Master's Thesis Proposal: Applying Contrastive Reinforcement Learning to Learn General Policies for Classical Planning Domains

November 23, 2025

Abstract

This proposal describes research to combine contrastive representation learning with reinforcement learning methods to learn *generalized policies* for classical planning domains. The aim is to learn policies that generalize across instance sizes and object counts by training on small to medium instances and testing on larger, unseen instances. The approach leverages contrastive objectives to produce RGNN-based state embeddings to learn the Q-function. Evaluation will be on planning benchmarks such as Blocksworld, Logistics, Gripper, Sokoban, etc. and compared to existing generalized planning baselines.

1 Introduction

Generalized planning asks for policies that solve multiple problem instances of the same domain rather than a single instance. Traditional symbolic approaches construct compact general plans or policies using domain knowledge and combinatorial search; more recently, learning-based approaches have attempted to learn policies that generalize across instances using supervised [6] and reinforcement learning [5].

Concurrently, *contrastive representation learning* [2] has proven effective at learning robust, sample-efficient representations in RL for robotics domains [3]. This proposal explores whether contrastive RL objectives can facilitate learning of domain-general policies for classical planning tasks.

2 Background and Related Work

Classical planning and generalized policies. Classical planning domains, modelled in PDDL and STRIPS, provide a structured, relational representation: objects, predicates, actions with preconditions and effects. Research on generalized policies seeks compact policies that work across instances. Representative work includes symbolic and combinatorial methods as well as learning-based policy methods [4, 6, 5, 7].

Contrastive representation learning and RL. Contrastive methods (e.g., Contrastive Predictive Coding) learn embeddings by pulling together positive pairs and pushing apart negatives; these have been widely adopted across modalities. In RL specifically, CURL showed that contrastive objectives on image observations greatly improve sample-efficiency when combined with off-policy RL. Other works interpret contrastive losses as goal-conditioned value or forward-predictive objectives [2, 1, 3]. One of the key benefits of Contrastive RL is that exploration is emergent in the task formulation (see [8]).

Bridging the two areas. This work will try to bridge these two areas, looking at the benefits of using Contrastive RL combined with relational encoders to learn policies that can generalize across instances in classical planning domains.

3 Research Questions

1. Can contrastive RL objectives help learning of generalized policies that transfer to larger unseen instances?
2. How do relational inductive biases (GNNs / relational encoders) interact with contrastive objectives for improved generalization?
3. How does contrastive-RL-based generalized policy learning compare to previous RL approaches (especially [7]) on classical planning benchmarks?
4. One key point will be to understand in which domains Contrastive RL works well and where is struggles.

4 Methodology

The initial pipeline will have the following components:

1. **Graph Neural Network (GNN) state encoder.** Represent PDDL states as graphs (objects as nodes, predicates/relations as typed edges or node features). Use a R-GNN, following the architectures in [6].
2. **Contrastive representation learning objective.** During training, generate positive and negative pairs to train the encoder with a contrastive loss, by sampling future states from trajectories, in a similar manner to what is done in [3].

5 Datasets and Benchmarks

- Classical planning domains: Blocksworld, Logistics, Gripper, Ferry, Sokoban, and others from IPC / established bench collections. Training on small/medium sized instances (e.g., low object counts) and testing on larger instances to measure generalization.
- Baselines: models from prior work [5, 7] and LAMA traces.

6 Expected Contributions

- Empirical evidence on whether contrastive RL yields transferable embeddings that facilitate generalized policies in classical planning, or if Contrastive RL struggles in this setting (if so, why?)
- Comparison of other strategies for long-horizon RL such as Hindsight Experience Replay (HER) [7] in the setting of generalized planning.

7 Recommended Reading

- [3] for an introduction to Contrastive RL as Goal-Conditioned RL and [8] for additional experiments with Contrastive RL.
- [6] for understanding Relational-GNNs and general policies in classical planning.
- [5] for an introdution to learning general policies with RL and [7] to understand the baseline approach.

8 Library Suggestions

- For encoding PDDL states as GNNs *mimir-rgnn* (`https://github.com/simon-stahlberg/mimir-rgnn`) can be used, which is based on the architectures in [6].
- For running baseline RL algorithms and to use as base for the new implementation, *mimir-rl* (`https://github.com/simon-stahlberg/mimir-rl`) can be used
- *mimir* (`https://github.com/simon-stahlberg/mimir`) is a library for generalized planning that includes PDDL parsing, state space expansion and is a good library for functionalities around classical planning.

References

[1] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive Unsupervised Representations for Reinforcement Learning. In *ICML Workshops / PMLR*, 2020. URL: `https://arxiv.org/abs/2004.04136`.

[2] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748*, 2018.

[3] Benjamin Eysenbach, Tom Erez, Sergey Levine, and Ruslan Salakhutdinov. Contrastive Learning as Goal-Conditioned Reinforcement Learning. In *NeurIPS*, 2022. URL: `https://proceedings.neurips.cc/paper/2022/file/e7663e974c4ee7a2b475a4775201ce1f-Paper-Conference.pdf`.

[4] Guillem Frances, Blai Bonet, and Hector Geffner. Learning General Policies from Small Examples Without Supervision. In *AAAI*, 2021. URL: `https://bonetblai.github.io/reports/AAAI21-learning-policies.pdf`.

[5] Simon Ståhlberg, Blai Bonet, and Hector Geffner. Learning General Policies with Policy Gradient Methods. 2023. (Conference / proceedings).

[6] Simon Ståhlberg, Blai Bonet, and Hector Geffner. Learning General Optimal Policies with Graph Neural Networks: Expressive Power, Transparency, and Limits. In *ICAPS*, 2022.

[7] Simon Ståhlberg, Blai Bonet, and Hector Geffner. First-Order Representation Languages for Goal-Conditioned RL. In *PRL*, 2025. URL: `https://prl-theworkshop.github.io/prl2025-icaps/papers/2.pdf`.

[8] Grace Liu, Michael Tang and Benjamin Eysenbach. A Single Goal is All You Need: Skills and Exploration Emerge from Contrastive RL without Rewards, Demonstrations, or Subgoals. In *ICLR*, 2024.