Bridging the Gap: Providing Post-Hoc Symbolic Explanations for Sequential Decision-Making Problems with Inscrutable Representations

Sarath Sreedharan, Utkarsh Soni, Mudit Verma, Siddharth Srivastava, Subbarao Kambhampati
Arizona State University

Explanation and Vocabulary Mismatch

Explanations for automated decisions needs to be framed in user understandable terms

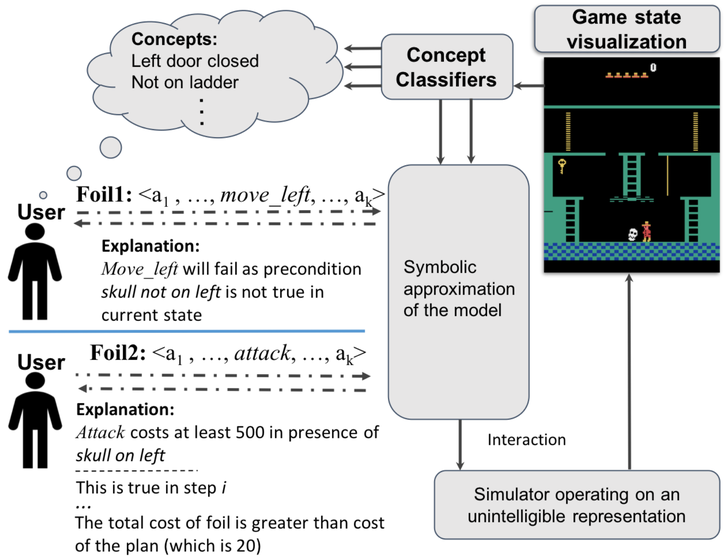
Challenging when the system is reasoning over high-dimensional states

Thus explanatory systems would need to overcome this **vocabulary mismatch**

Existing works mostly focus on handling single-shot decision making

Lime (Ribeiro'16)
TCAV(Kim'17)

Explanation in sequential decision-making settings still needs to be explored



Explaining Sequential Decisions

Decisions can no longer be evaluated in isolation

The systems now need to explain plans or policies

Help user understand why the proposed plan may be better than alternatives/foils they expected

May involve providing information like:

Why certain action is infeasible in certain states?
Why certain plans are costlier than others?

This could effectively mean providing information about underlying **model dynamics**

Symbolic Local Approximation of Models

User queried for a set of task relevant **concepts**

User provides positive and negative example for each concept

User specified concepts used to train a classifier over task states

A symbolic model can be constructed in terms of these concepts through interaction with a simulator

- Each action captured in terms of preconditions and effects
- Similarly, action costs are captured in terms of concepts

Focus on states relevant to current problem/explanatory query

Explanations can be presented in terms of this symbolic model

Concise Model Information and Explanation Confidence

Learning the entire model may be unnecessary

Consider ***contrastive explanation*** cases where user presents an alternative plan

Need to explain, either
a) Why the alternative will not succeed?
b) Why the alternative may be more expensive?

Explaining a) requires identification of a missing precondition
Explaining b) requires identification of an abstract action costs

Can be done in isolation

Confidence of explanation = Confidence over the estimated model component

User Study

Hypothesis 1: Missing precondition information is a useful explanation for action failures.

Hypothesis 2: Abstract cost functions are a useful explanation for foil suboptimality.

Hypothesis 1 tested on Montezuma's revenge over author specified concepts. Study involved four unique examples and 20 participants
19/20 participants chose our explanations

Hypothesis 2 tested on variation of Sokoban over concepts collected from users. Study involved two settings and 20 participants
14/20 participants chose our explanations

Acknowledgement: Kambhampati's research is supported in part by ONR grants N00014-16-1-2892, N00014-18-1-2442, N00014-18-1-2840, N00014-9-1-2119,AFOSR grant FA9550-18-1-0067, DARPA SAIL-ON grant W911NF-19-2-0006,NSF grants 1936997 (C-ACCEL), 1844325, NASA grant NNX17AD06G, and a JPMorgan AI Faculty Research grant.