BISCUIT: Causal Representation Learning from Binary Interactions

Phillip Lippe[1], Sara Magliacane[2,3], Sindy Löwe[2], Yuki M. Asano[1], Taco Cohen[4], Efstratios Gavves[1]

[1]University of Amsterdam, QUVA lab, [2]University of Amsterdam, AMLab, [3]MIT-IBM Watson AI Lab, [4]Qualcomm AI Research, The Netherlands

We identify **causal variables** from **interactions**, e.g. in Embodied AI.

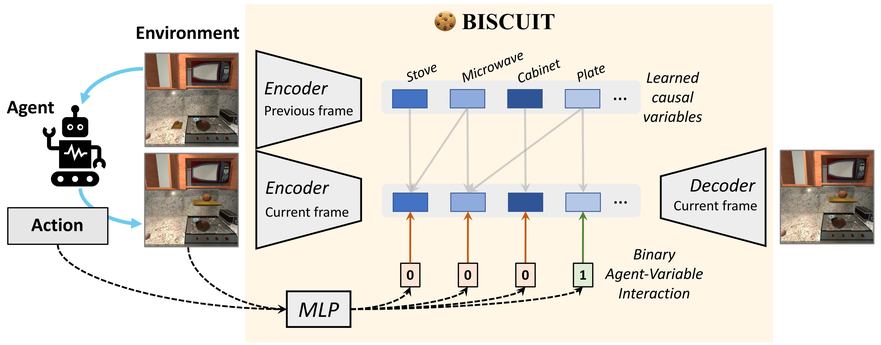
Paper and Demo

What is BISCUIT?

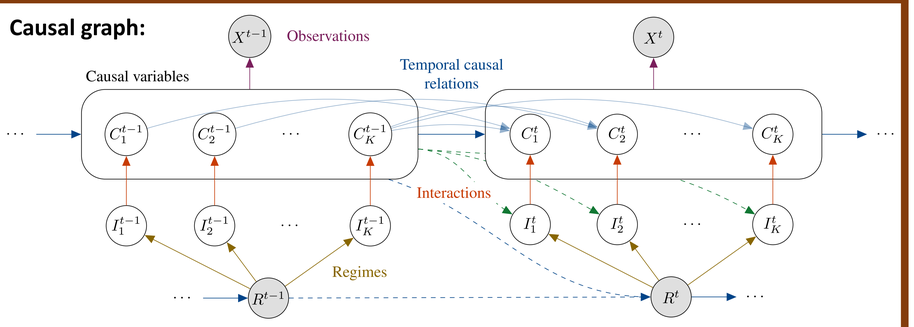
- BISCUIT learns **causal representations** from videos of **interactive systems**
  - Example: identify the causal variables (e.g. microwave state, plate position, etc.) of the kitchen environment
- Key assumption: interactions between agent and a causal variable can be described by a **binary interaction variable**
  - Interventional (e.g. open microwave) vs observational

How does BISCUIT work?

- Temporal VAE with causal vars in latent space and MLPs learning interaction vars
- Alternative setup: normalizing flow applied on autoencoder representation



BISCUIT provably identifies causal variables

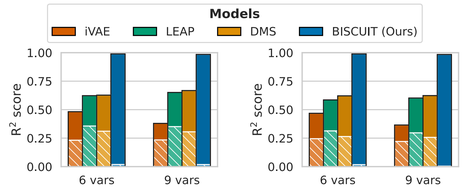


- Assumption 1: each causal variable has a distinct interaction pattern
- Assumption 2: mechanisms sufficiently vary on intervention or over time
- Allows for additive Gaussian noise models if mean changes over time

Experiments

- Evaluating accuracy of identifying causal variables from high-dimensional videos
- Actions being clicks or robotic input

Synthetic Environment (additive Gaussian noise)



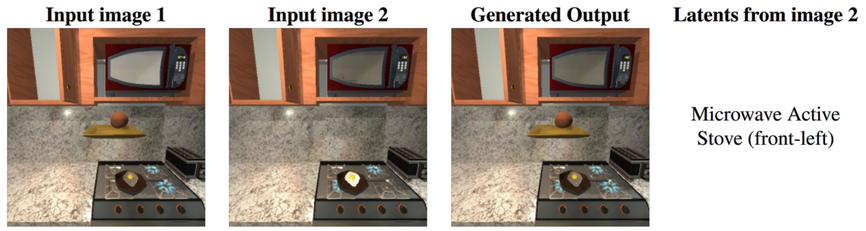
Robotic Environments

Table 1: R^2 scores (diag ↑ / sep ↓) for the identification of the causal variables on CausalWorld and iTHOR.

Models	CausalWorld	iTHOR
iVAE (Khemakhem et al., 2020a)	0.28 / 0.00	0.48 / 0.35
LEAP (Yao et al., 2022b)	0.30 / 0.00	0.63 / 0.45
DMS (Lachapelle et al., 2022b)	0.32 / 0.00	0.61 / 0.40
BISCUIT-NF (Ours)	**0.97** / 0.01	**0.96 / 0.15**

Simulating Interventions in Latent Space

- Latent interventions by (1) encoding two images, (2) replacing latents of first image by latents of second image for respective causal variables, (3) decoding new latents
- Achieves novel combinations of causal vars, e.g. uncooked egg on burning stove



Interaction Maps

- In iTHOR, an action is a random x-y position of object interacted with
- Visualizing learned interaction variables for each causal variable segments objects