Invariant Representation Learning for Generalizable Imitation

Mohamed Khalil Jabri, Panagiotis Papadakis, Ehsan Abbasnejad, Gilles Coppin, Javen Shi

Invariant Representation Learning for Generalizable Imitation

Mohamed Khalil Jabri[1,2,3], Panagiotis Papadakis[1,3], Ehsan Abbasnejad[2,3], Gilles Coppin[1,3] and Javen Shi[2,3] *

1- IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France

2- Australian Institute for Machine Learning, The University of Adelaide, Australia

3- IRL CROSSING, CNRS, Adelaide, Australia

Abstract. We address the problem of learning imitation policies that generalize across environments sharing the same underlying causal structure between the system dynamics and the task. We introduce a novel loss for learning invariant state representations that draws inspiration from adversarial robustness. Our approach is algorithm-agnostic and does not require knowledge of domain labels. Yet, evaluation in visual and non-visual environments reveals improved zero-shot generalization in the presence of spurious features compared to previous works.

1 Introduction

Imitation Learning (IL) has emerged as a promising approach for sequential decision-making, leveraging expert demonstrations to reduce the need for exploration and reward engineering in Reinforcement Learning (RL). However, deploying learned policies in real world presents challenges in generalizing across diverse environments, making most works prioritize RL over IL.

Among such methods, bisimulation-based approaches have shown promise in achieving better generalization [1, 2, 3]. They, however, depend on reward information, unavailable in IL settings. Other techniques aim to remove task-irrelevant features without relying on reward signals. Works like [4, 5] employ domain labels to unlearn domain-dependent features using techniques like gradient reversal layers or increasing a domain discriminator entropy. Meanwhile, approaches like [6, 7] eschew domain labels relying instead on task-specific assumptions like time irrelevance or goal proximity as inductive bias to regularize the learned representations. Similarly, data augmentation methods like [8, 9] improve generalization but are only suited for visual inputs. In all these approaches, dependence on reward information, domain labels, or specialization to specific tasks or data modalities limits their broader applicability.

A standout among generalization approaches is Mixreg [10], being task and modality-agnostic and requiring neither reward nor domain labels. Mixreg encourages linear behavior between training examples, leading to smoother policies and improved generalization. We aim to match such a versatility level while

*This work was funded by the region of Brittany under the ROGAN project, the University of Adelaide and IRL CROSSNIG.

improving the zero-shot generalization performance. To this end, we generate adversarial data that breaks invariance properties and use it to formulate an adversarial invariance loss to regularize the learned representations.

In essence, our work bears similarities with PAADA [11] in viewing the environment irrelevant changes as adversaries. However, a drawback of PAADA lies in its objective for adversarial data generation, which revolves around minimizing the estimated advantage function. This is not suitable for IL settings where rewards are not given but estimated. In contrast, our work introduces a novel adversarial invariance loss aimed at learning invariant state representations devoid of reward or domain labels, leveraging adversarial data generalization. Furthermore, our method does not require explicit domain information yet achieves superior zero-shot generalization performance in both visual and non-visual environments with simulated spurious features.

2 Proposed Method

We focus on generalization in online IL where environments are sampled from a distribution $p(E)$ sharing a common causal structure, yet presenting domain-dependent differences. The agent is provided with a set of demonstrations collected in a set of environments $\mathscr{E}_{\text{demo}}$ and denoted $\mathcal{D}^{\mathscr{E}_{\text{demo}}}$, then interacts with a distinct set of environments $\mathscr{E}_{\text{inter}}$. We aim to learn a policy π that emulates the demonstrator behavior in unseen environments $\mathscr{E}_{\text{test}}$, maximizing the zero-shot cumulative reward $\mathbb{E}_{\tau\sim\mathcal{D}_{\pi}^{\mathscr{E}_{\text{test}}}} \sum_{t=0}^{T} \gamma^t r_t$ where τ is a trajectory, $\mathcal{D}_{\pi}^{\mathscr{E}_{\text{test}}}$ are trajectories generated under π, and r_t is the true reward only available at test time.

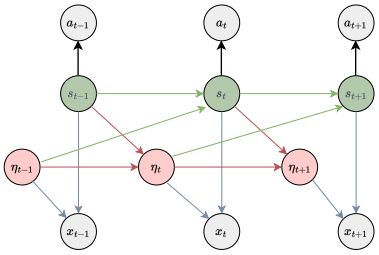

Fig. 1: Causal graph under optimal policy: Green denotes invariance across environments and red environment-specific mechanism. Actions are based on invariant states.

2.1 Disentangling the Causal State and the Noise Representations

We seek to learn a causal state encoder $\phi : \mathcal{X} \to \mathcal{S}$ that maps observation space $\mathcal{X}$ to a latent representation space $\mathcal{S}$, such that a policy with support on $\mathcal{S}$ maintains consistent performance across environments following $p(E)$. We denote as $s = \phi(x)$ the causal state of observation x and consider a noise encoder $\mu : \mathcal{X} \to \mathcal{H}$ to capture environment-dependent observation noise, with $\eta = \mu(x)$ denoting the noise representation of x. We also introduce a decoder $dec : \mathcal{S}\times\mathcal{H} \to \mathcal{X}$ to reconstruct observations from their state and noise latent representations. ϕ, μ, and dec, are parameterized by the vectors θ_ϕ, θ_μ, and θ_{dec}, respectively. An optimal policy is based on the causal state as shown in Fig. 1.

To learn the desired state and noise representations, the encoders must satisfy

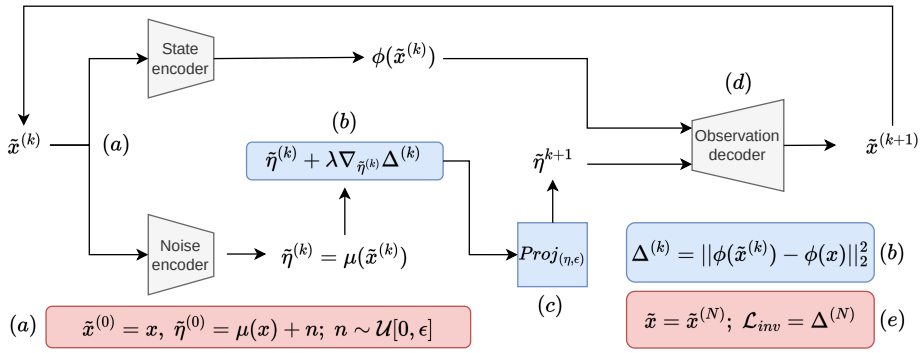

Fig. 2: Adversarial data generation pipeline

three conditions: (1) the state and noise representations need to be independent to capture distinct information; (2) the observation decoder must be able to restore the original, ensuring full information capture; (3) the learned causal state should be invariant across environments in $p(E)$.

To satisfy Condition 1, Mutual Information Neural Estimation (MINE) [12] is first employed to estimate the Mutual Information (MI) between the learned causal state and noise representations. MINE provides a way to estimate the MI of two random variables from their empirical distributions by iterative gradient ascent on some loss function. The estimated MI between the state variable s and the noise η is then minimized, giving the following independence loss:

$$\mathcal{L}_{ind}(\theta_\phi, \theta_\mu) = \mathbb{E}_{\tau \sim \mathcal{D}^{\mathscr{E}_{\text{demo}}} \cup \mathcal{D}_\pi^{\mathscr{E}_{\text{inter}}}} I(\phi(x; \theta_\phi), \mu(x; \theta_\mu); \theta_I) \tag{1}$$

For Condition 2, a reconstruction loss is minimized:

$$\mathcal{L}_{rec}(\theta_\phi, \theta_\mu, \theta_{dec}) = \mathbb{E}_{\tau \sim \mathcal{D}^{\mathscr{E}_{\text{demo}}} \cup \mathcal{D}_\pi^{\mathscr{E}_{\text{inter}}}} ||dec(\phi(x; \theta_\phi), \mu(x; \theta_\mu)) - x|| \tag{2}$$

Finally, to enforce Condition 3, we propose a novel adversarial loss that does not require knowledge of domain indexes as explained in the following.

2.2 State Representation Invariance via Adversarial Observations

Similarly to [11], policy generalization can be viewed as robustness against adversarial environments which act as adversaries by adding noise to observations. The noise is the source of the adversarial vulnerability, as it distracts the policy. A natural approach to address this problem is adversarial training, wherein generated adversarial observations are used as training data. Alternatively, one can explicitly impose the invariance condition on the state representation by minimizing the discrepancy between the representations of the original observations and their adversarial counterparts. This lays the ground for our invariance loss.

A natural choice for adversarial observations is observations whose noise component has been altered in a way that alters their learned causal states. Recall

that each observation is split into a supposedly invariant state and a noise using ϕ and η, and can be reconstructed using *dec*. We generate adversarial examples by iteratively decomposing and reconstructing the observations while adversarially perturbing the noise component to alter the causal state representation.

Fig. 2 explains in detail the steps involved in the generation of an adversarial version of observation x, which we denote as $\tilde{x}$. At each iteration k of the process, and **(a)** upon the decomposition the observation using the state and noise encoders, **(b)** the noise term is perturbed in the direction of the gradient of the distance between the causal state of the adversarial observation and the original causal state of the previous iteration, **(c)** then projected back onto a constraint set to ensure the alteration is minimal. This operation is defined by a function $\text{Proj}_{(\eta,\epsilon)}$ that projects the obtained noise onto the $l2$ ball of radius ϵ centered on the original noise η. At step **(d)** the next adversarial sample is then generated by the decoder via $\tilde{x}^{(k+1)} = dec(\phi(\tilde{x}^{(k)}), \tilde{\eta}^{(k+1)})$, and the process goes on. Note that in the first iteration, the noise term is perturbed through random sampling from the same constraint set, and that as we train the decoder, random noise is added to its input to ensure the reconstructed adversarial observations do not leave the original manifold. **(e)** After N iterations, the adversarial observation is defined as $\tilde{x} = \tilde{x}^{(N)}$ and the invariance loss is given as follows:

$$\mathcal{L}_{inv}(\theta_\phi, \theta_\mu) = \mathbb{E}_{\tau \sim \mathcal{D}^{\mathscr{E}_{\text{demo}}} \cup \mathcal{D}_\pi^{\mathscr{E}_{\text{inter}}}} ||\phi(x; \theta_\phi) - \phi(\tilde{x}; \theta_\phi)||_2^2 \tag{3}$$

where $\tilde{x}$ can be thought as the observation yielding the worst case divergence of the learning state from the original one, under minimal noise perturbation.

2.3 Learning the Imitation Policy with Invariant Representation

Our representation loss is the combination of the three previously defined losses: $\mathcal{L}_{rep} = \mathcal{L}_{ind} + \mathcal{L}_{rec} + \mathcal{L}_{inv}$. We pick GAIL [13] as an imitation baseline wherein a discriminator D is trained to discriminate expert trajectories from agent's ones and its output is used as reward to train the policy via RL. The policy is conditioned on the causal state: $\pi : \mathcal{S} \to \mathcal{A}$ where $\mathcal{A}$ denotes the action space, as well as the discriminator: $D : \mathcal{S} \times \mathcal{A} \to [0, 1]$. GAIL objective then becomes:

$$\min_\pi \max_D \mathbb{E}_{\tau \sim \mathcal{D}_\pi^{\mathscr{E}_{\text{inter}}}} [\log D(\phi(x; \theta_\phi), a)] + \mathbb{E}_{\tau \sim \mathcal{D}^{\mathscr{E}_{\text{demo}}}} [\log(1 - D(\phi(x; \theta_\phi), a))] \tag{4}$$

While GAIL alternates between iteratively updating π and D, we further update the state and noise encoders along with the decoder. At interaction time, a target state encoder $\hat{\phi}$ whose weights are exponentially averaged over time is utilized to reduce the variance of the learned state and stabilize the training.

3 Results

We conduct experiments in various generalization scenarios with deliberate injection of spurious correlations and visual distractions. In the first set, inspired by [14, 15], we augment observations in MuJoCo tasks ("Inverted Pendulum",

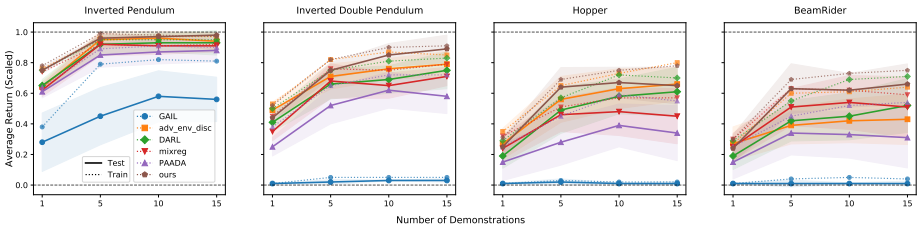

Fig. 3: Evaluation of generalization performance on different tasks.

"Inverted Double Pendulum", "Hopper") with n_{spur} independent noise components simulating spurious correlations, obtained by multiplying the last n_{spur} dimensions in the original observation by a full-rank random matrix defining a unique domain. Gaussian noise is also added to the last n_{spur} original dimensions to incentivize the agent to focus on spurious features. We also exclude domain labels from observations. In the second setting presented by an Atari game ("BeamRider"), domains are simulated and uniquely defined by random frame rotations. In all settings, demonstrations are obtained using pretrained RL policies on original observations.

We compare our approach with **(1) Mixreg** [10], similar in applicability and orthogonality to our method; **(2) PAADA** [11], treating changing environments as adversaries; **(3) DARL** [4], using Gradient Reversal Layer to unlearn domain-specific features; and **(4) adv_env_disc**, maximizing the entropy of a domain discriminator. (3),(4) have access to domain labels, (2) uses learned rewards instead of true ones, and all methods are based on **GAIL** [13].

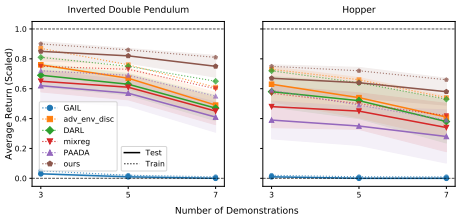

Fig. 4: Robustness to an increasing number of spurious dimensions.

The approaches are evaluated on average return in unseen test environments, scaled between 0 and 1, with 1 representing the expert's return. Varying numbers of demonstrations are given to the agent across 5 domains, and in each episode, the agent interacts with a random domain. At test time, results represent the average return of the learned policy across 10 runs, 2 in each of 5 unseen domains, averaged over 5 random seeds. For MuJoCo environments, $n_{\text{spur}} = 3$. As shown in Fig. 3, our method outperforms others in train and test environments, even compared to methods using domain labels. Concerning robustness to increasing number of spurious dimensions, tests are performed on two MuJoCo environments, the Double Inverted Pendulum and Hopper, with 10 demonstrations. We increase the difficulty by adding more spurious dimensions to the observations. Results in Fig. 4 show our method's superior robustness against these added complexities.

4 Conclusion

We tackled the generalization challenge in environments with spurious changes affecting policy decisions. Our approach performs invariant representation learning, inspired by adversarial robustness. The introduced invariance loss, incorporated into online IL, exhibited superior generalization capabilities without requiring true rewards or domain labels. This versatility makes our method valuable across diverse decision-making scenarios. Nevertheless, our approach relies on the presumption that causal features remain invariant across environments. Future work will explore scenarios where this condition does not hold.

References

[1] Norman Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. *arXiv preprint arXiv:1207.4114*, 2012.

[2] Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, 2019.

[3] Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

[4] Bonnie Li, Vincent François-Lavet, Thang Doan, and Joelle Pineau. Domain adversarial reinforcement learning. *arXiv preprint arXiv:2102.07097*, 2021.

[5] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 2015.

[6] Roberta Raileanu and Rob Fergus. Decoupling value and policy for generalization in reinforcement learning. In *International Conference on Machine Learning*, 2021.

[7] Youngwoon Lee, Andrew Szot, Shao-Hua Sun, and Joseph J Lim. Generalizable imitation learning from observation via inferring goal proximity. *Advances in neural information processing systems*, 2021.

[8] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *ICML*, 2020.

[9] Roberta Raileanu, Maxwell Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. Automatic data augmentation for generalization in reinforcement learning. *Advances in Neural Information Processing Systems*, 2021.

[10] KAIXIN WANG, Bingyi Kang, Jie Shao, and Jiashi Feng. Improving generalization in reinforcement learning with mixture regularization. In *Advances in Neural Information Processing Systems*, 2020.

[11] Hanping Zhang and Yuhong Guo. Generalization of reinforcement learning with policy-aware adversarial data augmentation. In *ICML Workshop on Decision Awareness in Reinforcement Learning Workshop*, 2022.

[12] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, 2018.

[13] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 2016.

[14] Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. Invariant causal prediction for block mdps. In *International Conference on Machine Learning*, 2020.

[15] Ioana Bica, Daniel Jarrett, and Mihaela van der Schaar. Invariant causal imitation learning for generalizable policies. *Advances in Neural Information Processing Systems*, 2021.