Neuro-Symbolic Generation of Explanations for Robot Policies with Weighted Signal Temporal Logic

Mikihisa Yuasa[1], Ramavarapu S. Sreenivas[2], and Huy T. Tran[1]

[1]Department of Aerospace Engineering / [2]Department of Industrial & Enterprise Systems Engineering, The Grainger College of Engineering, University of Illinois Urbana-Champaign

LIRA@illinois
Lab for Intellignet Robots and Agents

INTRODUCTION

Black-box nature of neural networks: While learning-based methods have advanced robot decision-making and control, their lack of interpretability raises concerns for **safety-critical applications** like autonomous vehicles.

Need for explainability: Formal methods, such as **Weighted Signal Temporal Logic (wSTL)**, offer a structured way to interpret robot policies by prioritizing constraints based on importance.

Limitations of existing approaches: Current methods mainly classify trajectories rather than explain the underlying **policy behavior**, often producing **overly complex** and **hard-to-interpret** explanations.

Contribution

- Develop a **neuro-symbolic method** to generate **concise, interpretable** wSTL explanations for robotic policies.
- Introduce a **simplification process** (predicate filtering, regularization, pruning) to improve clarity without sacrificing accuracy.
- Propose **new evaluation metrics—conciseness, consistency, and strictness**—to better assess explanation quality.
- Demonstrate the effectiveness of our approach in **three robotics environments** with diverse challenges.

Experimental Setup

The experiments were designed to evaluate the effectiveness of our neural network simplification method in generating interpretable and policy-aligned explanations. We compared our method against three **baseline** approaches: **Greedy pruning** and two **top-*k*** methods (top-3 and top-5).

We tested all approaches across **seven scenarios** in **three distinct environments.**

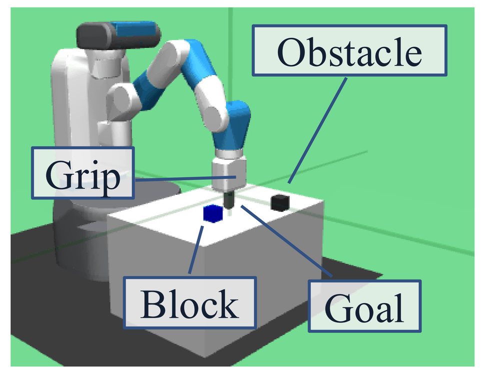

Fig 3.Obstructed Fetch Push

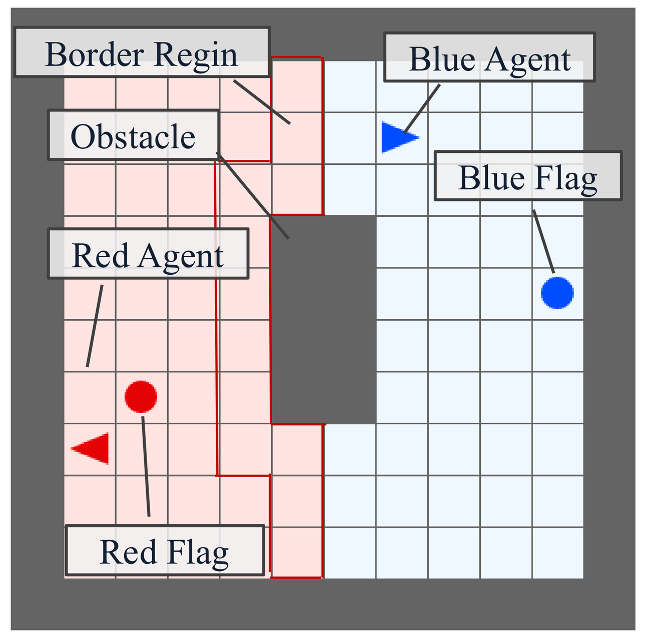

Fig 2. Capture-the-Flag

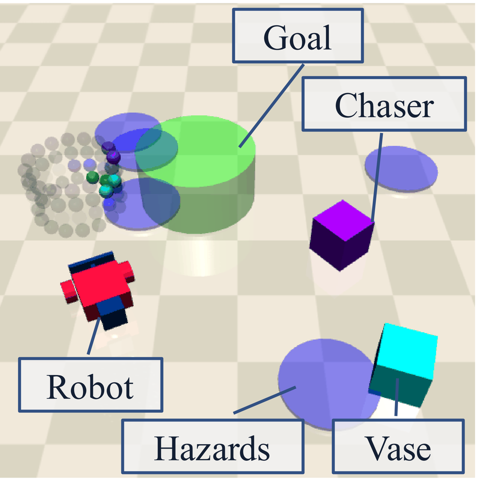

Fig 4. Chased Robot Navigation

METHOD

Predicate Filter:

- Removes predicates with similar trajectory distributions in positive and negative trajectories
- Uses a trajectory distribution vector (ratio of all-positive, mixed, all-negative robustness values).
- Applies cosine similarity as the metric and removes predicates above a user-provided threshold.

Regularization:

- Introduces two complementary regularizers to improve neural network optimization:
- **Temporal Clause Regularizer**: Enforces different conjunctive structures between eventual and global clauses.
- **Disjunctive Clause Regularizer**: Forces different structures between disjunctive clauses within both temporal clauses.
- Both regularizers are added to the loss function with adjustable weights (λ).

Weight Pruning:

- Two-step process to simplify the network:
- First prunes weights with zero values (ensuring they remain zero).
- Then removes the smallest *N* weights specified by the user.
- Eliminates least contributing weights from the optimization process.

Neural Network Architecture:

- Designed to match with the following explanation format:

$$0.5\mathcal{F}_I^{w^U}\left(\bigwedge_{i=1}^{n_{\mathrm{AP}}}\bigvee_{j=1}^{n_{\mathrm{AP}}} w_i^{\mathcal{F}} w_{ij}^{\mathcal{F}}\psi_{ij} \vee \bigvee_{j=n_{\mathrm{AP}}+1}^{2n_{\mathrm{AP}}} w_i^{\mathcal{F}} w_{ij}^{\mathcal{F}}\neg\psi_{ij}\right)$$
$$\wedge\, 0.5\mathcal{G}_I^{w^U}\left(\bigwedge_{i=1}^{n_{\mathrm{AP}}}\bigvee_{j=1}^{n_{\mathrm{AP}}} w_i^{\mathcal{F}} w_{ij}^{\mathcal{F}}\psi_{ij} \vee \bigvee_{j=n_{\mathrm{AP}}+1}^{2n_{\mathrm{AP}}} w_i^{\mathcal{F}} w_{ij}^{\mathcal{F}}\neg\psi_{ij}\right).$$

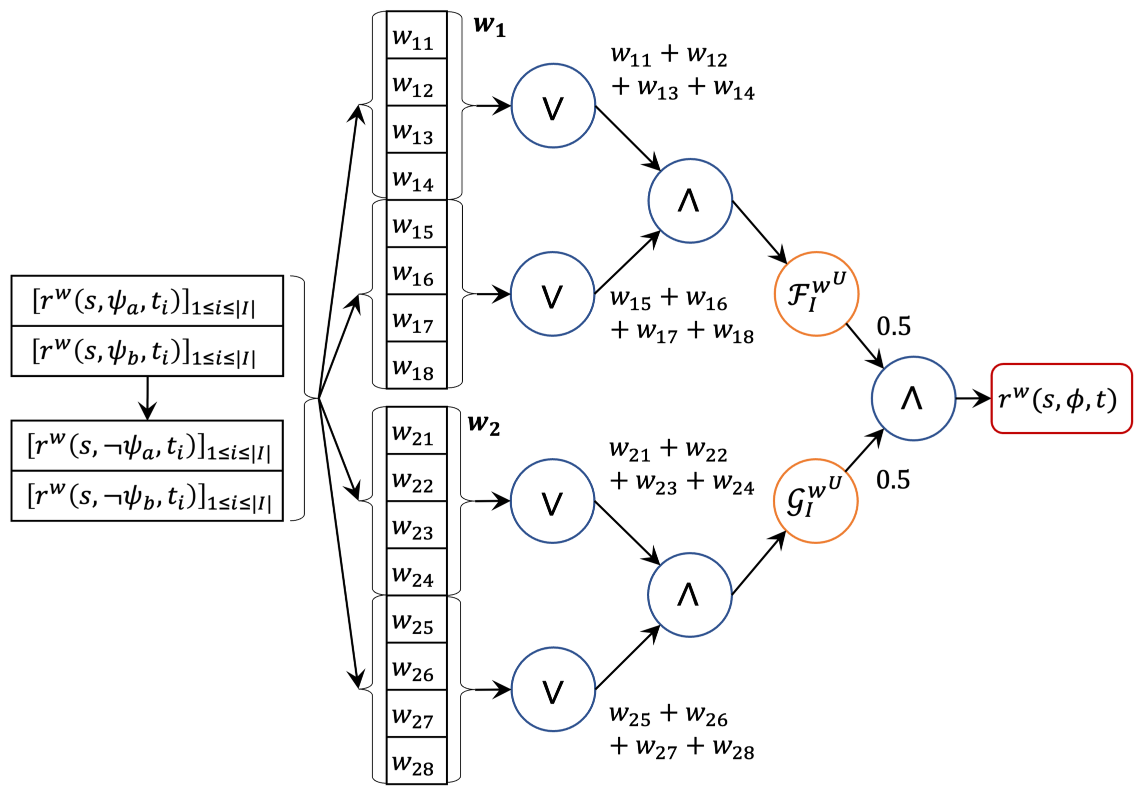

Fig 1. Neural Network Architecture for Two Predicates

RESULTS

Table I. Baseline Comparison of Representative Generated Explanations

Scenarios	Ours	Greedy	Top-3	Top-5
CtF Capture	$0.5\mathcal{F}[1.0\psi_{\mathrm{ba,rf}}]\wedge 0.5\mathcal{G}[0.30\psi_{\mathrm{ba,rf}} \vee 0.70\neg\psi_{\mathrm{ra,bf}}]$	$0.5\mathcal{F}[0.26\psi_{\mathrm{ba,rf}} \wedge (0.25\psi_{\mathrm{ba,rf}} \vee 0.33\neg\psi_{\mathrm{ra,bt}})\wedge (0.09\neg\psi_{\mathrm{ba,bt}} \vee 0.07\neg\psi_{\mathrm{ra,bt}})]\wedge 0.5\mathcal{G}[0.36\psi_{\mathrm{ba,rf}} \vee 0.64\neg\psi_{\mathrm{ra,bf}}]$	$\mathcal{F}[0.73\neg\psi_{\mathrm{ra,bt}} \wedge 0.27\neg\psi_{\mathrm{ba,bt}}]$	$\mathcal{F}[0.83\neg\psi_{\mathrm{ra,bt}} \wedge 0.17\psi_{\mathrm{ba,rf}}]$
CtF Capture 0	$0.5\mathcal{F}[1.0\psi_{\mathrm{ba,rf}}]\wedge 0.5\mathcal{G}[0.31\psi_{\mathrm{ba,rf}} \vee 0.69\neg\psi_{\mathrm{ra,bf}}]$	$0.5\mathcal{F}[0.08\psi_{\mathrm{ba,rf}} \wedge (0.40\psi_{\mathrm{ba,rf}} \vee 0.31\neg\psi_{\mathrm{ra,bt}})\wedge (0.07\neg\psi_{\mathrm{ba,bt}} \vee 0.14\neg\psi_{\mathrm{ra,bt}})] \wedge 0.5\mathcal{G}[(0.33\psi_{\mathrm{ba,rf}} \vee 0.58\neg\psi_{\mathrm{ra,bf}}) \wedge (0.05\psi_{\mathrm{ba,rf}} \vee 0.04\neg\psi_{\mathrm{ra,bt}})]$	$\mathcal{F}[0.67\neg\psi_{\mathrm{ra,bt}} \wedge 0.33\neg\psi_{\mathrm{ba,bt}}]$	$\mathcal{F}[0.83\neg\psi_{\mathrm{ra,bt}} \wedge 0.17\psi_{\mathrm{ba,rf}}]$
CtF Fight	$0.5\mathcal{F}[1.0\psi_{\mathrm{ba,rf}}]\wedge 0.5\mathcal{G}[0.33\psi_{\mathrm{ba,rf}} \vee 0.67\neg\psi_{\mathrm{ba,ra}}]$	$0.5\mathcal{F}[0.77\psi_{\mathrm{ba,rf}} \wedge (0.15\psi_{\mathrm{ba,rf}} \vee 0.08\psi_{\mathrm{ra,bf}})]\wedge 0.5\mathcal{G}[(0.15\neg\psi_{\mathrm{ba,bt}} \vee 0.11\neg\psi_{\mathrm{ra,df}}) \wedge 0.06\psi_{\mathrm{ba,rf}} \vee 0.04\neg\psi_{\mathrm{ra,df}}) \wedge (0.06\psi_{\mathrm{ba,rf}} \vee 0.06\neg\psi_{\mathrm{ba,bt}}\vee 0.04\neg\psi_{\mathrm{ra,df}}) \wedge (0.02\psi_{\mathrm{ba,ra}} \vee 0.19\psi_{\mathrm{ba,rf}}\vee 0.16\neg\psi_{\mathrm{ba,bt}} \vee 0.11\neg\psi_{\mathrm{ra,df}})]$	$\mathcal{F}[1.0\psi_{\mathrm{ba,rf}}]$	$\mathcal{F}[1.0\psi_{\mathrm{ba,rf}}]$
CtF Patrol	$0.5\mathcal{F}[1.0\psi_{\mathrm{ba,rf}}]\wedge 0.5\mathcal{G}[0.52\psi_{\mathrm{ba,rf}} \vee 0.48\neg\psi_{\mathrm{ra,bt}}]$	$0.5\mathcal{F}[0.35\psi_{\mathrm{ba,rf}} \wedge (0.09\psi_{\mathrm{ba,rf}} \vee 0.55\psi_{\mathrm{ra,df}})]\wedge 0.5\mathcal{G}[(0.29\psi_{\mathrm{ba,rf}} \vee 0.04\psi_{\mathrm{ra,bf}} \vee 0.07\neg\psi_{\mathrm{ra,df}})\wedge (0.30\psi_{\mathrm{ba,rf}} \vee 0.19\psi_{\mathrm{ra,bf}} \vee 0.02\neg\psi_{\mathrm{ba,bt}})\wedge 0.04\psi_{\mathrm{ra,bf}} \wedge 0.05\neg\psi_{\mathrm{ra,bf}}]$	$\mathcal{F}[1.0\psi_{\mathrm{ra,df}}]$	$\mathcal{F}[1.0\psi_{\mathrm{ra,df}}]$
CtF Roomba	$0.5\mathcal{F}[1.0\psi_{\mathrm{ba,rf}}]\wedge 0.5\mathcal{G}[0.35\psi_{\mathrm{ba,rf}} \vee 0.65\neg\psi_{\mathrm{ra,bf}}]$	$0.5\mathcal{F}[1.0\psi_{\mathrm{ba,rf}}] \wedge 0.5\mathcal{G}[0.21\neg\psi_{\mathrm{ra,bf}} \wedge 0.47\psi_{\mathrm{ba,rf}} \wedge(0.04\neg\psi_{\mathrm{ra,bf}} \vee 0.05\neg\psi_{\mathrm{ra,df}}) \wedge (0.18\psi_{\mathrm{ba,rf}}\vee 0.03\psi_{\mathrm{ra,df}} \vee 0.04\neg\psi_{\mathrm{ra,bf}} \vee 0.02\neg\psi_{\mathrm{ra,bt}})]$	$\mathcal{F}[1.0\psi_{\mathrm{ba,bt}}]$	$\mathcal{F}[0.48\neg\psi_{\mathrm{ba,bt}} \wedge 0.52\psi_{\mathrm{ba,rf}}]$
Fetch Push	$0.5\mathcal{F}[1.0\psi_{\mathrm{bt}}]\wedge 0.5\mathcal{G}[0.41\psi_{\mathrm{gb}} \vee 0.59\psi_{\mathrm{bt}}]$	$0.5\mathcal{F}[0.74\psi_{\mathrm{bt}} \wedge (0.10\psi_{\mathrm{bt}} \vee 0.16\psi_{\mathrm{od}})]\wedge 0.5\mathcal{G}[1.0\neg\psi_{\mathrm{bd}}]$	$\mathcal{G}[1.0\psi_{\mathrm{bt}}]$	$\mathcal{G}[1.0\psi_{\mathrm{bt}}]$
Robot Navi.	$0.5\mathcal{F}[1.0\psi_{\mathrm{eg}}] \wedge 0.5\mathcal{G}[1.0\neg\psi_{\mathrm{ec}}]$	$\mathcal{F}[1.0\psi_{\mathrm{eg}}]$	$\mathcal{F}[1.0\psi_{\mathrm{eg}}]$	$\mathcal{F}[1.0\psi_{\mathrm{eg}}]$

Table II. Baseline Comparison of Evaluation Metrics

	Conciseness				Consistency				Strictness			
Scenarios	Ours	Greedy	Top-3	Top-5	Ours	Greedy	Top-3	Top-5	Ours	Greedy	Top-3	Top-5
CtF Capture	**0.408**	0.207	0.196	0.223	**0.325**	0.078	0.083	0.125	**0.325**	0.078	0.083	0.125
CtF Capture 0	**0.417**	0.211	0.200	0.304	**0.450**	0.108	0.150	0.242	**0.450**	0.108	0.150	0.242
CtF Fight	**0.392**	0.181	0.250	0.250	**0.675**	0.088	0.500	0.500	**0.675**	0.088	0.500	0.500
CtF Patrol	**0.375**	0.181	0.250	0.275	**0.625**	0.046	0.500	0.525	**0.625**	0.046	0.500	0.525
CtF Roomba	**0.394**	0.119	0.213	0.162	**0.267**	0.061	0.100	0.088	**0.267**	0.061	0.100	0.088
Fetch Push	**0.417**	0.308	0.250	0.250	**1.00**	0.275	0.500	0.500	**1.00**	0.275	0.500	0.500
Robot Navi.	**0.475**	0.222	0.250	0.250	**0.675**	0.150	0.500	0.500	**0.675**	0.150	0.500	0.500

ANALYSIS

Baseline Comparisons

- Our method achieved higher mean accuracy with shorter explanation lengths.
- Lower variance in explanation quality across scenarios.
- Exception: "roomba" scenario due to suboptimal policy.

Qualitative Analysis

- **Our method**: Successfully inferred both task ($\mathcal{F}$) and constraint ($\mathcal{G}$) clauses.
- **Top-k methods**: Only inferred either task OR constraint, not both.
- **Greedy method**: Generated overly complex explanations.

Environment-Specific Insights

- **CtF scenarios**: Captured core task of flag capture and enemy behaviors.
- **Fetch push**: Correctly inferred block-target relationship.
- **Robot navigation**: Accurately captured goal-reaching while avoiding chaser.

Quantitative Results

- **Conciseness**: Up to 1.9 × improvement.
- **Consistency**: Up to 2.6 × improvement.
- **Strictness**: Up to 2.7 × improvement.

Limitations

- Approximated min/max functions affected constraint inference.
- Binary classification approach limited detection of rarely violated constraints in the positive and negative trajectories.

CONCLUSIONS

- Developed a **neuro-symbolic framework** for wSTL-based policy explanations.
- Improved **conciseness** and **interpretability** using predicate filtering, regularization, and pruning.
- Outperformed baselines in **seven robotics scenarios** with accurate, interpretable explanations.
- Limitation: approximated min/max functions, inferring a constraint with identical distributions.
- Future directions: **higher-order wSTL, human-in-the-loop refinement, real-world applications**.

ACKNOWLEDGEMENTS

This work was supported in part by ONR grant and JASSO Study Abroad Fellowship Program (Graduate Degree Program).

ILLINOIS