Causal Imitation Learning with Unobserved Coufounders

Deyu Liu
caliu17@cs.toronto.edu

Xudong Liu
xudong@cs.toronto.edu

Abstract:
Imitation learning is known to suffer from distributional shifts induced by expert and learned policies. [1] is among the first works to study this effect under a causal lens, where they aim to identify the direct causes of action from observed variables under the assumption of a linear energy model. We seek to consider the case where both observed and unobserved confounders exist in the observation space and how can we apply the factor model to improve the final performance of imitation learning.

Keywords: Imitation Learning, Robotics, Causal Inference

1 Introduction

Mimicking experts is a natural and efficient way to learn. Behavioral cloning, the simplest form of imitation learning, reduces policy learning to supervised learning by training a discriminative model to predict expert actions given observations. Despite the fact that imitation learning has been successfully applied in a variety of applications [2, 3], it still experiences the suffer from a distribution shift caused by expert and learned policies that induce the distribution of the training and testing states to be different from one another[4]. A Study has shown that the issue of distribution shift is frequently caused by causal misidentification. It can be resolved by an intervention method that identifies a causal structure from the observed variables under the assumption of a linear energy model to eliminate the potential adverse impact on IL performance resulting from more observations [1]. However, except for the distribution shift, the assumption which claims that covariates used by the expert are fully observed is another potential factor that contributes to unsatisfactory performance [5].

In this paper, we aim to extend the work that addresses causal confusion by weakening the ignorability assumption to enable the existence of unobserved confounders. We propose a deconfounder to identify and substitute the hidden confounder with an estimation of it in order to enhance the observational space for imitation learning. Then, we incorporate the deconfounder with the intervention algorithm to enhance the performance of imitation learning.

2 Related Work

The literature we explored for this project can be categorized into 2 different categories. One is the work about dealing with observed confounders in the observation that can cause causal confusion. [6]. The other is the works about dealing with unobserved confounders caused by data collection [7, 5]. In the first category, a mask is generated by the training data to hide the potential confounder in the observation space. This is based on the intuition that only a subset of observation represents the true causal factors that lead the model to the right result. The mask is found by Soft-Q learning and is based on the intervention of a mixed policy of all possible graphs and the environment. In the second category, a graphical criterion for imitability is proposed based on the partially observable structural causal model (POSCM) which models the unobserved nature of some endogenous variables.

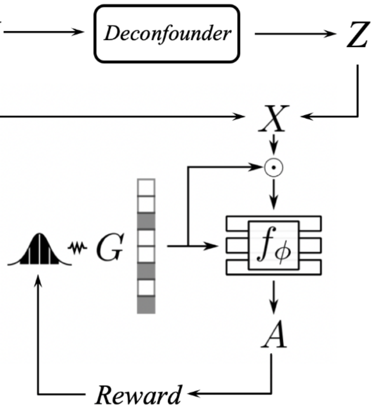

Figure 1: Overview of Our Framework

It determines the feasibility of imitation followed by presenting an algorithm to find the imitation policy [5]. To deal with the existence of hidden confounders in the estimation of treatment effect, the Factor Model Deconfounder is designed to replace the hidden confounders with the inferred latent variable and condition on it for computing the estimation [7]. Extending to the case in the time series setting, a Time Series Deconfounder is proposed to provide an unbiased estimation of treatment effects. It is implemented relied on building a factor model over time to obtain latent variables as the substitute for the unobserved confounders [8].

3 Methodology

In general, confounders can be classified into two categories: (1) the unobserved confounders, caused by failing to be recognized as a factor in the data collection; (2) the observed confounders, which are mistakenly identified as causes. To tackle the unobserved confounders, we propose to substitute them with an inferred latent variable derive from our deconfounders. In the presence of the observed confounders, we apply a mask learning procedure that identifies the causal structure of the observations and the action to eliminate the causal misidentification. A schematic of our proposed methodology is shown in Figure 1, which starts by substituting unobserved confounders and then masking out the observed confounders.

3.1 Deconfounder

To identify the hidden confounders, fitting a probabilistic factor model to represent the joint distribution of the assigned causes would be a feasible solution. Due to the fact that each cause is conditionally independent given the latent variable Z, then from Figure 2, we can show that by d-separation, there cannot be a confounder U (other than Z) that is the parent of multiple A_is. Hence, Z is proved to capture all unobserved multi-cause confounders U. Therefore, we aim to infer Z to replace U to expand the observational space for imitation learning.

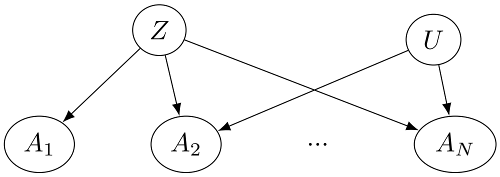

Figure 2: Graphic Model for the Deconfounder

In causal discovery, the traditional identification method is built based on three assumptions: (1) stable unit treatment value assumption (SUTVA), which states that the potential outcome of one individual is independent of the assigned causes of another individual [9, 10]; (2) Overlap, which claims that each individual has a positive probability of receiving each treatment level [11, 12]; (3) Ignorability, which requires no unobserved confounders and joint independence between causes and the potential outcome [11]. Relying on the idea of maintaining the SUTVA and Overlap assumption while weakening the Ignorability assumption by only requiring no unobserved single-cause confounder [7], we implement two deconfounders, Probabilistic PCA Deconfounder and VAE Deconfounder, to infer a latent variable Z as a substitute for unobserved multiple-cause confounders U.

The deconfounding procedure for the factor model can be summarized into three steps [7]. First, fitting the factor model M to capture the joint distribution of the assigned causes $P(A_{i1}, A_{i2}, ..., A_{iD})$. Second, inferring the latent variable for each individual from the fitted factor model by computing the conditional expectation of each individual's local factor weights $\hat{Z}_i = \mathbf{E}_M[Z_i|A_i]$. Finally, substituting the unobserved multiple-cause confounders U with the inferred variable Z, followed by performing imitation learning on it. Both of the deconfounders we applied in our experiment follow the same patterns.

3.2 Mask Learning

In this subsection, we learn a policy corresponding to the causal graph G. In the previous subsection, we identify unobserved confounders in observations. There are still observed confounders that need masking out by a causal graph. We parameterized the causal graph G as a vector of n binary variables, each representing the presence of an arrow from observation X_i to action A. We then optimize causal graph G so that policy $f_\phi(X \cdot G)$ minimizes imitation learning error.

Since the correct causal graph will avoid causal confusion and lead to appropriate decisions on testing, we can evaluate causal graph G by intervening with the environment and collecting the corresponding reward R_G. However, the candidate causal graphs increase exponentially in the observation dimension n. We instead sample G from causal graph distribution $p(G)$ by assuming a linear energy based model, $E(G) =< w, G > + b$. So causal graph distribution is

$$p(G) = \Pi_i \left(\mathbb{1}_{G_i=1}\sigma(\frac{w_i}{\tau}) + \mathbb{1}_{G_i=0}(1 - \sigma(\frac{w_i}{\tau}))\right).$$

We can utilize Soft Q-Learning to optimize graph distribution $p(G)$ when interacting with the environment as illustrated in Algorithm 1.

Algorithm 1 Policy Mask Intervention

```
Input: Policy network $f_\phi$
Initialize mask distribution $p(G)$
for $i = 1, 2, \dots, N$ do
    Sample mask $G \sim p(G)$
    Collect reward $R_G$ by executing policy $\pi_G$
    Update $p(G)$ with Soft Q-learning
end for
```

4 Experiment Results

To evaluate our approach in the proper problem setting, extensive experiments are conducted on the Hopper-v2 dataset [13] since it's a relatively simple dataset with a limited number of dimensions (11) in its observation space and 3-dimensional action space. The reward function will return a positive reward when the hopper object's shape remains in a pre-defined range and it's moving towards the correct direction. Thus, to obtain a high reward, our policy should apply the correct number of torque on the joint of the hopper object to maximize the reward value.

The observed confounder in our observation is created by appending the previous action to the observation space and the unobserved confounder is done by manually dropping dimension from

Mask Learning	# of Latent Dimensions	Deconfounder	Reward
×	1	VAE	1041.72
×	1	PPCA	1016.42
×	-	-	360.55
✓	1	VAE	1540.64
✓	1	PPCA	1056.65
✓	-	-	437.65

Table 1: 3rd observation is not available

Mask Learning	# of Latent Dimensions	Deconfounder	Reward
×	1	VAE	1282.13
×	1	PPCA	813.16
×	-	-	163.34
✓	1	VAE	1666.70
✓	1	PPCA	994.24
✓	-	-	415.31

Table 2: 2nd observation is not available

our observation. This is to create a confounded observation such that we can see the effect of our deconfounder and mask learning in the final reward.

In our experiments we seek to compare these cases: **(1)** Performance when we use observation with/without deconfounder when no mask learning, **(2)** Performance when we use observation with/without deconfounder with mask learning, **(3)** Performance when we dropped 2 dimensions with different number of latent dimension

From the results shown in Table 1 and Table 2, we can observe that the after we substitute the observation with our inferred latent variables, the performance of imitation learning is improved. More cases where we drop 2 dimensions are shown in Table 3.

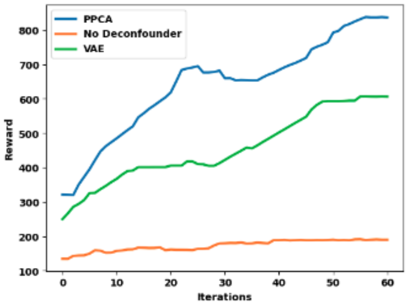

Figure 3: Reward vs Mask Learning Iterations

Mask Learning	# of Latent Dimensions	Deconfounder	Reward
✓	1	VAE	943.12
✓	2	VAE	1010.83
✓	1	PPCA	641.56
✓	2	PPCA	1363.82
✓	-	-	589.11

Table 3: 3rd, 4th observation is not available

Figure 3 demonstrates the relationship between the reward and the number of iterations in our mask learning procedure. We can see that the reward is higher when the deconfounder is applied and it's positively correlated with the number of iterations.

5 Conclusions, Limitations and Future Work

5.1 Conclusions

In this paper, we have presented an effective solution to address the shortcomings of traditional imitation learning caused by the confounders. First, to overcome the negative effect resulting from the strong ignorability assumption, we introduce two parametric deconfounders to enhance the observational space by substituting the multiple-cause hidden confounders with a latent variable inferred from the observed data. Next, to eliminate the causal confusion in observed data, we apply a mask learning algorithm to filter out the observed confounders that lead to misidentification in causal discovery. Experiment results carried out on Hopper-v2 dataset indicate that our method achieves a satisfying improvement in the performance of imitation learning.

A potential improvement we can apply is to try different kinds of factor models such as deep exponential family or Poisson factorization. A recurrent neural network (RNN) can also be a good choice when we have a time-series task instead of the current Markov environment. We can also try different kinds of input with the same approach. Image as input data can be very interesting as the input data is entangled such that the identification of multi-cause confounders can be difficult.

5.2 Limitations and Future Work

In this subsection, we will discuss some limits of our project and future work addressing them.

Posterior Distribution Approximation. In PPCA Deconfounder, the marginal likelihood $P(A)$ is tractable. Hence, it is feasible to directly derive the $P(W, Z|A)$ from Bayes' Rule, and it has a potential improvement in the IL outcome since a more accurate latent variable would be inferred due to no approximation.

Time Series Observations. In our setting, the policy is independent of previous state observations and actions, given current observations. This assumption is not necessary for some applications like medical treatment. In future work, we can extend our factor model based on recurrent neural networks to capture confounders encoded in time series.

Entangled Inputs. We assume observations X_i are disentangled in the factor model and mask learning. But observations can be entangled like image observations in some environments. In future work, we can train a VAE model to construct disentangled representations from these entangled observations.

6 Appendix: Contribution

Equal contribution. Listing order is random. Xudong was the first one to propose to apply mask learning to handle observed confounder and decided the direction of the project is to further solve unobserved confounder. Deyu did a significant amount of literature review and implement the imitation learning pipeline and chose the proper experiment settings. Both members were involved in nearly every detail and spent countless of time discussing.

References

[1] P. De Haan, D. Jayaraman, and S. Levine. Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4693–4700. IEEE, 2018.

[3] J. Hua, L. Zeng, G. Li, and Z. Ju. Learning for a robot: Deep reinforcement learning, imitation learning, transfer learning. *Sensors*, 21(4):1278, 2021.

[4] M. Bansal, A. Krizhevsky, and A. Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *arXiv preprint arXiv:1812.03079*, 2018.

[5] J. Zhang, D. Kumor, and E. Bareinboim. Causal imitation learning with unobserved confounders. *Advances in neural information processing systems*, 33:12263–12274, 2020.

[6] P. de Haan, D. Jayaraman, and S. Levine. Causal confusion in imitation learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/947018640bf36a2bb609d3557a285329-Paper.pdf`.

[7] Y. Wang and D. M. Blei. The blessings of multiple causes, 2018. URL `https://arxiv.org/abs/1805.06826`.

[8] I. Bica, A. M. Alaa, and M. van der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. 2019. doi:10.48550/ARXIV.1902.00450. URL `https://arxiv.org/abs/1902.00450`.

[9] D. B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593, 1980.

[10] D. B. Rubin. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

[11] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[12] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[13] H. Durrant-Whyte, N. Roy, and P. Abbeel. *Infinite-Horizon Model Predictive Control for Periodic Tasks with Contacts*, pages 73–80. 2012.