Goal-Conditioned Recommendations of AI Explanations

Saptarashmi Bandyopadhyay
Department of Computer Science
University of Maryland, College Park
Maryland, MD 20742
saptab1@umd.edu

Vibhu Agrawal
Department of Computer Science
University of Maryland, College Park
Maryland, MD 20742
vagrawa1@terpmail.umd.edu

Sarah Savidge
Department of Defense
joseph.sarahk.work@gmail.com

Eric Krokos
Department of Defense
EricPKrokos@gmail.com

John Dickerson
Department of Computer Science
University of Maryland, College Park
Maryland, MD 20742
johnd@umd.edu

Abstract

The large-scale usage of Artificial Intelligence (AI) models has made it important to explain their outputs subject to requirements and goals for using these models. The definition of goals in Goal-conditioned Reinforcement Learning (GCRL) aligns with the task of recommending an appropriate explanation among Explainable AI (XAI) models like SHAP or LIME that is most interpretive for specific AI models. We focus on two goals of training random forest classifier to classify different training data in order to find appropriate explanations. SlateQ recommendation system is used for simulation where the underlying RecSim environment has a slate of documents with different quantity scores representing different goals.

1 Introduction

In the realm of artificial intelligence, the pursuit of Explainable AI (XAI) models has never been more crucial for transparency. Understanding the inner workings of AI models has become imperative to ensure their reliability, fairness, and accountability. XAI tools have emerged to provide insights into AI model behavior and decisions. There exist a variety of XAI models that aim to provide insight into the behavior of complex AI models, which presents a challenge: how can we effectively recommend the most appropriate explanation that elucidates the behavior of AI models as per user goals?

We explore this challenge by connecting Goal-Conditioned Reinforcement Learning (GCRL) to a recommender system that can select relevant explanations. The recommender system's objective is to learn a policy that selects an XAI model's outputs most aligned to user goals. Goals are the specific task done by the AI model.

It can be challenging to represent goals in recommender systems [1, 2]. There can be unified goals while selecting a set of items such as selecting different XAI model outputs for a meta-task like agriculture planning using precipitation predictions from an AI model. Likewise, there can be individual goals in selecting different XAI model outputs. Different tasks like disaster recovery or irrigation planning need different characteristics of XAI model outputs based on an interpretation

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

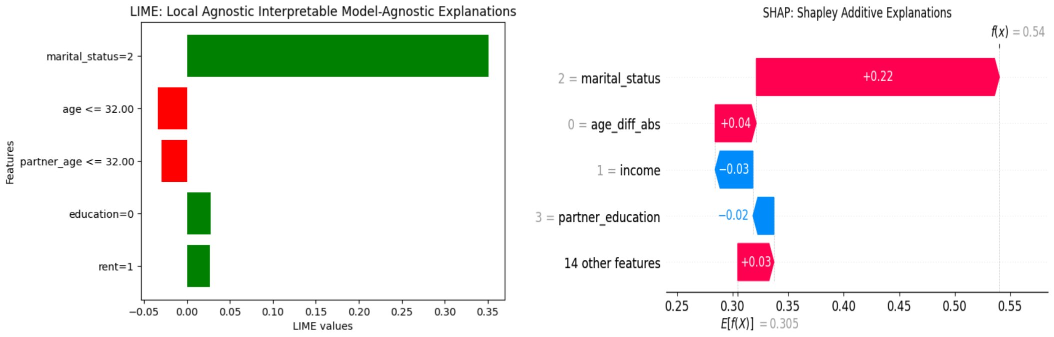

Figure 1: (a) LIME explanation and (b) SHAP explanation for a test sample in the couples dataset

of the individual preferences of the users in different scenarios. This makes the GCRL based recommender system pertinent to recommend the appropriate XAI model output for different tasks.

The SlateQ [3] recommender system is used to simulate this in the RecSim environment [4], recommending from a slate of different documents, which are selected by their relevance to the user. This motivates us to recommend slates of XAI explanations elucidating AI model outputs. We have user preferences and generated labels from LLMs preferring explanations generated from XAI models.

Shapley Additive Explanations (SHAP), calculates a score for each feature in the AI model, which represents its weight to the model output. It does this by approximating the effect of removing a feature from the model and then doing this for all subsets of features in the model. [5]. LIME (Locally Interpretable Model-Agnostic Explanations) calculates scores for each feature in the model by measuring model behavior in response to applying small perturbations to an input, thereby generating local explanations for test samples [6]. It is important to decide when to use SHAP or LIME depending on the goal of which machine learning model is to be explained [7].

2 Offline Explanations Dataset

We generate explanations using the XAI models LIME and SHAP on the corresponding test samples for random forest classifiers trained on the Diabetes dataset [8] and the HCMST dataset [9] as different goals. Figure 1 represents the LIME and SHAP XAI plots for a test sample explaining the goal of random forest classification on couples dataset.

Goal	# Train	# Test
Explaining Random Forest Classifier trained on Diabetes dataset [8]	537	231
Explaining Random Forest Classifier trained on Couples dataset [9]	1030	442

Table 1: Statistics of Training & Testing Datasets for two AI model goals

3 Goal-Conditioned Recommendations

State representations in SlateQ recommender system (RecSys) are static user features and historical user preference. These user features are attributes of the LIME and SHAP XAI models like latency, training data sparsity, consistency and other semantic features, which vary across different test samples. The current state is updated when an item like an XAI model output is selected by the user from the slate of explanations, updating the known user preferences for the aforementioned features. The user selects item i from the slate A with unnormalized probability $v(s, i)$ where v is some function [3] which leads to a transition to the next state in the trajectory of the user agent. The components for a goal-conditioned Markov Decision Process (MDP) in the RecSys setup for explanations are summarized in Table 2 with simulation results in Figure 2. Let any XAI model output be θ_{x_i} for the AI model output x_i, or goal. Let the user feature values be represented as $f_u^1, f_u^2, ... \in F_u$ where F_u is the set of all feature values for the user. Let the XAI output features be

	Descriptions
States	User preference for explanations
Actions	User choice of XAI model explanation
Transitions	Deterministic updates to the known user preferences for explanations
Reward	Assigned when the model selects an explanation that the user also prefers
Goals	Recommending XAI model outputs appropriate for specific AI models

Table 2: Summary of Goal-conditioned Reinforcement Learning RecSys for Explanations

represented as $f^1_{\theta_{x^i}}, f^2_{\theta_{x^i}}, ... \in F_{\theta_{x^i}}$. Then $\theta^*_{x_i}$ is defined as the XAI model output minimizing the distance between F_u and $F_{\theta_{x^i}}$ among a slate of other XAI model outputs for x_i. The reward can be defined as the influence of user engagement in selecting recommended explanations from a particular slate that is aligned with their goals.

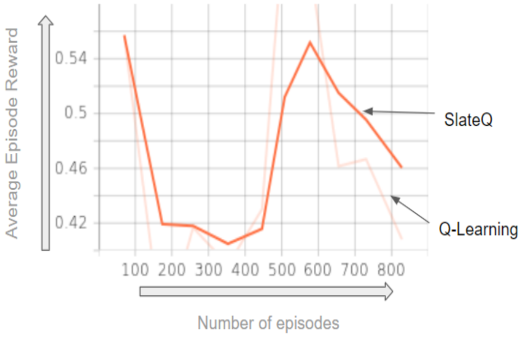

Figure 2: Average Episode Reward across time-steps for two goals while simulating with SlateQ RecSys in comparison with a Q-learning approach

Future work involves AI explanation goal representations aligned with requirements of user agents in multiple scenarios which can be validated with human feedback along with the development of benchmarks to evaluate the quality of goals.

References

[1] Shoujin Wang, Longbing Cao, Yan Wang, Quan Z. Sheng, Mehmet A. Orgun, and Defu Lian. A survey on session-based recommender systems. *ACM Comput. Surv.*, 54(7), jul 2021.

[2] Xiaocong Chen, Lina Yao, Julian McAuley, Guanglin Zhou, and Xianzhi Wang. A survey of deep reinforcement learning in recommender systems: A systematic review and future directions. *arXiv preprint arXiv:2109.03540*, 2021.

[3] Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Tushar Chandra, and Craig Boutilier. Slateq: A tractable decomposition for reinforcement learning with recommendation sets. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2592–2599. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[4] Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. Recsim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847*, 2019.

[5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[7] Ahmed Salih, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E Petersen, Gloria Menegaz, and Karim Lekadir. Commentary on explainable artificial intelligence methods: Shap and lime. *arXiv preprint arXiv:2305.02012*, 2023.

[8] Mehmet Akturk. Diabetes dataset. `https://www.kaggle.com/datasets/mathchi/diabetes-data-set`, 2020. [Accessed 04-10-2023].

[9] Reuben J. Thomas Michael J. Rosenfeld and Maja Falcon. How Couples Meet and Stay Together (HCMST). `https://data.stanford.edu/hcmst`, 2018. [Accessed 04-10-2023].