahi Shafiullah, Zichen Jeff Cui, Ariuntuya Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone. *arXiv preprint arXiv:2206.11251*, 2022.

[22] Adam R Villaflor, Zhe Huang, Swapnil Pande, John M Dolan, and Jeff Schneider. Addressing optimism bias in sequence modeling for reinforcement learning. In *International Conference on Machine Learning*, pages 22270–22283. PMLR, 2022.

[23] Abraham P Vinod and Meeko MK Oishi. Probabilistic occupancy function and sets using forward stochastic reachability for rigid-body dynamic obstacles. *arXiv preprint arXiv:1803.07180*, 2018.

[24] Binyu Wang, Zhe Liu, Qingbiao Li, and Amanda Prorok. Mobile robot path planning in dynamic environments through globally guided reinforcement learning. *IEEE Robotics and Automation Letters*, 5(4):6932–6939, 2020.

[25] Binyu Wang, Zhe Liu, Qingbiao Li, and Amanda Prorok. Mobile robot path planning in dynamic environments through globally guided reinforcement learning. *IEEE Robotics and Automation Letters*, 5(4):6932–6939, 2020.

[26] Xiaolin Zhao, Yu Zhang, and Boxin Zhao. Robust path planning for avoiding obstacles using time-environment dynamic map. *Measurement and Control*, 53(1-2):214–221, 2020.

[27] Boyu Zhou, Fei Gao, Jie Pan, and Shaojie Shen. Robust real-time uav replanning using guided gradient-based optimization and topological paths. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1208–1214. IEEE, 2020.