Supplementary Materials for "ASAP: Attention-Based State Space Abstraction for Policy Summarization"

Yanzhe Bekkemoen YANZHE.BEKKEMOEN@NTNU.NO
Helge Langseth HELGE.LANGSETH@NTNU.NO
Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway

Editors: Berrin Yanıkoğlu and Wray Buntine

Appendix A. Considerations

This section discusses aspects of ASAP, including scalability, choosing the number of hyperstates, human evaluation, and how state semantics can be evaluated.

A.1. Scalability

The method assumes states are low-dimensional with interpretable features, in line with state-of-the-art methods. To scale the method to environments with high-dimensional symbolic states, one can show the most important features to minimize the complexity of explanations.

Several things need to be considered to apply this method to more complex environments. One is how to represent states while still being human-understandable. For example, we cannot directly work in pixel space like with Atari games since ASAP does not provide a way to represent the hyperstates interpretably when features are uninterpretable. Thus, we must first design alternative representations for environments with high-dimensional states to scale ASAP. Another consideration is the number of hyperstates so explanations are faithful and capture agent behavior without overwhelming the end-user. This is a difficult trade-off, and a solution would be to create explanations for subspaces by first segmenting the state spacing before applying ASAP.

A.2. Hyperstate Numbers

The number of hyperstates is determined by the environment's complexity and the user's needs. The more complex environments need more hyperstates to capture the agent's behavior. When it comes to the user's needs, we need to consider several aspects. Whether the user wants hyperstates to represent a single action each, state similarity, and/or feature importance similarity within hyperstates. In the experiments, we minimized the number of hyperstates and simultaneously captured the agent's behavior without directly clustering on actions.

A.3. State Semantics

To understand what a hyperstate represents, the set of states representing the hyperstate must be similar. Accordingly, this makes it easier for humans to interpret the hyperstates as they represent a single situation or a few similar situations. This has previously been achieved by directly clustering on state features.

We use t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) and the silhouette score (Rousseeuw, 1987) in the experiments to validate if our method captures this property. t-SNE visually shows us how similar states are and trends in a low-dimensional space. In this way, we can visually verify if the hyperstate assignment produced by ASAP aligns with human intuition. We use default hyperparameters for t-SNE specified by Scikit-learn v1.2.2 (Pedregosa et al., 2011) to avoid overfitting and misreading patterns in the visualizations. Silhouette score is often used as a metric to evaluate clustering. As hyperstates are clusters of states, we can use this score to assess the hyperstate assignment. Although the score gives a higher value to a spherical hyperstate assignment, it enables a quantitative way to evaluate hyperstates without subjective interpretations, like when using t-SNE to evaluate.

A.4. Human Evaluation

Human evaluation provides important new insights and is important to understand the method's limitations. However, there are some shortcomings. First, the evaluations are not uniform across papers, making replicating and comparing results difficult. Second, due to cost, researchers often use Amazon Mechanical Turk and university students for evaluation. They are not necessarily the intended end-users for the explanations, making it hard to assess the real-world value and the impact of explanations. Third, it is challenging to design user studies, and often, best practices from the human-computer interaction literature are not used in explainable artificial intelligence papers (Abdul et al., 2018).

Future work can design and execute user studies to measure the effectiveness of ASAP using best practices from the human-computer interaction literature. In addition, comparisons to other explainable reinforcement learning (XRL) methods can be made to see how they complement each other. Moreover, investigate when it is suitable to use explanations produced by ASAP versus other XRL methods.

References

Ashraf M. Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan S. Kankanhalli. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proc. of CHI*, 2018. doi: 10.1145/3173574.3174156.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 2011. doi: 10.5555/1953048.2078195.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 1987. doi: 10.1016/0377-0427(87)90125-7.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *JMLR*, 2008.