Explaining a Reinforcement Learning Agent via Prototyping

Ronilo Ragodos, Tong Wang, Qihang Lin, Xun Zhou

University of Iowa

NeurIPS 2022

Goal & Challenges for Prototype-Based Explanations

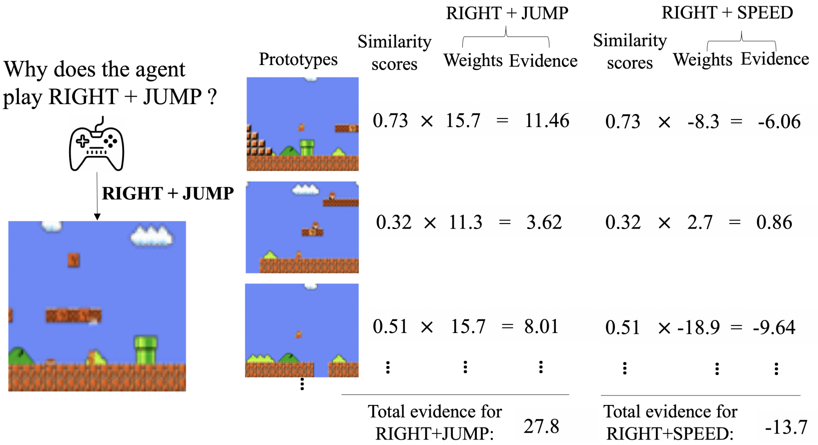
Goal: A policy explainer that does not need access to expert agent internals, only demonstrations.

Idea: *Find set of prototypical situations and relate state-action pairs to those prototypes*.

Challenge 1: How to find the set of prototypical states that represent an agent's prototypical behaviors.

Challenge 2: How to define similarity between states and prototypes.

Explaining Why Agent Jumps Rightward Using Prototypes



Why does the agent play RIGHT + JUMP ?

RIGHT + JUMP

Prototypes	RIGHT + JUMP Similarity scores		RIGHT + JUMP Weights		RIGHT + JUMP Evidence	RIGHT + SPEED Similarity scores		RIGHT + SPEED Weights		RIGHT + SPEED Evidence
	0.73	×	15.7	=	11.46	0.73	×	-8.3	=	-6.06
	0.32	×	11.3	=	3.62	0.32	×	2.7	=	0.86
	0.51	×	15.7	=	8.01	0.51	×	-18.9	=	-9.64
⋮	⋮		⋮		⋮	⋮		⋮		⋮

Total evidence for RIGHT+JUMP: 27.8

Total evidence for RIGHT+SPEED: -13.7

Figure 1: ProtoX's action choice depends on weighted sums of similarity scores between the input and each prototype.

ProtoX = Encoder + Prototypes + Linear Classifier

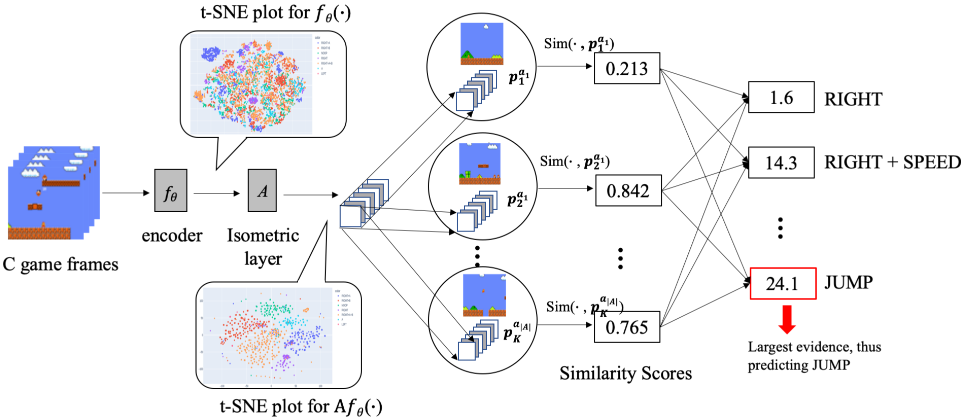
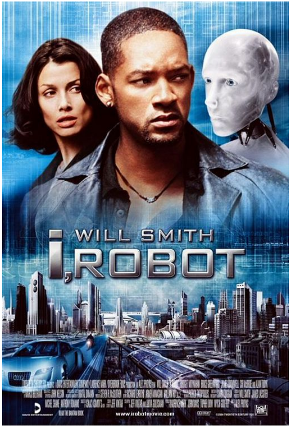
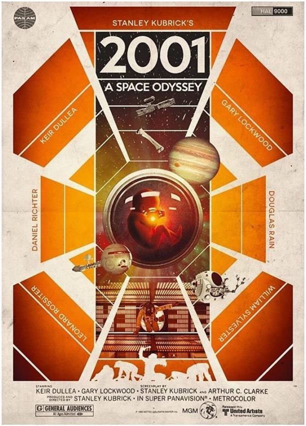


Figure 2: ProtoX Model Architecture

Potential Future Applications

- Explain a self-driving car
- Explain a (malfunctioning) robot




HAL 9000
PAN AM
STANLEY KUBRICK'S
2001
A SPACE ODYSSEY
KEIR DULLEA
GARY LOCKWOOD
DANIEL RICHTER
DOUGLAS RAIN
LEONARD ROSSITER
WILLIAM SYLVESTER
STARRING KEIR DULLEA · GARY LOCKWOOD · SCREENPLAY BY STANLEY KUBRICK AND ARTHUR C. CLARKE
PRODUCED AND DIRECTED BY STANLEY KUBRICK · IN SUPER PANAVISION® · METROCOLOR
G GENERAL AUDIENCES All Ages Admitted
MGM
Released thru United Artists A Transamerica Company