Learning Hierarchical Policies from Unsegmented Demonstrations using Causal Information

Mohit Sharma*, Arjun Sharma*, Nicholas Rhinehart, Kris M. Kitani

Introduction

- Learning complex tasks require learning sub-task specific policies.
- In this work, use a Directed Graphical Model to learn the interaction between sub-tasks and resulting state-action trajectory sequences
- Our algorithm CausalInfo-GAIL learns sub-task policies from unsegmented demonstrations by maximizing the directed information flow in the resulting graphical model.
- We also show how our approach connects with existing `Options' framework commonly used to learn hierarchical policies

Imitation Learning

- ***Generative Adversarial Imitation Learning (GAIL)*** **[1]**
  Objective,
  $$\min_{\pi}\max_{D}\mathbb{E}_{\pi}[\log D(s,a)] + \mathbb{E}_{\pi_E}[1-\log D(s,a)] - \lambda H(\pi)$$
- ***GAIL for mixture of experts*** **[2, 3]**
  c : Latent variable denoting expert

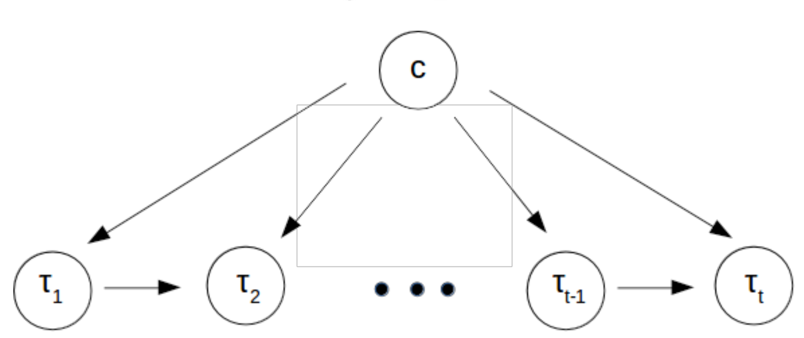


Figure 1: Graphical model in [2, 3]

Maximize lower bound to mutual information,

$$L_1(\pi, Q) = \mathbb{E}_{c\sim p(c), a\sim\pi(\cdot|s,c)}\log Q(c|\tau) + H(c) \le I(c;\tau)$$

Overall objective,

$$\min_{\pi,q}\max_{D}\mathbb{E}_{\pi}[\log D(s,a)] + \mathbb{E}_{\pi_E}[1-\log D(s,a)]$$
$$-\lambda_1 L_1(\pi,q) - \lambda_2 H(\pi)$$

Options framework

- Option: $o \in \mathcal{O}$
- Sub-policy: $\pi(a|s,o)$
- Termination policy: $\pi(b|s,\bar{o})$
- Option activation policy: $\pi(o|s)$

* - Equal contribution

Method

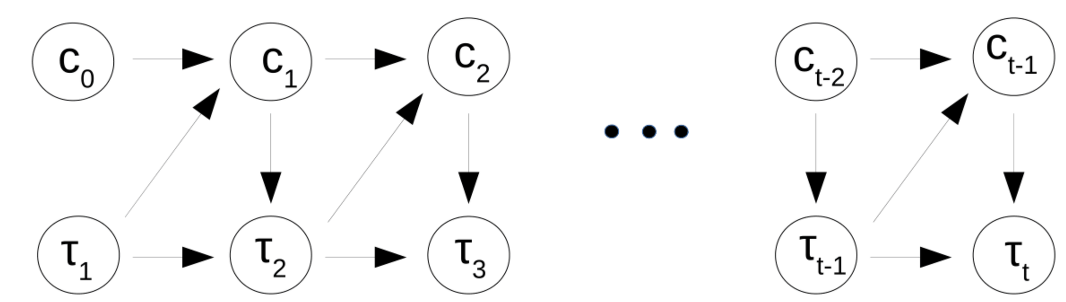


Figure 2: Graphical model used in this work

- *Using lower bound to mutual information,*

$$L(\pi,q) = \sum_t \mathbb{E}_{c^{1:t}\sim p(c^{1:t}), a^{t-1}\sim\pi(\cdot|s^{t-1},c^{1:t-1})}\Big[\log q(c^t|c^{1:t-1},\boldsymbol{\tau})\Big] + H(\boldsymbol{c}) \le I(\boldsymbol{\tau};\boldsymbol{c})$$

Dependence of q on the entire trajectory $\boldsymbol{\tau}$ precludes its use at test time where only trajectory up to current time is known

- *Causal Information*

$$I(\boldsymbol{\tau}\to\boldsymbol{c}) = H(\boldsymbol{c}) - H(\boldsymbol{c}\|\boldsymbol{\tau})$$
$$= H(\boldsymbol{c}) - \sum_t H(c^t|c^{1:t-1},\tau^{1:t})$$

- *Using lower bound to causal information,*

$$L_1(\pi,q) = \sum_t \mathbb{E}_{c^{1:t}\sim p(c^{1:t}), a^{t-1}\sim\pi(\cdot|s^{t-1},c^{1:t-1})}\Big[\log q(c^t|c^{1:t-1},\tau^{1:t})\Big] + H(\boldsymbol{c}) \le I(\boldsymbol{\tau}\to\boldsymbol{c})$$

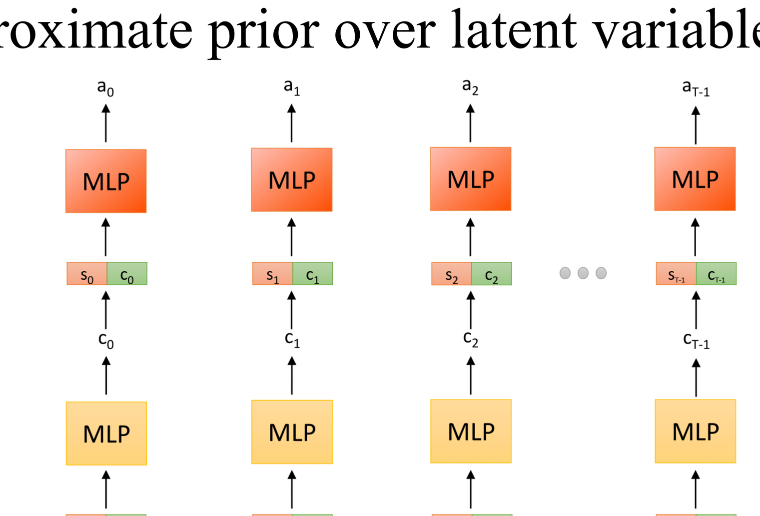
Using causal information removes dependence of q on future unobserved trajectory. Thus, q can now be used as a macro-policy to select the next sub-task latent variable.

Overall *Causal-Info GAIL* objective,

$$\min_{\pi,q}\max_{D}\mathbb{E}_{\pi}[\log D(s,a)] + \mathbb{E}_{\pi_E}[1-\log D(s,a)]$$
$$-\lambda_1 L_1(\pi,q) - \lambda_2 H(\pi)$$

where L_1 is the lower bound to *causal information.*

- *Variational Auto-encoder (VAE) pre-training*
  Learn approximate prior over latent variables using VAE



Experiments

- *Discrete environment*
  15x11 grid with 4 rooms connected via corridors. An object is placed at the center of a random room at the beginning of the episode. The agent spawns at a random location in the grid.

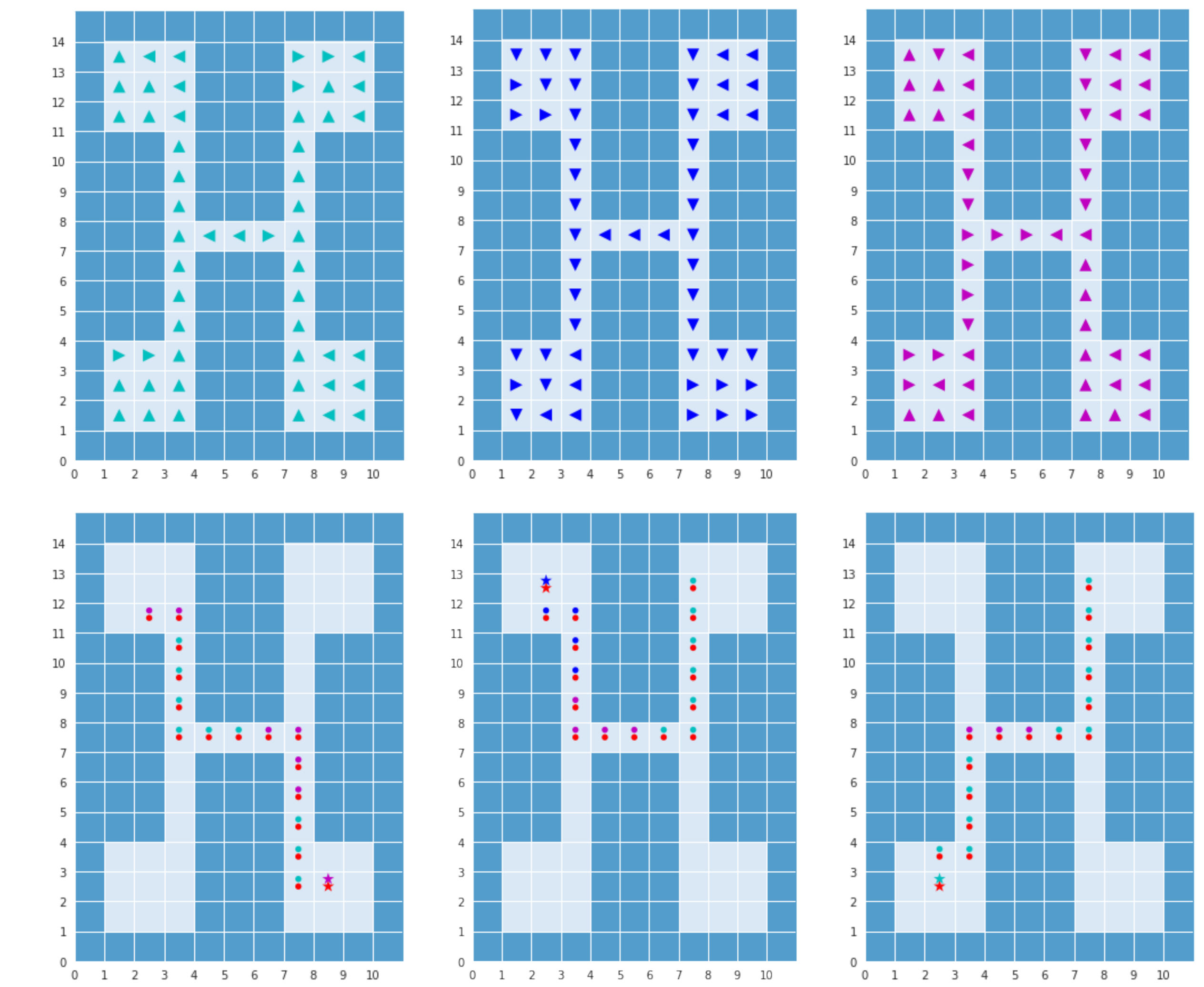


Figure 3: Visualization of sub-policy actions (top) and macro-policy actions (bottom)

- *Continuous environments*

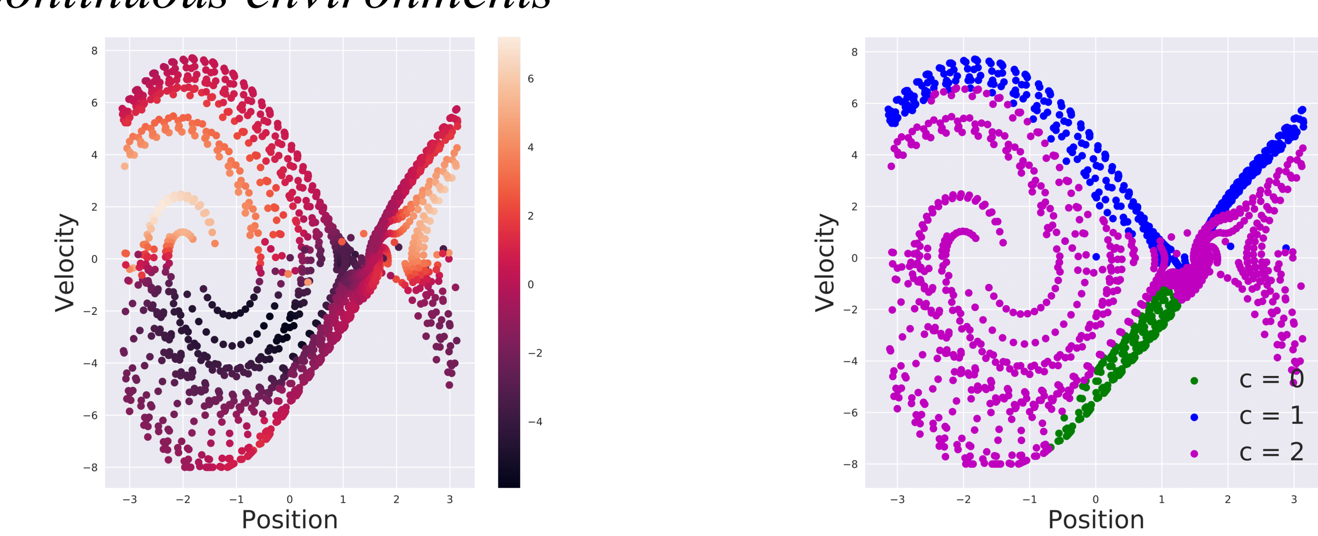


Figure 4: Visualization of sub-policy actions (left) and macro-policy actions (right) on Pendulum-v0

Environment	GAIL	VAE	Causal-Info GAIL
Pendulum	-121.4 ± 94.1	-142.9 ± 95.6	-125.4 ± 103.8
Inverted Pendulum	1000.0 ± 15.2	218.8 ± 8.0	1000.0 ± 15.0

Table 1: Returns over 300 episodes on continuous environments

References

[1] J. Ho and S. Ermon. "Generative Adversarial Imitation Learning." *NIPS*, 2016.
[2] Y. Li, J. Song and S. Ermon. "InfoGAIL: Interpretable Imitation Learning from Visual Demonstrations." *NIPS*, 2017.
[3] K. Hausman, Y. Chebotar, S. Schaal, G. Sukhatme and J. J. Lim. ''Multi-modal Imitation Learning from Unstructured Demonstrations using Generative Adversarial Nets.'' *NIPS*, 2017.
[4] C. Daniel, H. V. Hoof, J. Peters and G. Neumann. "Probabilistic Inference for determining Options." *Machine Learning*, 2016.
[5] R. Sutton, D. Precup and S. P. Singh. "Intra-option learning about temporally abstract actions." *ICML*, 1998
[6] J. Massey. "Causality, Feedback and Directed Information." *ISITA*, 1990