SCIENCE CHINA
Information Sciences

• LETTER •

June 2025, Vol. 68, Iss. 6, 169101:1–169101:2
https://doi.org/10.1007/s11432-024-4353-9

Graph decision transformer for offline reinforcement learning

Shengchao HU[1,2], Li SHEN[3*], Ya ZHANG[1,2] & Dacheng TAO[4]

[1]*School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*
[2]*Shanghai Artificial Intelligence Laboratory, Shanghai AI Laboratory, Shanghai 200233, China*
[3]*School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China*
[4]*School of Computer and Data Science, Nanyang Technological University, Singapore 639798, Singapore*

Received 20 October 2024/Revised 3 January 2025/Accepted 17 March 2025/Published online 13 May 2025

Citation Hu S C, Shen L, Zhang Y, et al. Graph decision transformer for offline reinforcement learning. Sci China Inf Sci, 2025, 68(6): 169101, https://doi.org/10.1007/s11432-024-4353-9

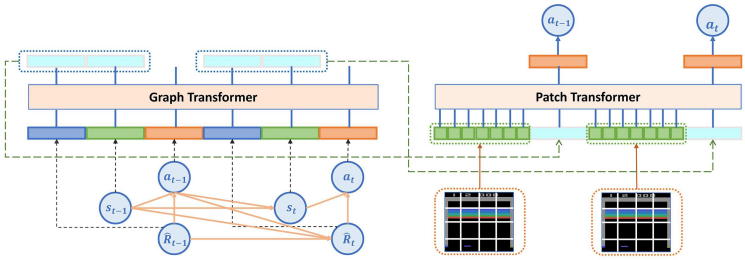
Recent advances [1,2] in offline reinforcement learning (RL) have taken a new perspective on the problem, departing from conventional methods that concentrate on learning value functions or policy gradients. Instead, the problem is viewed as a generic sequence modeling task, where past experiences consisting of state-action-reward triplets are input to the Transformer. The model generates a sequence of action predictions using a goal-conditioned policy, effectively converting offline RL to a supervised learning problem. This approach relaxes the Markov decision process assumption by considering multiple historical steps to predict an action, allowing the model to be capable of handling long sequences. Furthermore, this framework unifies multiple components in offline RL, such as estimating the behavior policy and predictive dynamics modeling, into a single sequence model, resulting in superior performance.

However, this approach faces three major issues. First, states and actions represent fundamentally different concepts. While the agent has complete control over its action sequences, the resulting state transitions are often influenced by external factors. Thus, modeling states and actions as a single sequence may indiscriminate the effects of the policy and world dynamics on the return, which can lead to overly optimistic behavior [3]. Second, in RL problems, the adjacent states, actions, and rewards are typically strongly connected due to their potential causal relationships. Specifically, the state observed at a given time step is a function of the previous state and action, and the action taken at that time step influences the subsequent state and reward. Simply applying Transformer to attend to all tokens without considering the underlying Markovian relationship can result in an overabundance of information, hindering the learning process in accurately capturing essential relation priors and handling long-term sequences of dependencies from scratch [4]. Finally, tokenizing image states as as-a-whole using convolutional neural networks can hinder the ability of Transformers to gather fine-grained spatial relations. This loss of information can be critical in visual RL tasks that require detailed knowledge of regions-of-interest. Therefore, it is necessary to find a more effective way to represent states and actions separately while still preserving their intrinsic relationships, and to incorporate the Markovian property and spatial relations in the modeling process.

To alleviate such issues, we propose a novel approach, namely graph decision transformer (GDT), which involves transforming the input sequence into a dependency graph structure. The graph representation explicitly incorporates the potential dependencies between adjacent states, actions, and rewards, thereby introducing a Markovian-like inductive bias to the learning process and differentiating the impact of different tokens. To process the input graph, we utilize the graph transformer to effectively handle long-term dependencies that may be present in non-Markovian environments. To gather fine-grained spatial information, we incorporate an optional patch transformer to encode image-based states as patches similar to vision transformer, which helps with action prediction and reduces the learning burden of the graph transformer. Our experimental evaluations conducted in Atari benchmark environments provide empirical evidence to support the advantages of utilizing a dependency graph representation as input to the graph transformer in RL tasks. The proposed GDT method achieves state-of-the-art performance in several benchmark environments and outperforms most existing offline RL methods without incurring additional computational overhead.

Methods. The proposed approach leverages both graph and sequence modeling techniques to create a deep learning-based model for offline RL tasks. The model consists of three main components: the graph representation, the graph transformer, and an optional patch transformer, as shown in Figure 1(a). Specifically, the graph representation is used to represent the input sequence as a graph with a dependency relationship, thereby better capturing the Markovian property of the input and differentiating the impact of different tokens. The graph transformer then processes the graph inputs using the relation-enhanced mechanism, which allows

* Corresponding author (email: mathshenli@gmail.com)



(a)

Game	CQL	QR-DQN	REM	BC	DT	GDT	StAR	GDT-plus
Breakout	211.1	17.1	8.9	10.4	267.5 ± 97.5	393.5 ± 98.8	436.1 ± 40.0	**441.7 ± 41.0**
Qbert	**104.2**	0	0	17.3 ± 14.7	15.4 ± 11.4	45.5 ± 14.6	51.2 ± 11.5	51.7 ± 20.8
Pong	95.8	18	0.5	20.3	106.1 ± 8.1	108.4 ± 4.7	110.8 ± 4.8	**111.2 ± 4.6**
Assault	40.4	-17.3	24.6	42.3	46.2 ± 16.0	87.9 ± 5.3	103.7 ± 18.3	**104.1 ± 15.3**
Seaquest	1.7	0.4	0.7	2.1	2.5 ± 0.4	2.8 ± 0.1	1.7 ± 0.3	**2.7 ± 0.1**
Average	90.6	3.6	6.9	18.5	87.5	127.6	140.7	**142.3**

(b)

Figure 1 (Color online) (a) The proposed model comprises three main components: the graph representation, the graph transformer, and an optional patch transformer. When employing the direct output of the graph transformer for action prediction, the resultant model is denoted as GDT, representing the left half of the depicted figure. Alternatively, if the output of the graph transformer undergoes additional processing by the patch transformer, the resulting model is identified as GDT-plus, encompassing the entire figure. (b) Results for 1% DQN-replay Atari datasets. We evaluate the performance of GDT on five Atari games using three different seeds, and report the mean and variance of the results. The best mean scores are in bold.

the model to acquire long-term dependencies and model the interactions between different time steps of the graph tokens given dependency relationships. The optional patch transformer is introduced to gather the fine-grained spatial information in the input, which is particularly important in visual tasks such as the Atari benchmark.

The proposed approach offers several advantages. First, it can effectively acquire the intricate dependencies and interactions between different time steps in the input sequence, making it well-suited for RL tasks. Second, it encodes the sequence as a dependency graph, which explicitly incorporates potential dependencies between adjacent tokens, thereby explicitly introducing the Markovian bias into the learning process and avoiding homogenizing all tokens. Finally, it can accurately gather fine-grained spatial information and integrate it into action prediction, leading to improved performance. Please refer to the appendix for more detailed methods.

Experiments. We conduct a comprehensive evaluation of the performance of the GDT model on a range of tasks. Specifically, we evaluate the performance of GDT on the widely used Atari benchmark [5], which consists of a set of discrete control tasks. Note that we refer to the model as GDT-plus when incorporating the patch transformer, and we categorize these methods into three groups based on the employed approach category and the number of parameters to ensure a fair comparison.

Figure 1(b) presents the comparison of our proposed method with these offline baselines on five games. The results demonstrate that our method achieves comparable or superior performance in all five games. This indicates that our GDT, which incorporates dependency relationships in the input and leverages the graph transformer accordingly, outperforms temporal difference-learning and sequence modeling methods. The observed performance highlights the significance of the Markovian bias in the input for the decision-making process. To ensure a fair comparison with StAR, we further introduce a patch transformer to incorporate fine-grained spatial information and report the results as GDT-plus. The results demonstrate that GDT-plus achieves comparable or superior performance to StAR on all five Atari games, emphasizing the significance of fine-grained information on these games. Furthermore, when compared to GDT, the effectiveness of the patch transformer in incorporating spatial information into action prediction is further highlighted. Please refer to the appendix for all detailed results.

Conclusion. In summary, this study introduces GDT, an innovative offline RL methodology that transforms input sequences into causal graphs. This approach effectively captures potential dependencies between distinct concepts and enhances the learning of temporal and causal relationships. The empirical results presented demonstrate that GDT either matches or outperforms existing state-of-the-art offline RL methods across image-based Atari benchmark tasks. Our work highlights the potential of graph-structured inputs in RL, which has been less explored compared to other deep learning domains. We believe that our proposed GDT approach can inspire further research in offline RL, especially in tasks where spatial and temporal dependencies are essential, such as robotics, autonomous driving, and video games.

Acknowledgements This work was supported by STI 2030—Major Projects (Grant No. 2021ZD0201405), Shenzhen Basic Research Project (Natural Science Foundation) Basic Research Key Project (Grant No. JCYJ20241202124430041).

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Chen L, Lu K, Rajeswaran A, et al. Decision transformer: reinforcement learning via sequence modeling. In: Proceedings of Advances in Neural Information Processing Systems, 2021
2 Janner M, Li Q, Levine S. Offline reinforcement learning as one big sequence modeling problem. In: Proceedings of Advances in Neural Information Processing Systems, 2021
3 Villaflor A R, Huang Z, Pande S, et al. Addressing optimism bias in sequence modeling for reinforcement learning. In: Proceedings of International Conference on Machine Learning, 2022
4 Shang J, Kahatapitiya K, Li X, et al. StARformer: transformer with state-action-reward representations for visual reinforcement learning. In: Proceedings of European Conference on Computer Vision, 2022
5 Bellemare M G, Naddaf Y, Veness J, et al. The arcade learning environment: an evaluation platform for general agents. J Artif Intell Res, 2013, 47: 253–279