Explainable AI for RL in federated learning

Ruizhi Huang, Dr. Beiyu Lin

Motivation

Nowadays, AI is everywhere, and it is increasing used in a variety of real-world applications like helping doctors and judges to make decisions. However, most of AI models are black-box, and sometimes it is difficult to extract and read the model decisions processes by human, especially in some complicated neutral models. Therefore, model understanding becomes important in some domains particularly those involving high stakes decisions, which can impact millions of individuals. Recently, some studies have already explored some technologies to improve the understandings and proposed systems and models to deploy the applications of explainable AI. The motivation of this research is to apply explainable AI into popular reinforcement learning and federated learning.

Objective

To solve federated learning unbalanced dataset problems and to extract global and local explanations via federated learning.

Global explanations:

- Explain complete behavior of the model.
- Help detect big picture model biases affecting larger subgroups.
- Help vet if the model, at a high level is suitable for deployment.

Local explanations:

- Explain individual predictions.
- Help unearth biases in the local neighborhood of a given instance
- Help detect if individual predictions are being made for the right reasons.

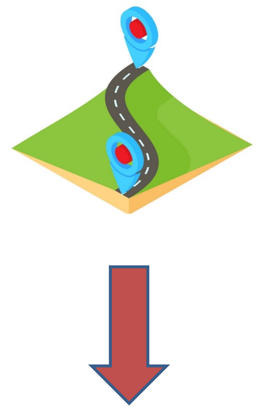
Proposed Methods

1.To generate samples from underrepresented groups by using Generative adversarial imitation learning. In the step, we use OpenAI to implement the algorithm.

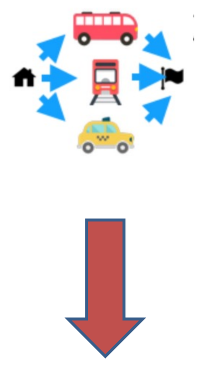
1.Generate expert trajectory data
2.Sample the expert trajectory data from the generated trajectories
3.Execute imitation learning –GAIL, run behavioral cloning
4.Test trained policy for GAIL

2. To extract both global and local features for XAI via Federated learning

Local explanations	Global explanations
Feature importance: LIME, SHAP value	Collection of Local Explanations
Rule based	Representation Based: Network Dissection, TCAV
Saliency Maps	Model Distillation
Prototypes/example based	Summaries of Counterfactuals



Trip distance, weather, speed



Rating bus. Rating weather, linha

Dataset

Variables:

- Speed – it represents the average speed (Km/H)
- Distance – it represent the total distance (Km)
- Rating – it represent the total distance (Km). Responds were divided in to three categories (3- good, 2- normal, 1 – bad).
- Rating_bus – it is other evaluation parameter for bus crowded. (1 – the amount of people inside the bus is little, 2 – the bus is not crowded, 3 – the bus is crowded)
- Rating_weather – there are two categories. (1 – raining, 2 – sunny).
- Car_or_bus – (1-car, 2-bus)
- Linha – information about the bus that does the pathway
- latitude latitude from where the point is
- Longitude: longitude from where the points is.
- Track_id: identify the trajectory which the point belong.
- Id_android – it represents the device used to capture the instance.
- Time: datetime when the point was collected (GMT-3)

Features		N=164
Car or bus	87 (car)	72 (bus)
Distance (average)	6.495 Km (car)	3.937 Km (bus)
Weather	21.28% (raining)	78.72% (sunny)
Speed (average)	20.553 Km/H (car)	12.299 Km/H (bus)

Acknowledgement

This work was funded by an unrestricted gift from Google. I would like to thank UTRGV for hosting this program and Beiyu Lin for her guidance in this work