ESPRIT: Salient Event Classification

Abhijit Gupta,[1] Dragomir Radev [1]

[1]Department of Computer Science, Yale University, New Haven, CT

LILY Lab

Introduction

PHYRE.ai is an open-source platform created by Facebook AI to facilitate and encourage the development of physical reasoning algorithms In intelligent agents. The platform provides a set of physics puzzles in a simulated 2D world. Each puzzle has a goal state and initial state. The goal state can be reached by placing new objects in the environment and running a simulation (Figure 1). There are 100 template puzzles, each with 100 slightly-varying tasks.

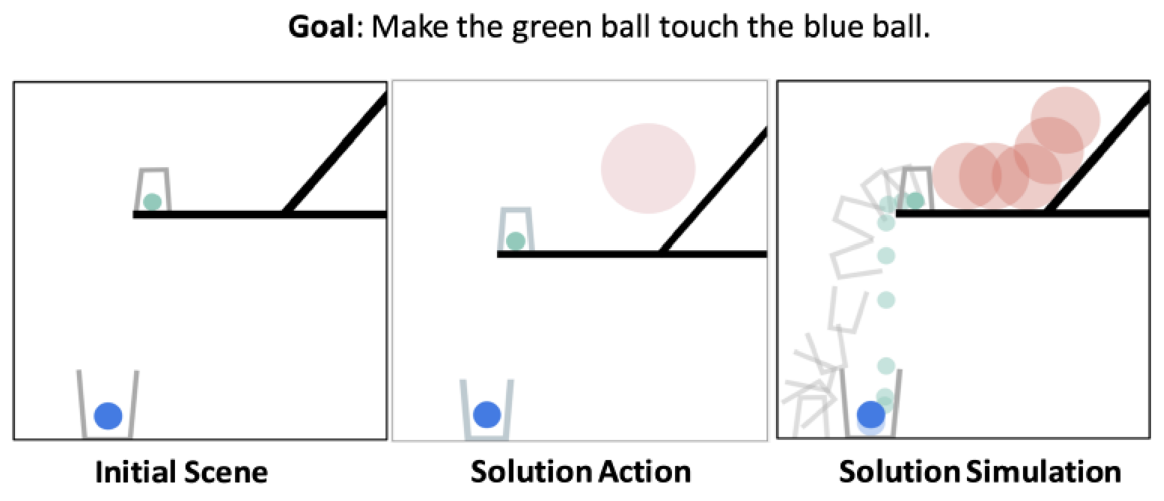
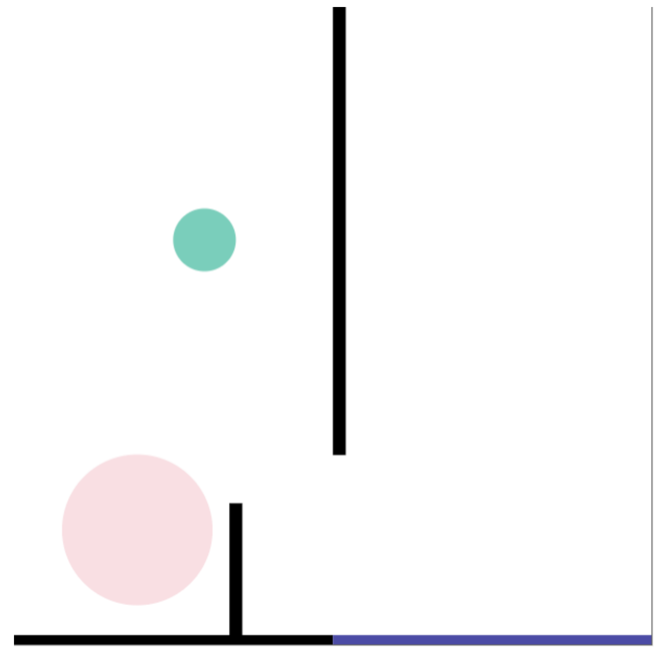


Figure 1. A demonstrative PHYRE task with Initial Scene, Solution Action, and Solution Simulation leading to the Goal State.

ESPRIT – Explaining Solutions to Physical Reasoning Tasks

We propose ESPRIT, a framework for commonsense reasoning about qualitative physics in natural language that generates interpretable descriptions of physical events. Using the PHYRE dataset, ESPRIT analyzes a PHYRE simulation to produce concise natural language descriptions on the initial scene summary and action summary. This information can be used to explain the intelligent agent's actions, or provide a hot start to additional reinforcement learning models. Figure 2 shows an example output of ESPRIT.



The red ball falls and knocks the green ball off of its curved black platform and to the left. It rolls leftwards and continues falling until it lands on the purple floor...

Figure 2: Natural Language Description of a PHYRE simulation

Salient Event Classification - Overview

ESPRIT consisted of four major steps, transforming the input simulation frames into the output natural language description. First, all the simulation frames were read and analyzed. Next, we catalogued all collisions between objects in the simulation. We then classified salient events, and sent this output into a data-to-text model. This outputted the natural language descriptions we needed. I focused primarily on the salient event classification step.

In a single simulation, there are many collisions yet only a handful are relevant to the final outcome of the simulation. Using machine learning algorithms, I sought to classify salient collisions to reduce the size of the input into the data-to-text model. We started by collecting an annotated dataset, shown in Table 1.

Templates	25
Tasks	2441
Objects / Task	13.6
Frames / Task	657.9
Collisions / Task	54.2
Annotated Tasks (train/dev/test)	625/84/76
Collisions / Annotated Task	24.5
Important Collisions / Annotated Task	3.9
Tokens / Initial State Description	38
Tokens / Simulation Description	44
Vocabulary Size	867

Table 1: Statistics for the ESPIRIT Dataset

Salient Event Classification - Methods

We collected over 8,000 annotated collisions using Amazon Mturk, separated into a train and test set. For each collision, we extracted from the PHYRE simulator thirteen input features focused on the time step and object characteristics at the time of collision. We used this data to train three machine learning models.

The baseline was a positive classifier that outputted all events as salient. I also used Python and Sklearn to create and train a Decision Tree and Support Vector Classifier. The classifiers were evaluated on three metrics: precision, recall, and F1-score. Results are shown in Table 2 below.

	Precision	Recall	F1
Positive	0.17	1.00	0.29
Decision Tree	0.99	0.96	0.97
Support Vector	0.99	0.97	0.98

Table 2: Results for Salient Event Classification

Results

As demonstrated in Table 2, the Decision Tree and Support Vector Classifier performed significantly better than the baseline, with a 99% accuracy on the testing set. Both models had high precision and recall, indicating they were successful at limiting false positives and false negatives. Using these models, the input into the data-2-text model was reduced greatly, increasing its effectiveness in creating concise descriptions of the PHYRE simulation.

The decision tree was also able to report feature importance for the thirteen inputs used to classify each collision. The time step of the collision had 18% relative importance,. Additionally, the more dynamic object in the two-body collision was weighed 350% more than the more static object when classifying salient events. These results can help guide future feature engineering.

Conclusion

Both the decision tree and support vector classifier were successful in classifying events as salient or not, useful in cleaning the input for the data-2-text models. Future improvements include better incorporating the salient event classification into the overall model, and testing new models and/or tuning hyperparameters for better classification results.

Acknowledgement

I would like to acknowledge the entire team working on the ESPRIT model – Dr. Dragomir Radev, Rui Zhang, Jeremy Weiss, Adit Vyas, Yi Chern Tan, and Nazneen Rajani.