Context Aware Policy Adaptation: Towards Robust Safe Reinforcement Learning

Phillip Odom
phillip.odom@gtri.gatech.edu
Georgia Tech

Eric Squires
eric.squires@gtri.gatech.edu
Georgia Tech

Zsolt Kira
zkira@gatech.edu
Georgia Tech

Abstract

Real world tasks are often safety-critical, requiring policies that respect safety constraints while also being able to safely adapt to novel situations. Typical safe reinforcement learning methods focus on adapting to shifts in the transition function but assume a fixed state space, limiting their ability to generalize to novel states. We consider the problem of safe reinforcement learning that must adapt to novel, potentially unsafe states. Our proposed approach for context aware policy adaptation leverages foundation models as a contextual representation that enables the agent to align novel observations with similar experience. We demonstrate empirically that our approach is able to generalize across different types of novelty that may include dangerous as well as safe states. We also show performance and safety are robust even when multiple types of novelty are introduced.

1 Introduction

Reinforcement learning (RL) has achieved broad success across a variety of tasks from complex games (OpenAI, 2018; Vinyals et al., 2019) to robotics (Tang et al., 2022). However, this is realized through extensive exploration in the environment, which may not be possible in safety-critical tasks where actions may lead to dangerous or costly outcomes. Furthermore, real-world tasks are complex, requiring agents to adapt to novel situations. Consider search and rescue where an agent is trained to navigate pre-disaster buildings. When deployed post-disaster, the agent may fail to adapt to hazards caused by novel objects or existing objects shifted within the environment. We consider the challenge of safely adapting to a distribution shift in the environment where novel situations are encountered.

Safe RL introduces constraints on the policy, requiring the agent to trade-off between higher reward and improved safety. Safety constraints are commonly satisfied by directly modeling the set of unsafe states (Thananjeyan et al., 2020) or through a safety critic that estimates the risk of unsafe behavior in the future (Srinivasan et al., 2020; Thananjeyan et al., 2021). These explicit models of safety are typically pre-trained on offline demonstrations of unsafe behavior. While demonstrations may come from the training task itself, there has been recent work that has explored leveraging related tasks to safely adapt to the training task (Zhang et al., 2020; Luo et al., 2021). However, previous work primarily consider related tasks with different dynamics but sharing the same set of states. In contrast, we consider how to safely adapt the policy in the presence of novel states.

Recent work has demonstrated the generalizability of foundation models for many applications including natural language processing, classification (Radford et al., 2021) and decision-making (Parisi et al., 2022; Khandelwal et al., 2022; Tam et al., 2022; Mu1 et al., 2022). These large pre-trained models incorporate natural context across data modalities and have been shown to improve exploration in RL (Tam et al., 2022; Mu1 et al., 2022; Gupta et al., 2022) and enable policy transfer (Xu et al., 2022). We aim to leverage the generalizability of foundation models and extend their use to safe RL for improving robustness to novel states.

In order to safely adapt to novel states, our approach for Context Aware Policy Adaptation (CAPA) aims to leverage foundation models to generalize an agent's policy across similar contexts by clustering over the states. The key intuition is that these clusters will align novel states with the agent's previous experience (i.e., novel states map to clusters of similar states) enabling the policy to execute similar behavior. We make the following key contributions: (1) We consider the problem of generalization in safe RL to novel states that may include unsafe as well as safe states. (2) We propose an approach for context-based safe RL that leverages foundation models to enable safe adaptation to this distribution shift. (3) Finally, our initial results demonstrate our approach can adapt to different types of novelty and even multiple types of novelty simultaneously.

2 Context Aware Policy Adaptation

RL is formulated as a Markov Decision Process (Sutton & Barto, 2018) that consist of a set of states S, a set of actions A, a reward function $R : S \times A \to \mathbb{R}$, a state transition function $P : S \times A \times S \to [0, 1]$, a discount factor $\gamma \in [0, 1]$ and a distribution of initial states $s_0 \sim \mu$. RL seeks to find a policy that maximizes the expected discounted reward, $R^\pi = E_{\mu,\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t R(s_{t+1}, a_t)\right]$. Safe RL extends this formulation to a constrained MDP (CMDP)(Altman, 1999), $M = (S, A, R, P, \gamma, \mu, C, \gamma_C)$, by including a cost function $C : S \to \{0, 1\}$ that indicates if a state violates the safety constraints, and a discount $\gamma_C \in [0, 1]$ associated with that cost. Similar to the reward, the expected discounted cost is $C^\pi = E_{\mu,\pi,P}\left[\sum_{t=0}^{\infty} \gamma_C^t C(s_{t+1})\right]$. Thus, the goal of safe RL is to find policy that maximizes the reward subject to the cost being less than ϵ, $\pi^* = \text{argmax}_\pi\{R^\pi : C^\pi \leq \epsilon\}$.

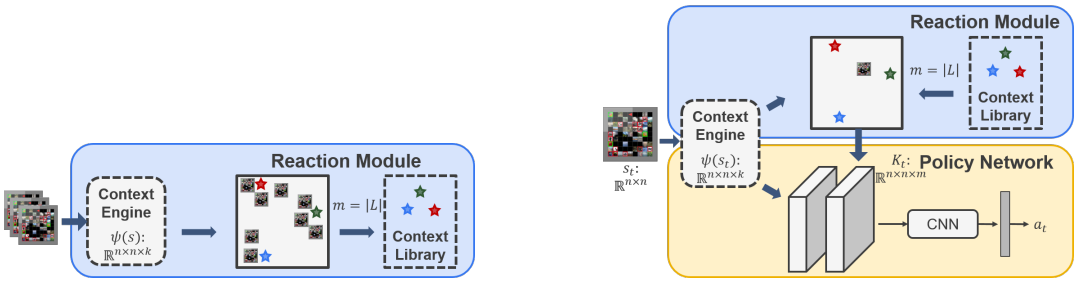
We consider the problem of generalization in reinforcement learning where an agent is trained in one environment, but must adapt to another environment where it may experience novel, potentially unsafe states. Specifically, consider a policy π trained on CMDP M but deployed in $M_2 = (S_2, A_2, R_2, P_2, ...)$ that extends M by introducing a novel set of states ($S_2 = S \cup S_2^{nov}$)[1]. Novel states include new objects that did not appear in M (e.g., external debris from disaster) or existing objects shifted to abnormal locations. While previous work typically focuses on changing dynamics but keeping the set of states fixed (i.e., $P_2 \neq P$ and $S_2 = S$), we learn a policy that effectively adapts to novel states, S_2^{nov}.

In order to generalize the notion of safety across unseen states, our approach for Context Aware Policy Adaptation (CAPA) leverages foundation models to enable efficient adaptation to distribution shift in the environment by incorporating contextual features into π. These features are generated by projecting observations onto a set of clusters learned from the training states, aligning novel states with the agent's previous experience and improving policy generalization. Our framework consists of two parts: the **reaction module**, which learns a set of clusters offline and the **policy network**, which leverages those clusters to make decisions.

Reaction Module: Motivated by the notion that the agent should react similarly to contextually similar states, the reaction module projects the current state into the space of clusters that are learned offline from the set of training states. Note that not all states encountered in the environment are unsafe so our approach does not limit the context library to safety-oriented features. The reaction module provides a contextual feature space that captures the relationship between states, enabling the policy to generalize behavior to novel states.

First, we build the context library (Figure 1a) by clustering over the set of states available during training. We generate a dataset $\{\psi(s)\}_{s \in S}$ where ψ is a contextual representation. The context library, L, consists of a set of m clusters. While any clustering algorithm could be used to construct the context library, we learn a clustering function via meta classification learning (MCL) (Hsu et al., 2019). MCL naturally yields a distribution over the clusters given pairwise information about the similarity of two observations. MCL is capable of discovering unseen classes (i.e., unknown m). However, we specify m in the experiments. Intuitively, we assume that the set of clusters will be significantly smaller than the number of states (i.e., $m \ll |S|$).

[1]Note that novel states also imply differences in P_2, R_2, and C_2



(a) Initialization of reaction module

(b) Policy adaptation

Figure 1: The reaction module is initialized by pre-training over a set of states from M. Leveraging the contextual representation, we cluster over the set of states and store these clusters in the context library (1a). CAPA leverages the reaction module to safely adapt to novel states by projecting the current state into space of clusters that represents the agent's experience. These clusters are used as features in the policy to generalize to novel states (1b).

During RL, the context library is fixed, enabling the agent to leverage this contextual information for decision-making. As shown in Figure 1b, the context engine takes in the current state s_t and produces a set of features $\psi(s_t)$ using a pre-trained foundation model. These contextual features are aligned with the context library by computing a distribution over the clusters via MCL.

Policy Network: The policy network, shown in Figure 1b, is a standard RL policy that takes in the state and produces an action. In addition to the inputs from the state, we incorporate a distribution over the clusters (K_t) from the reaction module as features, aligning it with previous context from the training environment and enabling the policy to generalize over novel states. We show empirically that the reaction module enables safe adaptation that is robust to novelty. This policy network and reaction module can be used with any safe RL framework.

3 Experiments

Our experiments aim to answer the following questions: **Q1:** Does the reaction module enable safe adaptation to novel, dangerous states? **Q2:** Can CAPA generalize to different types of novelty? **Q3:** Is CAPA robust as the amount of novelty increases? In order to evaluate our approach, we implement CAPA within PPO-Lagrangian (Achiam & Amodei, 2019) on RLlib (Liang et al., 2018), which solves the CMDP through an adaptive penalty. The policy network consist of a convolutional layer followed by fully connected layers. We use CLIP (Radford et al., 2021) as the contextual representation. The baseline (PPO-Lagr) uses the same network architecture except that it does not incorporate the features from the reaction module. We train PPO-Lagr for 25 million timesteps and CAPA, which converges more quickly, for 7.5 million timesteps. The resulting policies are then evaluated on 500 episodes across several settings. Each method is averaged over 10 runs. In order to demonstrate that our approach is able to adapt to different types of novelty, we design a scenario with multiple types of objects that require different actions.

Domain: We evaluate our approach on a modified MiniGrid (Chevalier-Boisvert et al., 2023) task, inspired by (Tsung-Yen Yang et al., 2021), with each cell in the grid represented by an image. The task objective is to navigate the grid to collect the fruit while avoiding the other vehicles. Grid cells may also be empty or contain an outdoor background scene that have no affect on the agent's reward. The agent has 4 actions: turn left, turn right, move forward, pick up. The agent's observation space includes a 7×7 view in front of the agent. Since the observation can be decomposed into a set of images, each image is passed through context engine and context library, producing an output for each grid cell. The input to the policy for each cell is a predicted image category.

The object types are each represented by superclasses from cifar100 (Krizhevsky, 2009): fruits and vegetables represent goals, vehicles_1 represent dangerous objects, large natural outdoor scenes

	Novelty	Reward (↑)		Cost (↓)	
		PPO-Lagr	CAPA	PPO-Lagr	CAPA
	No Novelty	5.29 ± 0.16	5.75 ± 0.08	2.28 ± 0.37	1.99 ± 0.25
Q1	Novel Danger	5.44 ± 0.10	**5.80 ± 0.08**	8.10 ± 2.28	**3.04 ± 0.29**
Q2	Novel Goals	3.22 ± 0.16	**5.70 ± 0.13**	3.17 ± 0.51	**2.10 ± 0.24**
Q2	Novel Background	5.32 ± 0.17	**5.85 ± 0.10**	3.67 ± 0.92	**2.19 ± 0.37**
	Comparing all settings with increasing amounts of novelty (Difference from No Novelty)				
Q3	1 type of novelty	4.66 (-0.63)	**5.79** (0.04)	4.98 (2.70)	**2.44** (0.46)
Q3	2 types of novelty	4.02 (-1.27)	**5.84** (0.09)	10.73 (8.45)	**2.81** (0.82)
Q3	3 types of novelty	3.26 (-2.03)	**5.89** (0.14)	21.62 (19.35)	**3.16** (1.18)

Table 1: We compare the reward and cost across settings with different types of novelty: novel danger, novel goals, and novel background. We also show average performance as the amount of novelty increases and compare it to the no novelty setting. The best performance is shown in bold.

represent background. In order to evaluate novelty, we split the categories for each superclass into those that appear during training or evaluation (4500 images of 9 categories) and those introduced only during evaluation (3000 images of 6 categories). Each episode consists of 3 goals. Dangerous and background objects are placed randomly throughout the environment ($\frac{1}{6}$ chance for each type). Note that within a single episode, the objects are sampled from one unique category for dangerous objects and two for background. The agent receives a reward of 2 for collecting each fruit and a cost of 1 for overlapping with another vehicle. Constraint violations do not terminate the episode.

Results - Adapting to Novelty: We compare CAPA to the baseline (PPO-Lagr) across different novelty settings in Table 1. Each setting may include multiple types of novelty: novel danger, novel goals and novel background. We evaluate the approaches based on their accumulated reward and cost. Note that higher reward and lower cost is better.

The first setting in Table 1 represents adaptation to dangerous obstacles. While the baseline maintains high reward, it has a significantly higher cost, resulting in a less safe policy. Alternatively, CAPA maintains high reward with a relatively lower increase in cost, indicating that CAPA enables safe adaptation to dangerous objects (**Q1**). Similarly, for novel goals, PPO-Lagr has a significant lower reward, while CAPA is able to effectively adapt and maintain high reward. For novel background, both approaches maintain their high reward, but our approach is also able to maintain lower cost. Collectively, these settings show that CAPA is able to capture different types of novelty (**Q2**).

We also explore settings where multiple types of novelty are introduced simultaneously. The lower block of Table 1 shows the average performance as you increase the amount of simultaneous novelty. Note the significant decrease in performance of PPO-Lagr in both reward and cost as the amount of novelty increases. CAPA is able to maintain its reward even as the amount of novelty increases and displays only minor degradation in terms of cost (**Q3**). Ultimately, CAPA is more robust across multiple types of novelty, achieving higher reward and lower cost.

4 Conclusion

We focus on the challenge of safe reinforcement learning in the presence of distribution shift where the agent must adapt to novel states. Our approach for context-aware policy adaptation leverages foundation models to cluster over the state space, representing the agent's experience. These clusters enable the policy to generalize to novel states. Our initial results suggest that our approach is able to adapt to multiple types and increasing levels of novelty even when introduced simultaneously.

Acknowledgments

The authors would like to thank the Georgia Tech Research Institute for funding this research out of Independent Research and Development (IRAD) funds.

References

Joshua Achiam and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. 2019.

Eitan Altman. *Constrained Markov Decision Processes*. 1999.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *CoRR*, abs/2306.13831, 2023.

Tarun Gupta, Peter Karkus, Tong Che1, Danfei Xu1, and Marco Pavone1. Foundation models for semantic novelty in reinforcement learning. In *NeurIPS Workshop on Foundation Model for Decision-Making*, 2022.

Ten-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *International Conference on Learning Representation (ICLR)*, 2019.

Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhav. Simple but effective: Clip embeddings for embodied ai. In *CVPR*, 2022.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Eric Liang, Richard Liaw, Philipp Moritz, Robert Nishihara, Roy Fox, Ken Goldberg, Joseph Gonzalez, Michael Jordan, and Ion Stoica. Rllib: Abstractions for distribution reinforcement learning. In *International Conference on Machine Learning*, 2018.

Michael Luo, Ashwin Balakrishna, Brijen Thananjeyan, Suraj Nair, Julian Ibarz, Jie Tan, Chelsea Finn, Ion Stoica, and Ken Goldberg. Mesa: Offline meta-rl for safe adaptation and fault tolerance. In *Workshop of Safe and Robust Control of Uncertain Systems at NeurIPS*, 2021.

Jesse Mu1, Victor Zhong, Roberta Raileanu, Minqi Jiang, Noah Goodman, Tim Rocktäschel, and Edward Grefenstette. Improving intrinsic exploration with language abstractions. In *NeurIPS*, 2022.

OpenAI. Openai five, 2018.

Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The (un)surprising effectiveness of pre-trained vision models for control. In *ICML*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Krishnan Srinivasan, Benjamin Eysenbach, Sehoon Ha, Jie Tan, and Chelsea Finn. Learning to be safe: Deeping rl with a safety critic. In *arXiv:2010.14603*, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Allison Tam, Neil Charles Rabinowitz, Andrew Kyle Lampinen, Nicholas Andrew Roy, Stephanie C.Y. Chan, DJ Strouse, Jane X Wang, Andrea Banino, and Felix Hill. Semantic exploration from language abstractions and pretrained representations. In *NeurIPS*, 2022.

Yang Tang, Chaoqiang Zhao, Jianrui Wang, Chongzhen Zhang, Qiyu Sun, Weixing Zheng, Wenli Du, Feng Qian, and Juergen Kurths. Perception and navigation in autonomous systems in the era of learning: A survey. *IEEE TNNLS*, 34, 2022.

Brijen Thananjeyan, Ashwin Balakrishna, Ugo Rosolia, Felix Li, Rowan McAllister, Joseph E. Gonzalez, Sergey Levine, Francesco Borrelli, and Ken Goldberg. Safety augmented value estimation from demonstrations (saved): Safe deep model-based rl for sparse cost robotic tasks. *IEEE Robotics and Automation Letters*, 5, 2020.

SBrijen Thananjeyan, Ashwin Balakrishna, Suraj Nair, Michael Luo, Krishnan Srinivasan, Minho Hwang, Joseph E. Gonzalez, Julian Ibarz, Chelsea Finn, and Ken Goldberg. Recovery rl: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 6, 2021.

Tsung-Yen Yang, Michael Hu, Yinlam Chow, Peter J. Ramadge, and Karthik Narasimhan. Safe reinforcement learning with natural language constraints. In *NeurIPS*, 2021.

Oriol Vinyals, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojciech M Czarnecki, Andrew Dudzik, Aja Huang, Petko Georgiev, Richard Powell, and et al. Alphastar: Mastering the real-time strategy game in starcraft ii, 2019.

Yifan Xu, Nicklas Hansen, Zirui Wang, Yung-Chieh Chan, Hao Su, and Zhuowen Tu. On the feasibility of cross-task transfer with model-based reinforcement learning. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022. URL https://openreview.net/forum?id=CAFK5R65IX.

Jesse Zhang, Brian Cheung, Chelsea Finn, Sergey Levine, and Dinesh Jayaraman. Cautious adaptation for rl in safety-critical settings. In *International Conference on Machine Learning*, 2020.