State trajectory abstraction and visualization method for explainability in reinforcement learning

Yoshiki Takagi [1 2] **Roderick Tabalba** [1] **Jason Leigh** [1]

Abstract

Explainable AI (XAI) has demonstrated the potential to help reinforcement learning (RL) practitioners to understand how RL models work. However, XAI for users who have considerable domain knowledge but lack machine learning (ML) expertise, is understudied. Solving such a problem would enable RL experts to communicate with domain experts in producing ML solutions that better meet their intentions. This study examines a trajectory-based approach to the problem. Trajectory-based XAI appears promising in enabling non-RL experts to understand a RL model's behavior by viewing a visual representation of the behavior that consists of trajectories that depict the transitions between the major states of the RL models. This paper proposes a framework to create and evaluate a visual representation of RL models' behavior that is easy to understand between both RL and non-RL experts.

1. Introduction

Reinforcement learning (RL) has evolved rapidly in the past decade and is now capable of achieving human capabilities, such as self-driving cars. Moreover, in the last few years, the performance of deep RL, which applies deep neural networks to RL, has surpassed that of skilled human players in areas of video games, chess, and Go. However, as the performance of deep RL models increases, the complexity of the model also increases. As a result, understanding and interpreting the models becomes a challenging problem for humans.

Explainable AI (XAI) research has shown the potential to close the gap between humans and RL models by providing explanations that help users understand how RL models work. So far, various aspects of explainability for both RL practitioners and non-RL practitioners has been illuminated. For RL practitioners, these aspects involve revealing the learning mechanism of a deep RL agent by enabling users to examine the internal parameters of an agent (Strobelt et al., 2017; Ming et al., 2017; Jaunet et al., 2020). There has been research in investigating how RL agents observe the state of the environment by using an attribution method such as saliency maps (Puri et al., 2019; Atrey et al., 2019; Rupprecht et al., 2019; Sundararajan & Najmi, 2020). Lastly, research has also looked into replacing a deep RL model with a more transparent model by using model distillation techniques (Bastani et al., 2018; Coppens et al., 2019; Liu et al., 2019). Although these methods provide an explanation of the model, they require knowledge of RL to understand. This makes it difficult for non-RL users to explore these methods. Other aspects of explainability for non-RL practitioners have been explored in the field of Human Robot Interaction. These methods include explaining the policy of an agent (Hayes & Shah, 2017; Struckmeier et al., 2019), justifying the action of an agent with reward (Tabrez et al., 2019; Juozapaitis et al., 2019), and explaining the dynamics of an environment '(Elizalde et al., 2007; Khan et al., 2009). Recently, counterfactual explanations that explain "*If A did not happen, B would not have happened*" and contrastive explanations that answering "*Why* B_1 *rather than* B_2*?*" have been actively considered as good explanations that are understandable for a wide variety of users (van der Waa et al., 2018; Madumal et al., 2020; Olson et al., 2021).

However, the gap between RL practitioners and non-RL practitioners still needs much attention. As a result, RL and non-RL experts try to understand the RL model using different XAI approaches, communication problems are likely to happen in the discussion of RL model's behavior. For instance, when discussing the proper decision boundary of autonomous UAVs between RL practitioners and pilots, the primary issues are as follows:

- Pilots, who are non-RL experts, have domain knowledge, but they cannot use XAI interfaces designed for RL experts in the assessment of the RL model;
- In order to obtain feedback from pilots, RL experts

[1]Information and Computer Science, University of Hawaii at Manoa, Honolulu, HI, USA [2] Acquisition, Technology & Logistics Agency, Ministry of Defense, Tokyo, Japan. Correspondence to: Yoshiki Takagi <takagiyo@hawaii.edu>.

AI & HCI Workshop at the 40^{th} *International Conference on Machine Learning (ICML)*, Honolulu, Hawaii, USA. 2023.

need to explain the behavior of the RL model while minimizing the use of RL terminology;

- Pilots may use domain specific terminology during the assessment and the RL expert needs to interpret the pilot's statements and apply them to the model;

Furthermore, the current XAI approaches for non-RL experts generally require a deep domain knowledge to understand. Not only does this make it difficult for RL-experts to communicate with non-RL experts, this also affects lay users without the domain knowledge to participate in the fundamental discussion of how RL models should be designed. For example, in a conventional automobile development, a prototype vehicle designed by designers is evaluated by a test driver with in-depth knowledge of driving, and modified by the designers as necessary. However, discussions on the more fundamental question of how a self-driving car should be driven in the first place, should not be conducted among a few group of people such as developers and test drivers, but should also involve the general public more broadly. This indicates the need for a new XAI approach that requires less descriptive knowledge to build the mental model of RL models.

One approach that may bridge the knowledge gap between a wide range of users is to use trajectories that represent the experience of an agent. With this approach, users can build a mental model of an agent by observing the agent's trajectory as a visualization. This is akin to estimating the personality of the driver by observing their driving performance from video recordings. This approach allows the user to make inferences based on their observations. Furthermore, this also does not require a deep understanding of RL, whereas other approaches do. Despite the advantages of using trajectories for XAI, there is little prior work on this approach in the context of XAI in RL. Amir *et al.* indicated that users are able to assess the proficiency level of agents by observing the summarized state trajectory of the agents (Amir & Amir, 2018).

The compact non-descriptive representation of the mental model acquired by the trajectory based approach can be used for decision making involving ethical issues. One example is body-worn cameras for police accountability (Coudert et al., 2015). When looking at a record of a body-worn camera of a contested situation, we can infer the mindset of the police officers in question, and assess whether their actions were justified.

Early insights into using a trajectory-based approach shows us that its advantage in the context of RL is gradually being recognized. However, the question of *can users more efficiently build the mental model of agents by abstracting trajectories, and how to visualize the abstracted trajectory to support a user's intuitive inference of the agent's behavior* remains open. The purpose of this paper is to propose a fundamental framework for the creation of an abstracted state trajectory visualization for various types of agents and a user study to evaluate user's comprehension when observing the abstracted state trajectory visualization. Specifically, the focus of this study is to evaluate the explainability of an abstracted state trajectory represented as a visualization.

2. Related Work

This study builds on three important streams of related work. In this section, we describe these in chronological order, that is, the study of state aggregation, which is the basis of state trajectory abstraction, the study of the adaptation of abstracted trajectories to improve explainability in the context of RL, and the study of conducting user study to verify whether trajectory visualization can actually provide explainability in the context of RL.

State aggregation is the oldest and a well-studied problem. The state aggregation is a process that replacing states S in the Markov Decision Process (MDP) with the set of clusters C identified by the similarity of the states S. In the 1990s, Singh *et al.* proposed state aggregation algorithms to address the curse of dimensionality that arose with scaling RL (Singh et al., 1994). They indicated that state aggregation can benefit agent learning since the transition probability matrix P, reward signal R, and policy dimensions decrease.

Recently, Zahvy *et. al.* proposed Semi Aggregated MDP (SAMDP), which abstract the trajectory of states and actions that enable a visualization of the abstracted trajectory (Zahavy et al., 2016). From the perspective of XAI, their contribution was to point out the potential that an abstracted trajectory can provide a better understanding of a RL model and support debugging task for RL practitioners. Although the target users in their study were RL practitioners, they have not conducted a user study to evaluate their method. Thus, it has not been investigated how or to what extent the abstracted trajectory visualization method provides a understandable explanation for a wide range of users. In addition, since the proposed method was designed for a specific type of RL model, it has not been examined whether the visualization using other RL models can improve explainability.

A user study of the trajectory visualization was conducted by Amir *et. al* (Amir & Amir, 2018). Their user study indicated that non-RL experts who do not have the knowledge of RL are able to infer the proficiency level of agents by observing the agent's trajectory. However, since their visualization method is limited to replays, other types of visualization designs that can convey information to a wide range of users as efficiently as replays have not been explored.

3. Methodology

To examine the potential of explainability using state trajectory abstraction and its visualization for various RL models, this section introduces a model agnostic state trajectory abstraction algorithm and design exploration of its visualization.

3.1. Model-agnostic State Trajectory Abstraction Algorithm

In this section, a model-agnostic trajectory abstraction algorithm using a Variational Autoencoder (VAE) is proposed. As shown in Figure 1, the algorithm consists of three steps: feature extraction, state aggregation, and graphical model inference. In the following paragraphs, we provide an overview of these three steps in the context of Breakout, one of Atari's most popular games. Dataset used in this study were obtained from Such *et al.* (2018). They trained deep RL models on various ATARI games to support research that investigates the properties of agents. Deep RL models used in this study are A2C, ApeX, DQN, ES, GA and Rainbow.

In step 1, the states, x_t, which are game images in each time frame of Breakout, are encoded in a low-dimensional latent variable z_t. The encoder is a part of the VAE trained on the same image dataset. In this process, the original high dimensional trajectory, which is the series of Breakout's images, is projected to a lower dimensional latent space. In step 2, the encoded states in the latent space are aggregated by clustering. In step 3, a graphical model is inferred by evaluating the transitions of the clusters and expressing these transitions as a directed graph.

In the following subsections, these three steps, feature extraction, state aggregation via clustering, and graphical model inference, will be described in more detail.

3.1.1. Feature Extraction

Feature extraction can be defined as a mapping function $q : x \rightarrow z \in \mathbb{R}^m$, where x is the input state, which can be an image or vector type data, z is what is called a latent variable, and m is a dimension of the latent variable. To obtain a disentangled representation of state transitions for different types of RL models, this study employs a β-VAE as the mapping function. The β-VAE is trained by optimizing the evidence lower bound (ELBO) instead of directly maximizing the intractable marginal log-likelihood. The ELBO is defined as follows:

$$\mathcal{L}_{ELBO} = -log\left(p(x|z)\right) + \beta D_{\mathbb{KL}}\left(q(z|x)||p(z)\right) \quad (1)$$

By imposing a strong constraint with $\beta > 1$, the β-VAE can learn a compressed representation that captures the statistically significant information in the data (Chen et al., 2018). Consequently, similar images are positioned closely together in the latent space, and as the state changes continuously due to the actions of the agent, the state transitions are represented as smooth trajectories in the latent space.

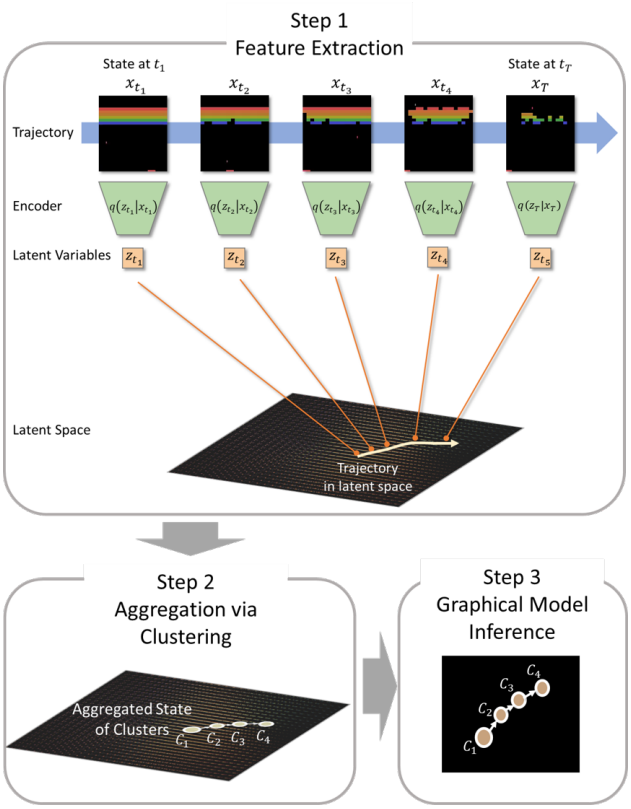


Figure 1. Overview of the proposed model-agnostic trajectory abstraction algorithm.

In this research, the architecture of the β-VAE follows the approach presented by Ha and Schmidhuber (2018). The dimensionality of the latent variable is set to 32, and the value of β is chosen as 5 for this study.

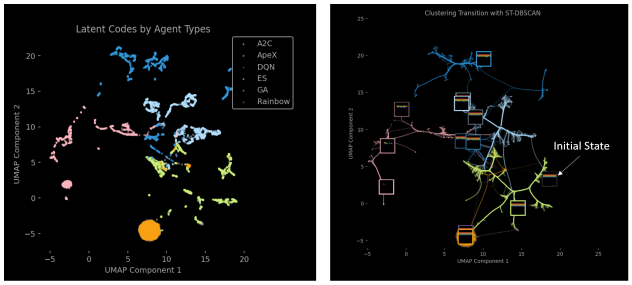
3.1.2. Aggregation via ST-DBSCAN

The encoded states are aggregated by clustering. In this research, we have chosen ST-DBSCAN as a clustering algorithm (Cakmak et al., 2021), which is a density based clustering technique designed for spatio-temporal data. Unlike many clustering methods that assume the independence and identically distributed (i.i.d.) assumption, ST-DBSCAN considers the temporal aspects. It forms a cluster when the number of points in a spatio-temporal region of specified radius around a point exceeds a specified threshold. This enables the clustered state transitions to capture the temporal structure inherent in the original data. Another important advantage of this clustering technique is to have the ability in discovering clusters with arbitrary shape such as linear,

concave, and oval. Furthermore, in contrast to some clustering algorithms, it does not necessitate the pre-determination of the number of clusters.

3.1.3. Graphical Model Inference

Finally, a graphical model is extracted from the clustered states. Figure 2 (1) shows a visualization of latent variable colored by the types of RL models. The latent variables with 32 dimension are projected into 2D UMAP component. As shown in the figure, the trajectories of six different RL models show distinct patterns. Figure 2 (2) shows the result of visualization applying clustering to the latent variables with 32 dimension and highlighting trajectories of the latent variables by edge-bundling. The minimal expression of cluster transition is obtained by sorting cluster label in chronological order and removing adjacent duplicate labels. In this step, the minimal expression of cluster transition is expressed as a directed acyclic graph. The examples of the graphical model are shown in Figure 3.



(1) Latent Codes projected into 2D UMAP component space

(2) Visualization emphasizing temporal state transition using edge bundling

Figure 2. (1) Visualization of latent variables from β-VAE colored by agent types. (2) Visualization emphasizing temporal state transition with edge bundling. The edges between two latent variables representing state transition were bundled based on its density and highlighted.

3.2. Visualization Design Exploration

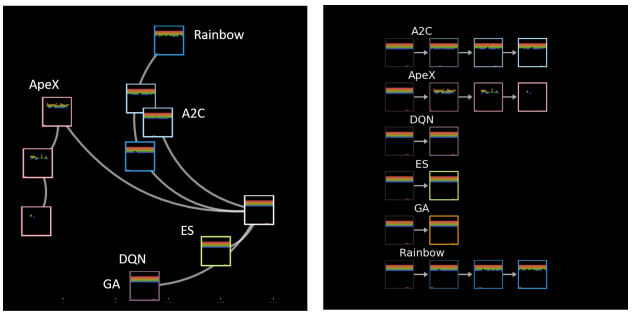
This section presents two visualization ideas for the abstracted state trajectory, which are Idea #1 and Idea #2, as shown in Figure 3.

3.2.1. Idea #1

First, Idea #1 is a method of displaying a graphical model on a two-dimensional latent space. By using β-VAE and ST-DBSCAN, the location of clustered state has some semantic meanings, since the distance between two clustered states is considered to be correlated with the proximity of the states. Therefore, since different policies have different state transitions, the user can estimate the agent's policy by comparing the differences in trajectories on the latent space. Also, users are able to zoom into the latent space to view each agent's game states and infer each of their behaviors. This idea enables users to freely explore the latent space according to their personal preference.

3.2.2. Idea #2

Idea #2 is a method in which representative images of clusters are displayed as nodes of a graphical model, and transitions between two clusters are visualized with directed edges. Compared to Idea #1, Idea #2 does not plot trajectories on a latent space, so users cannot obtain information about the distance in latent space between each cluster. However, in situations where trajectories overlap when using Idea #1, Idea #2 avoids the overlapping. Although this idea restricts the user's ability to explore the latent space, the user is allowed to understand the trajectories generated by the agent individually and linearly.



(1) Idea #1

(2) Idea #2

Figure 3. Two ideas for visualizing abstracted trajectories: (1) how to visualize an abstracted trajectory on a latent space, and (2) how to visualize a graphical model corresponding to the abstracted trajectory.

4. User Study Plan

A user study is needed to evaluate the performance of our abstracted trajectory visualization method. The following describes the questions guiding the design of tasks and evaluation metrics for the user study. We also discuss the anticipated outcomes of the user study.

4.1. Questions

This section presents specific questions that serves as design guidelines for this user study.

- Q1- Can a graphical model visualization of abstracted trajectories provide accurate agent policy awareness to users than visualization of entire trajectories?
- Q2- Can visualization of abstracted trajectories by graphical models provide efficient agent

policy recognition for users?

Q3- How confidently can users infer an agent policy from the abstracted trajectory?

Q4- Which of the proposed trajectory visualization methods do users prefer and why?

4.2. Task Design

To evaluate the abstracted trajectory visualization method, the following two types of tasks, task #1 and #2, are designed. As shown in Figure 4, these tasks #1 and #2 will be performed for visualization ideas #1 and #2, respectively.

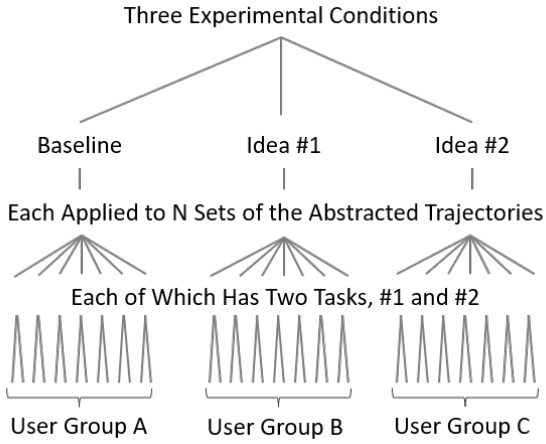


Figure 4. Overview of the user study trials

4.2.1. Task #1

The aim of task #1 is to measure how correctly the user is able to generalize an agent's policy from a proposed visualization idea. Participants will be randomly assigned to the baseline case, idea #1, or idea #2. In each case, a user is instructed to look at three trajectories generated by three different agents named Player #1, #2, and #3. Then, the user is asked to look at another trajectory and infer which of the three agents generates this trajectory.

Figure 5 shows the proposed method and the baseline case. The baseline on the left side of Figure 5 uses the method of showing the full-length trajectories of the agents of choice. This corresponds to showing a whole movie of each agent playing the game. In the proposed method, the trajectories are abstracted from the full-length trajectories and visualized as a graphical model, as shown in the right side of Figure 5. By comparing users' accuracy between the case of baseline and proposed methods, this task can provide a measure of how correctly the user is able to generalize an agent's policy from a proposed visualization idea compared to baseline.

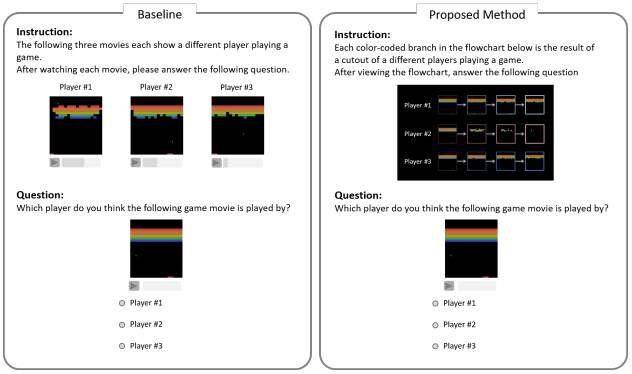


Figure 5. Comparison of baseline and proposed methods using Idea #2 in Task #1; this task corresponds to the task of decoding human abstracted trajectories into complete trajectories. The comparison of the baseline and the proposed method in the user study will show the extent to which the proposed method is successful in abstracting the complete trajectory.

4.2.2. Task #2

Task #2 is the reverse of Task #1. The aim of the task #2 is to evaluate how well a user's mental model obtained from the proposed abstract visualization ideas agrees with agents' complete trajectory. Likewise in task #1, participants will be randomly assigned to either the case of base line or proposed method. In both cases, a user is instructed to look at one trajectory generated by an anonymous agent. Then, the user is asked to look at three trajectories played by Player #1 to #3, and asked to infer which of the three agents is generating the trajectory that is displayed in the instruction session.

While the instructions in both cases for the baseline and proposed method are the same, the setting of questions of each case are different. Figure 6 shows the baseline used in Task #2 and the proposed method. In the baseline, the question section shows a whole movie of each agent playing the game, as shown in the left side of Figure 6. The proposed method on the right side of Figure 6 shows the graphical model of trajectories abstracted from the whole movie. By comparing users' accuracy between the case of baseline and proposed method, this task can provide a measure of how well a user's mental model obtained from the proposed abstract visualization ideas agrees with the user's mental model obtained from agents' complete trajectory.

4.3. Evaluation Metrics

In addition to the user's accuracy described in the task design, this user study evaluate users objectively and subjectively.

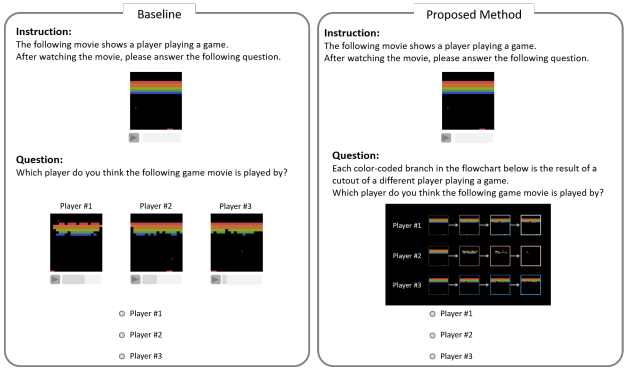


Figure 6. Comparison of the baseline and the proposed method using Idea #2 in Task #2; this task corresponds to the task of a human encoding a trajectory from a complete trajectory to an abstracted trajectory. By comparing the baseline and the proposed method in the user study, we can see how well the proposed method compresses the information.

4.3.1. Objective Evaluation

Users' accuracy for each task can be considered a good objective measure of how successfully the users construct the mental model of agents from trajectory visualizations. By analyzing the accuracy for each task, the different aspects of visualization performance can be evaluated. The accuracy in the task #1 represent how correctly the user is able to generalize an agent's policy from a trajectory visualization. In task #2, the accuracy can be a good measure of how well a user's mental model obtained from a trajectory visualization. Also, by comparing the accuracy of different trajectory visualization methods, better trajectory visualization methods from a user's perspective can be identified.

As another objective metric, we will also measure the task completion time for each user. This will provide an efficiency metric defined as the accuracy divided by the user's completion time.

4.3.2. Subjective Evaluation

As a subjective measure, users will be asked to rate their confidence in their inference with a 5-point Likert scale. This corresponds to evaluating the subjective likelihood of encoding and decoding of abstracted trajectories. For reference, user's preferences and comments for the proposed visualization ideas will also be examined.

4.4. Expected Outcomes

The following sections describe the expected outcomes from the evaluation of accuracy, expected outcomes from the evaluation of efficiency, and expected outcomes from the subjective evaluation.

4.4.1. Expected outcomes from the evaluation of accuracy

As mentioned in the aim of the task design, the inference accuracy of users in Tasks #1 and #2 evaluates the performance of the users' encoding and decoding for an agent's policy. In addition, by comparing the accuracy for different trajectory visualization ideas, the best visualization idea from the users' perspective can be identified.

Furthermore, by changing the number of agents in each task, we can see how well the abstract trajectory visualization idea succeeds in compressing the agent's whole trajectory. We expect that as the number of agents in the tasks is increases, it would become more difficult for the users to keep all of the abstracted trajectories in their mind. This may be salient in the baseline case because the baseline case requires the user to visually see much more of the agents states than our proposed method.

4.4.2. Expected outcomes from the evaluation of efficiency

Efficiency, calculated from accuracy and task completion time is another important metric to evaluate the performance of the proposed visualization ideas. Comparing the efficiency of different proposed ideas will lead to the identification of useful visualization ideas for the XAI interface for real-time monitoring.

4.4.3. Expected outcomes from the subjective evaluation

Subjective evaluation metrics will also provide important insights into XAI interfaces for non-RL experts to be designed in the future. For instance, the aspect of users to confidently estimate the policy of an agent is one of important aspects of explainability. If users can make correct inference with confidence in a short amount of time, the visualization ideas can be considered ideal. Furthermore, as a result of careful statistical analysis, if users have arrive at the correct answer quickly by just guessing, the visualization might be providing some intuitive information about the agent's policy to the users. User preferences are also important to improve the acceptability of the XAI interface.

5. Limitations

In some cases, the trajectory abstraction approach can easily lead to biased and incorrect mental model of an agent since this approach is heavily dependent on human cognitive biases. For example, if we see exaggerated TV advertisement of a product, in other words a trajectory extracted from the skewed outline of the product, the impression of the product obtained from the advertisement certainly differ from the image of actual product. Consumers without knowledge

about the product have no way of knowing before actually using it that it is significantly different. In order to gain rich and accurate mental model of a subject, one must possess deep knowledge about the subject. However, it is practically impossible that everyone has deep knowledge of all subjects. Therefore, in the context of RL, it is important to assess the boundaries and limitations of explainability provided by this and any approach.

Despite such significant limitation, this limitation also suggests the potential to attract many people, as demonstrated by the aforementioned example of TV advertisements. This has the potential to involve a wide range of people in the debate over ethical issues regarding AI, which is still being discussed in a limited knowledge community.

6. Concluding Remarks

In this paper, we introduced a model-agnostic state abstraction algorithm, explored the design of the abstracted trajectory visualization and proposed an evaluation framework for the abstracted trajectory visualization. Our approach abstracts the trajectory of states that various agents has observed using β-VAE and ST-DBSCAN, and provides a high level view, in the form of a visualization from abstracted trajectories that characterize the various patterns of agents' behavior. Also, we designed a user study that can evaluate the visualization design in terms of objective metrics and subjective evaluations. Future work includes applying our algorithm to various types of applications such as Atari games other than breakout, and conducting the user study to evaluate the performance of the visualization.

References

Amir, D. and Amir, O. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1168–1176, 2018.

Atrey, A., Clary, K., and Jensen, D. Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. *arXiv preprint arXiv:1912.05743*, 2019.

Bastani, O., Pu, Y., and Solar-Lezama, A. Verifiable reinforcement learning via policy extraction. *Advances in neural information processing systems*, 31, 2018.

Cakmak, E., Plank, M., Calovi, D. S., Jordan, A., and Keim, D. Spatio-temporal clustering benchmark for collective animal behavior. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Animal Movement Ecology and Human Mobility*, HANIMOB '21, pp. 5–8, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450391221. doi: 10.1145/3486637.3489487. URL https://doi.org/10.1145/3486637.3489487.

Chen, R. T., Li, X., Grosse, R., and Duvenaud, D. Isolating sources of disentanglement in vaes. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, volume 2615, pp. 2625, 2018.

Coppens, Y., Efthymiadis, K., Lenaerts, T., and Nowé, A. Distilling deep reinforcement learning policies in soft decision trees. In *International Joint Conference on Artificial Intelligence*, 2019.

Coudert, F., Butin, D., and Le Métayer, D. Body-worn cameras for police accountability: Opportunities and risks. *Computer law & security review*, 31(6):749–762, 2015.

Elizalde, F., Sucar, L. E., Reyes, A., and Debuen, P. An mdp approach for explanation generation. In *ExaCt*, pp. 28–33, 2007.

Ha, D. and Schmidhuber, J. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

Hayes, B. and Shah, J. A. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pp. 303–312, 2017.

Jaunet, T., Vuillemot, R., and Wolf, C. Drlviz: Understanding decisions and memory in deep reinforcement learning. In *Computer Graphics Forum*, volume 39-3, pp. 49–61. Wiley Online Library, 2020.

Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., and Doshi-Velez, F. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on explainable artificial intelligence*, 2019.

Khan, O., Poupart, P., and Black, J. Minimal sufficient explanations for factored markov decision processes. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 19, pp. 194–200, 2009.

Liu, G., Schulte, O., Zhu, W., and Li, Q. Toward interpretable deep reinforcement learning with linear model u-trees. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part II 18*, pp. 414–429. Springer, 2019.

Madumal, P., Miller, T., Sonenberg, L., and Vetere, F. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34-3, pp. 2493–2500, 2020.

Ming, Y., Cao, S., Zhang, R., Li, Z., Chen, Y., Song, Y., and Qu, H. Understanding hidden memories of recurrent neural networks. In *2017 IEEE conference on visual analytics science and technology (VAST)*, pp. 13–24. IEEE, 2017.

Olson, M. L., Khanna, R., Neal, L., Li, F., and Wong, W.-K. Counterfactual state explanations for reinforcement learning agents via generative deep learning. *Artificial Intelligence*, 295:103455, 2021.

Puri, N., Verma, S., Gupta, P., Kayastha, D., Deshmukh, S., Krishnamurthy, B., and Singh, S. Explain your move: Understanding agent actions using specific and relevant feature attribution. *arXiv preprint arXiv:1912.12191*, 2019.

Rupprecht, C., Ibrahim, C., and Pal, C. J. Finding and visualizing weaknesses of deep reinforcement learning agents. *arXiv preprint arXiv:1904.01318*, 2019.

Singh, S., Jaakkola, T., and Jordan, M. Reinforcement learning with soft state aggregation. *Advances in neural information processing systems*, 7, 1994.

Strobelt, H., Gehrmann, S., Pfister, H., and Rush, A. M. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics*, 24(1):667–676, 2017.

Struckmeier, O., Racca, M., and Kyrki, V. Autonomous generation of robust and focused explanations for robot policies. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1–8. IEEE, 2019.

Such, F. P., Madhavan, V., Liu, R., Wang, R., Castro, P. S., Li, Y., Zhi, J., Schubert, L., Bellemare, M. G., Clune, J., et al. An atari model zoo for analyzing, visualizing, and comparing deep reinforcement learning agents. *arXiv preprint arXiv:1812.07069*, 2018.

Sundararajan, M. and Najmi, A. The many shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR, 2020.

Tabrez, A., Agrawal, S., and Hayes, B. Explanation-based reward coaching to improve human performance via reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 249–257. IEEE, 2019.

van der Waa, J., van Diggelen, J., Bosch, K. v. d., and Neerincx, M. Contrastive explanations for reinforcement learning in terms of expected consequences. *arXiv preprint arXiv:1807.08706*, 2018.

Zahavy, T., Ben-Zrihem, N., and Mannor, S. Graying the black box: Understanding dqns. In *International conference on machine learning*, pp. 1899–1908. PMLR, 2016.