



## Learning Causal Variables from Temporal Observations

Phillip Lippe PhD Student, University of Amsterdam

21. February 2023

### Introduction Representation Learning

#### Vision



#### Not interpretable Unknown robustness

• • •

Figure credits: [1] Waymo tech block, 2017 [2] Cordts et al., The Cityscapes dataset. CVPR 2016.

### Introduction **Representation Learning**

Vision



**Structured Representation** AD: Human guidance what to model, causal factors



Figure credits: [1] Waymo tech block, 2017 [2] Cordts et al., The Cityscapes dataset. CVPR 2016.

Learning Causal Variables from Temporal Observations - Phillip Lippe

### **Causal Representation Learning**

- Given high-dimensional observations of a (dynamical) system, what is its latent causal structure?
  - Causal variables
  - Their cause-effect relations



### Causal Representation Learning

- Given high-dimensional observations of a (dynamical) system, what is its latent causal structure?
  - Causal variables
  - Their cause-effect relations



## CITRIS: Causal Identifiability from Temporal Intervened Sequences Setup



### **CITRIS** Architecture **CITRIS-VAE**

 $I^{t+1}$ 

 $x^t$ 

1 0 0 1

0:0\_

 $x^{t+1}$ 



069 **Causal structure assumptions:** We assume that the un-

070 derlying latent causal process is an unobserved dynamic

Bayesian network (DBN (Dean & Kanazawa, 1989; Murphy,

2002)) over the random variables  $(C_1, C_2, ..., C_K)$  with no

instantaneous effect and first-order Markov (i.e. the causal

parents of a factor at time t can only be in the previous

z

 $p_{\phi}(z^{t+1}|z^t, I^{t+1})$ 

Latent to causal

variable assignment

 $= \prod p_{\phi} \left( z_{\Psi_i}^{t+1} | z^t, I_i^{t+1} \right)$ 

075 time step t-1), for

071

072

073

074

077

083

084

085

087

088

089

094

095

Encoder  $q_{\theta}$ 

076 (i.e. the time serie

assume that the c 078 there are no additi

Encoder  $q_{\theta}$ Availability of int

> sume in each time intervened targets. We vector  $I^t$

on the causal variable argensition/prior prior state:  $I_i^t \perp I_i^t | C_1$ 

In this setup we tan model interventions with an arbitrary number of targets, including the empty s 0:0\_ data). Moreover, it can model both *perfect* which the target variable becomes indepen parents) and soft interventions (in which on distribution  $P(C_i | pa(C_i))$  of the target  $C_i$  $pa(C_i)$  changes).  $\hat{x}^{t+1}$ 

 $,..., \cup_K$  .

096 **Observation assumptions:** We assume that each latent 097 causal factors can be uniquely identified from the observa-098 Learning Causal Variable tions, Temphere exists a surjective mapions XPhilic Ifrom the charaction areas  $\mathcal{V} \subset \mathbb{D}N$  to the second factor areas  $\mathcal{C}$ 

004 variables (

t+1 respectively, and  $I^{t+1}$  describes the in at time step t + 1. We aim to leave a spring from observations to a latent space, observations gling the different causal factors. Thereby latent space to be larger than the Causalist i.e.  $\mathcal{Z} \subseteq \mathbb{R}^M, M \geq K$ , such that yanging can be modeled in multiple late Bradiniansi the encoding of multidimensionaddagoorse

ocess. Since somercarie al factors with among cat s when modelied enterpot up, we model apthoah  $p_{\phi}(z^{\dagger})^{\dagger}z^{D}$ , as the vit iables for xtomerer are m s a disentanglement ov nt variable of vailabilit rtsume in ea  $\mathcal{R}^{t}$  intervened  $085 \, {}^{i} \hat{x}^{0+1}$  on the cause where  $\Psi_i = \{j \in [1..M]\}$ latent variables assigned to the consider tata  $\emptyset$ .  $\psi(i)$  is thereby a learnable assignment

maps each latent variable to ppe of the interior  $\psi : \llbracket 1..M \rrbracket \rightarrow \llbracket 0..K \rrbracket, \text{ with } \psi(i) = 1$ latent variable  $z_j$  does not belong to any find the field of the second seco variable. Then, the objective of the mode the likelihood  $p_{\phi}(g_{\theta}(x^{t+1}))$ 29 distribution in  $\mathcal{D}$ .  $095\hat{r}^{t+p}a(C_i)$  cha

Before discussing the identifiability results tional case, we first state that: Observati



ere a ball can

Learning Causal Variables from Temporal Observations - Phillip Lippe

# CITRIS Experiments Pong

- CITRIS identifies the causal variables accurately
- Identified cause-effect relations closely follow ground truth





### CITRIS Experiments Visualizing the latent space



Novel combinations of causal factors













### Causal Representation Learning for Instantaneous Effects



Learning for Instantaneous and Temporal Effects." ICLR, 2023.



- **Causal Representation Learning** aims to learn generalizable, robust representations of causal variables in an environment
- **CITRIS** identifies causal variables in variety of environments by information about interventions
- Allows for interpretable, controllable latent spaces
- Opportunity for learning representations in complex, interactive environments like Embodied AI



Figure credit: [1] Szot, Andrew, et al. "Habitat 2.0: Training home assistants to rearrange their habitat." NeurIPS 2021.

Learning Causal Variables from Temporal Observations - Phillip Lippe

### References

ſ	D

<u>Phillip Lippe</u>, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. "**CITRIS: Causal Identifiability from Temporal Intervened Sequences**." In International Conference on Machine Learning (ICML). PMLR, 2022.

<u>Phillip Lippe</u>, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. "**Causal Representation Learning for Instantaneous and Temporal Effects.**" In International Conference on Learning Representations (ICLR), 2023.



Johann Brehmer, Pim de Haan, <u>Phillip Lippe</u>, Taco Cohen. **"Weakly supervised causal representation learning**." In Advanced in Neural Information Processing Systems (NeurIPS), 2022. **Slides and Papers** 

