Foundation model for Reinforcement Learning

Volodymyr Mudryi

Independent Researcher, Lviv, Ukraine vovamudruy@gmail.com

1 Introduction

The focus of my Ph.D. research proposal centers around an innovative approach within the domain of Reinforcement Learning, specifically focusing on the concept of building an In-Context model via Policy Distillation. This approach involves sequential modeling, notably the transformers architecture, in combination with offline reinforcement learning. The central objective is to thoroughly investigate and advance the methodology of training in-context foundation model - Causal Model.

This model is trained using a large dataset that captures agents' historical interactions with a dynamic environment while solving various tasks. This dataset is commonly accumulated during the training process of a basic RL algorithm. As a result, the trained Causal Model enables autoregressively predict actions based on prior learning histories.

This method offers several distinct advantages in contrast to conventional Reinforcement Learning (RL) approaches. Importantly, it enhances data efficiency compared to models that generated the source training data. Moreover, it leverages the architecture of transformer models, thereby affording access to a multitude of effective optimization techniques [1].

This research holds significance in potentially revolutionizing task learning in Reinforcement Learning. By transitioning from isolated RL policy models for each task to a more comprehensive approach focused on "learning how to learn," and building foundation models. These models can subsequently be more adaptive and efficient via fine-tuning using minimal task-specific data, which leads to a reduction in the temporal and resource requirements associated with training RL solutions.

The potential applications of this research extend across various domains, offering solutions to challenges and complexities in high-risk sectors such as medicine. By capitalizing on the capacity for offline reinforcement learning within authentic environments, this methodology can use the treatment histories of real patients, guided by actions prescribed by medical professionals to train the model and then use it to enhance medical decision-making processes utilize multi-context robustness to dynamical environments.

Moreover, this method can be used in domains with resource-intensive environments, such as goal-oriented conversational chatbots. Constructing elaborate systems for agents to interact with real individuals incurs substantial costs. However, through offline reinforcement learning, historical conversational data can be distilled, facilitating the development of a robust model capable of not only

optimizing diverse end goals across varied contexts but performing well in other environment-related tasks.

The potential applications are not confined to these scenarios; they extend to domains that require quick policy adaptations within familiar environments. This appears as a use case for a foundational model, fine-tuned with limited expert interaction data. These applications encompass domains like trading, portfolio optimization, and fraud detection, where the ability to rapidly adapt policies while harnessing minimal expert input can yield enhanced performance and effective decision-making.

2 The current state of research

Recent studies demonstrate the potential of learning policies from offline data using imitation learning and sequential modeling. Noteworthy contributions demonstrate the effectiveness of such approaches in accomplishing singular tasks, employing architectural frameworks like the Decision Transformer [2] and Trajectory Transformer [3]. Additionally, these methodologies exhibit promise in addressing multi-task problems within a unified domain, exemplified by the Multi-Goal Decision Transformer (MGDT) [4] approach applied to Atari games [5]. Moreover, the power of transformers is exemplified by their role in constructing a unified generalist agent (Gato) [6], capable of navigating diverse modalities, tasks, and embodiments.

However, these methodologies face limitations in augmenting policies through additional interaction with the environment. MGDT's dependence on fine-tuning for new tasks which require resource-intensive data collection, prevents its practical viability. Instead, the Gato framework requires expert prompting for learning new tasks, which can be expensive.

The latest advancements in the field of Policy Distillation [7] provide an interesting hypothesis about the difficulties in policy improvement. They suggest that the challenge in policy improvement stems from the lack of policy progress within the training data. This occurs because of the short training context window, which fails to capture the learning dynamics between episodes. To mitigate this issue, the study introduces the Algorithm Distillation method [7], leveraging the historical learning of the RL algorithm across various tasks. This methodology entails training a causal transformer network through the optimization of a causal sequence prediction loss.

Significantly, Algorithm Distillation [7] not only rectifies this limitation but also showcases remarkable out-of-distribution generalization capabilities. This observation makes a promising expectation, suggesting the potential for policy distillation to yield a universal and robust model capable of performing in dynamic environments. However, this method shared a problem of using a large enough contextual window size in transformer input. As indicated in [7] a substantial correlation exists between window size and performance. Consequently, a crucial task lies in developing efficient strategies to process long window inputs, since most modern reinforcement learning tasks involve long episodes.

3 Motivation

This research proposal is motivated by the need to address critical challenges in the Reinforcement Learning domain specific in-context foundation models. Based on existing research gaps, this study aims to address the following key research questions.

3.1 Efficient Handling of Large Window Inputs

I will explore methods to efficiently process large window inputs, considering alternative sequential architectures like LSTM [8] for improved memory efficiency. Another direction of investigation involves exploring the optimization of input efficiency through modern attention mechanisms. For instance, Longformer [9] presents an attention mechanism that scales linearly with input sequence length. Similarly, Nystromformer [10] introduces the application of the Nyström method for approximating self-attention matrices. These techniques offer promising pathways for enhancing the handling of larger inputs in our research.

3.2 Semi-Supervised Learning for Data Subsampling and Distillation

The research will utilize semi-supervised learning techniques for subsampling significant states, actions, and rewards from large windows of input. By distilling crucial information, this approach seeks to accelerate model training while maintaining or improving performance, potentially using aggregation mechanisms or hierarchical representations layer before transformer attention layers.

3.3 Self-Generated Data for Enhanced Training

The research will explore the concept of using trained models to self-generate additional data for offline RL training. In [11] proposed leveraging the trained model to generate more offline data to further boost the sequence model training.

These research questions are carefully chosen to address gaps in the current state of research, as detailed in the proposal. By tackling these questions, the research not only aims to contribute scientifically and technically to the RL field but also aligns with motivations of pushing boundaries and leaving a lasting impact on the advancement of Reinforcement Learning.

4 Methodology

The proposed research will employ a structured methodology to investigate and advance the concept of building an In-Context model via Policy Distillation within the domain of Reinforcement Learning.

4.1 Literature Review and Theoretical Foundation

My approach begins with a comprehensive review of literature encompassing Reinforcement Learning, sequential modeling, and transformer architectures. This pivotal step will lay a robust theoretical groundwork, to work with in-context foundation models, policy distillation, and their interconnected techniques.

4.2 Performance Evaluation of LSTM Architecture

The next step will be the rigorous assessment of the LSTM [8] architecture's viability as a Causal Model to effectively handle extensive window inputs. I will conduct a comparative study, evaluating the performance, speed, and memory efficiency of the LSTM solution against the established Transformer approach.

4.3 Integration of Modern Contextual Methods

The methodology will embrace the integration of modern techniques derived from Large Language models, notably Longformer [9] and Nystromformer [10], to address large window contexts. Implementing these methods practically will enable empirical testing on benchmark tasks, like Adversarial Bandit and Dark Room [7]. This will provide a valuable comparison with existing Foundation model strategies and fundamental RL-Algorithms such as off-policy DQN [12] and A3D [13].

4.4 Leveraging Semi-Supervised Learning for Distillation

An important step is the exploration of Semi-Supervised Learning techniques for distilled data subsampling. By selectively distilling states, actions, and rewards from extensive input windows, the goal is to shape a simplified training dataset. This phase might involve using Triplet loss [14] or other Contrastive Loss mechanisms [15], reducing the need for complex data labeling while improving the model's ability to capture essential embeddings.

4.5 Scaling to Complex Environments

If promising results are obtained, there is the potential to allocate additional resources toward training the foundational model in more intricate pixel-based environments. An example of such an environment is the extensive collection of Atari games [5]. This extension aims to test the model's adaptability to complex and diverse scenarios.

4.6 Self-Generated Data for Augmentation

An exploration into the utilization of trained models for self-generating supplementary offline data is another integral part of the methodology. This iterative process will ensure the generated data aligns seamlessly with the inherent distribution of the environment, thus improving the training dataset.

4.7 Comprehensive Discussion and Future Paths

As the methodology culminates, a comprehensive discussion will ensue, interlinking the findings with the initial research questions and the prevailing research landscape. The implications for the field of Reinforcement Learning, particularly for in-context foundation models, will be explored in detail. This will lay the groundwork for potential avenues of future research.

The iterative nature of this methodology will allow the fine-tuning and optimization of the proposed strategies. As a machine learning research, this approach signifies a proactive effort, aimed at contributing to the field of Reinforcement Learning through innovative insights.

5 Conclusion

In summary, this Ph.D. research proposal presents an innovative approach to enhance existing Reinforcement Learning techniques by refining In-Context models through Policy Distillation. By leveraging sequential modeling, transformers, and offline reinforcement learning, the objective is to bridge gaps, improve data efficiency, and speed-up transition from isolated policy models approach to a comprehensive "learning how to learn" framework.

The research questions target key gaps, aiming to optimize the handling of large windows, implement semi-supervised data distillation, and adapt the usage of self-generated data. This proactive strategy aligns to advance Reinforcement Learning while building upon established foundation models.

The structured methodology combines established and novel techniques, promising transformative insights into refining in-context models. The possible success holds the potential to contribute to revolutionizing approaches in the Reinforcement Learning landscape, optimizing resource utilization, and fostering adaptable solutions across diverse domains.

References

1. Amoiralis, Eleftherios I., Marina A. Tsili, and Antonios G. Kladas. "Transformer design and optimization: a literature survey." IEEE Transactions on power delivery 24.4 (2009): 1999-2024.
2. Chen, Lili, et al. "Decision transformer: Reinforcement learning via sequence modeling." Advances in neural information processing systems 34 (2021): 15084-15097.
3. Janner, Michael, Qiyang Li, and Sergey Levine. "Offline reinforcement learning as one big sequence modeling problem." Advances in neural information processing systems 34 (2021): 1273-1286.
4. Lee, Kuang-Huei, et al. "Multi-game decision transformers." Advances in Neural Information Processing Systems 35 (2022): 27921-27936.
5. Bellemare, Marc G., et al. "The arcade learning environment: An evaluation platform for general agents." Journal of Artificial Intelligence Research 47 (2013): 253-279.
6. Reed, Scott, et al. "A generalist agent." arXiv preprint arXiv:2205.06175 (2022).

7. Laskin, Michael, et al. "In-context reinforcement learning with algorithm distillation." arXiv preprint arXiv:2210.14215 (2022).
8. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.
9. Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." arXiv preprint arXiv:2004.05150 (2020).
10. Xiong, Yunyang, et al. "Nyströmformer: A nyström-based algorithm for approximating self-attention." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 16. 2021.
11. Wang, Kerong, et al. "Bootstrapped transformer for offline reinforcement learning." Advances in Neural Information Processing Systems 35 (2022): 34748-34761.
12. Mnih, Volodymyr, et al. "Playing atari with deep reinforcement learning." arXiv preprint arXiv:1312.5602 (2013).
13. Asynchronous Methods for Deep Reinforcement Learning
14. Hoffer, Elad, and Nir Ailon. "Deep metric learning using triplet network." Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3. Springer International Publishing, 2015.
15. Khosla, Prannay, et al. "Supervised contrastive learning." Advances in neural information processing systems 33 (2020): 18661-18673.