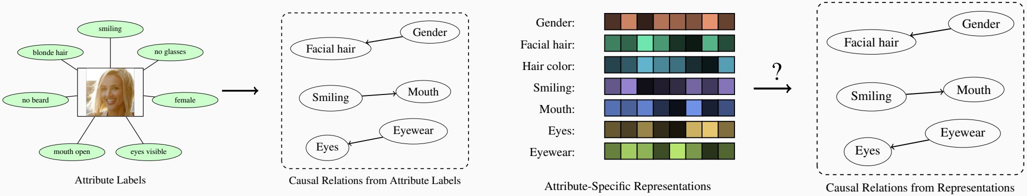


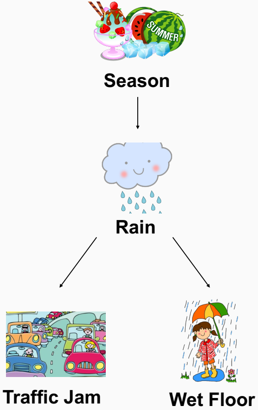
Do learned representations respect causal relationships?

Lan Wang and Vishnu Boddeti
Michigan State University

XAI4CV: Explainable Artificial Intelligence for Computer Vision

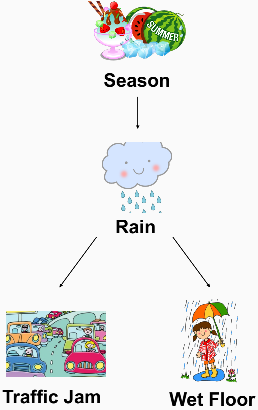
What is a causal relationship?

What is a causal relationship?

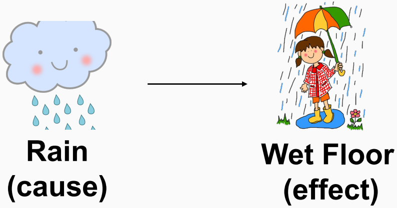


- Causal models capture the relations between random variables, which can be represented as graphs.

What is a causal relationship?



- Causal models capture the relations between random variables, which can be represented as graphs.



- Causal direction between two variables.

How to Infer Causal Relations?

Causal Discovery: Intervention

Example: Inferring the effect of treatment on some outcome.[1]

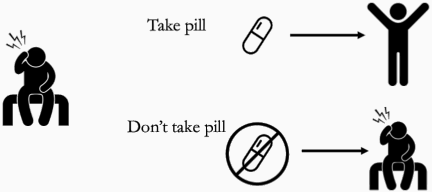


Figure 1: Causal effect

[1] https://www.bradyneal.com/causal-inference-course

Causal Discovery: Intervention

Example: Inferring the effect of treatment on some outcome.[1]

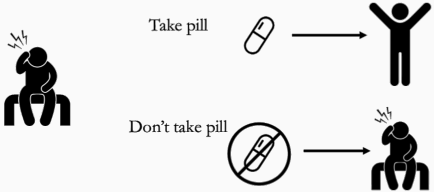


Figure 1: Causal effect

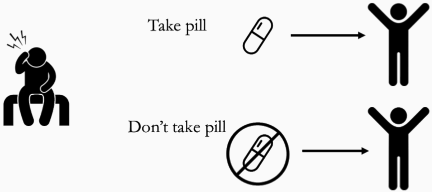


Figure 2: No causal effect

[1] https://www.bradyneal.com/causal-inference-course

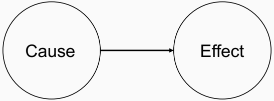
Causal Discovery from Observational Data

- In applications like computer vision: A large amount of data has already been collected.

Causal Discovery from Observational Data

- In applications like computer vision: A large amount of data has already been collected.
- It is difficult and even impossible to conduct control experiments (intervention) in most case.

Causal Discovery from Observational Data



Learning-based causal discovery

Exploit the manifestations of *causal footprint* present in real-world observational data.[a]

[a] Peters, Jonas, et al. "Causal discovery with continuous additive noise models." (2014).

- Relationships in causal direction are "simpler" than those in the anti-causal direction.
- Complexity metrics: MSE[2], Renyi Entropy[3], Kolmogorov complexity[4].

[2] Blöbaum, Patrick, et al. "Cause-effect inference by comparing regression errors." AISTATS 2018.

[3] Kocaoglu, Murat, et al. "Entropic causal inference." AAAI 2017.

[4] Marx, Alexander, and Jilles Vreeken. "Telling cause from effect using MDL-based local and global regression." ICDM 2017

Causal Discovery from Observational Data

Dataset $\longrightarrow$ $\boxed{\text{Model}}$ $\longrightarrow$ Causal Relation

Supervised methods

Exploit any and all possible causal signals in the observational data through learning.

[5] Lopez-Paz, David, et al. "Discovering causal signals in images." CVPR 2017.

[6] Goudet, Olivier, et al. "Learning functional causal models with generative neural networks." arXiv 2017

[7] Louizos, Christos, et al. "Causal effect inference with deep latent-variable models." arXiv 2017.

Causal Discovery from Observational Data

Dataset $\longrightarrow$ $\boxed{\text{Model}}$ $\longrightarrow$ Causal Relation

Supervised methods

Exploit any and all possible causal signals in the observational data through learning.

- NCC[5], GNN[6] ,CE-VAE [7]

[5] Lopez-Paz, David, et al. “Discovering causal signals in images.” CVPR 2017.

[6] Goudet, Olivier, et al. “Learning functional causal models with generative neural networks.” arXiv 2017

[7] Louizos, Christos, et al. “Causal effect inference with deep latent-variable models.” arXiv 2017.

Causal Discovery from Observational Data

Unsupervised methods

- Applied directly to the observational data which are agnostic to the data domain.
- Exploit only one type of causal footprint.
- Difficult to decide no causal relation case.

Causal Discovery from Observational Data

Unsupervised methods

- Applied directly to the observational data which are agnostic to the data domain.
- Exploit only one type of causal footprint.
- Difficult to decide no causal relation case.

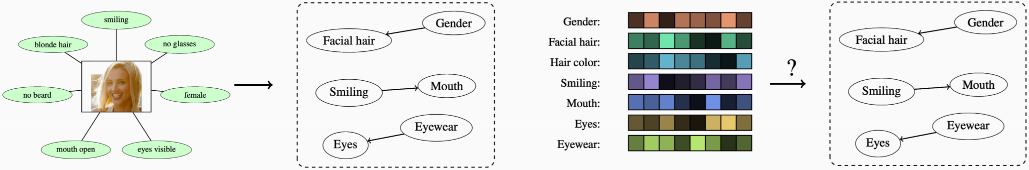
Supervised methods

- Can infer no causal relation case.
- Need groundtruth causal labels to train the causal classifier.

This paper:

Do learned representations respect causal relationships?

Do learned representations respect causal relationships?



Attribute Labels

Causal Relations from Attribute Labels

Attribute-Specific Representations

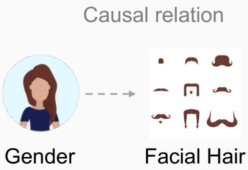
Causal Relations from Representations

Do learned representations respect causal relationships?

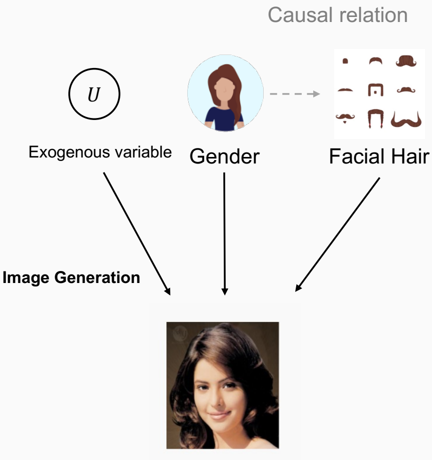
Gender

Facial Hair

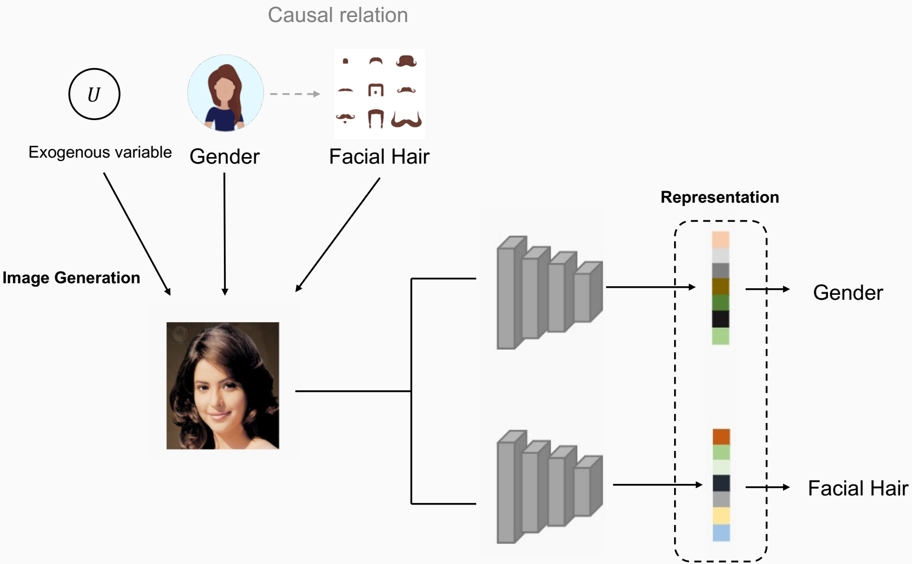
Do learned representations respect causal relationships?



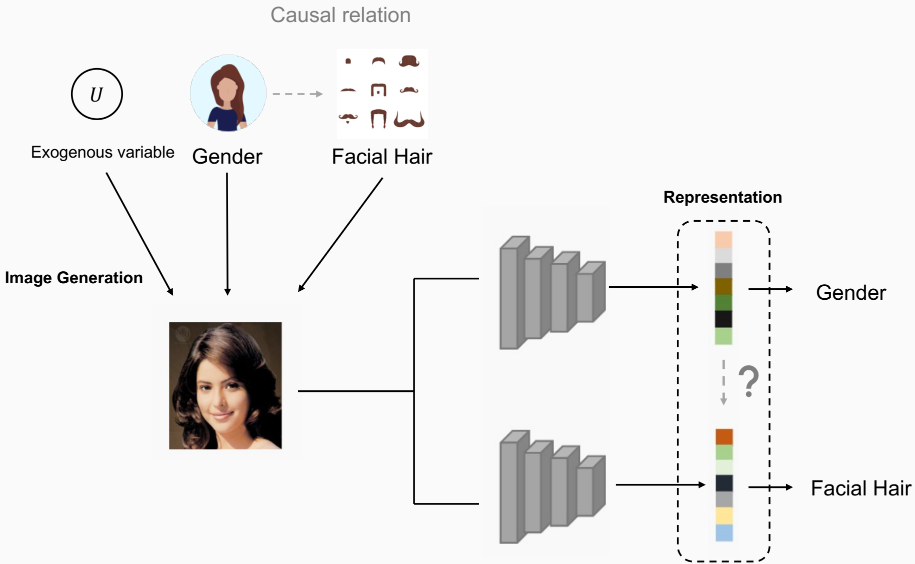
Do learned representations respect causal relationships?



Do learned representations respect causal relationships?

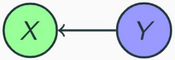


Do learned representations respect causal relationships?

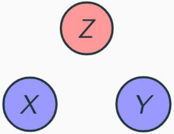


Synthetic Causal Feature Generators: six causal scenarios[8]

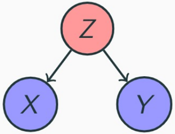
G_1
label=1



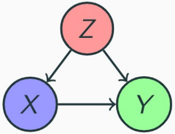
G_2
label=2



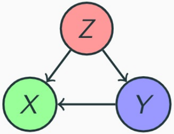
G_3
label=0



G_4
label=0



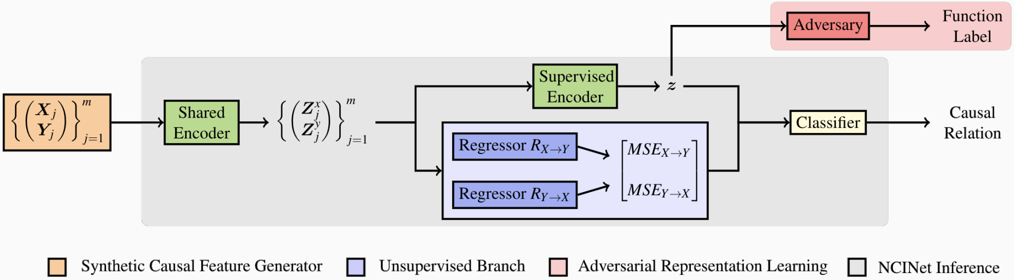
G_5
label=1



G_6
label=2

[8] Kalainathan, Diviyan, et al. "Discriminant Learning Machines." Cause Effect Pairs in Machine Learning. Springer, Cham, 2019. 155-189.

Neural Causal Inference Net (NCINet)



- L_C: Cross-entropy loss
- L_R: Regression loss
- L_A: Adversarial loss

$$Loss = L_C + L_R + \lambda L_A$$

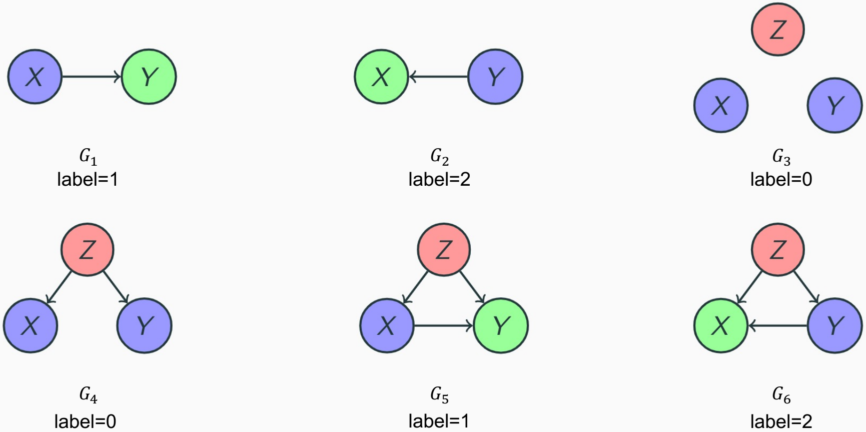
Experimental Design

Causal Inference with known relations between labels

In Real World Scenario:

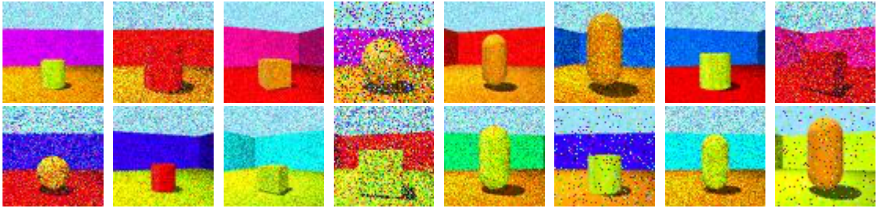
- **Problem:**
  - Ground truth causal relations are not known.
- **Our Solution:**
  - Generate controlled data with known causal relations.

Step 1: Generate Labels



Generate labels of 6 graphs using Gibbs sampling (3 classes for both X Y and Z)

Step 2: Generate images: 3D shape



- Two factors are decided by generated X and Y: floor hue, wall hue.
- The other factors are exogenous variables: object hue, scale, shape, and orientation.
- Add random noise: Gaussian, Shot, or Impulse.

Step 2: Sample images: CASIA-WebFace

- Annotations: color of hair, eyes, eye wear, facial hair, forehead, mouth, smiling, etc.
- Sample images with attributes consistent with generated labels.

Causal Inference with known relations between labels: Causal consistency

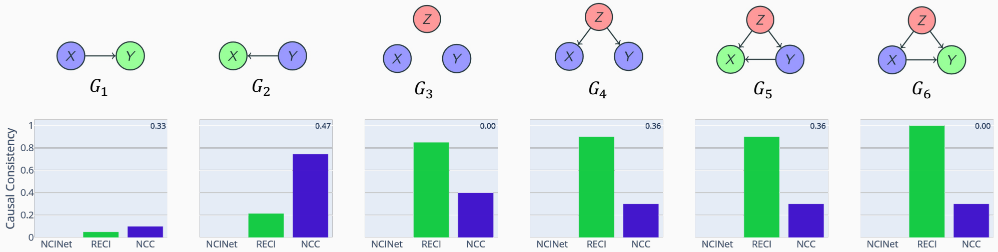
- Given $(x_j, y_j)_{j=1}^{m}$, split it into multiple non-overlapping subsets.

- Measure how many subsets are consistent with the causal relation between the labels.

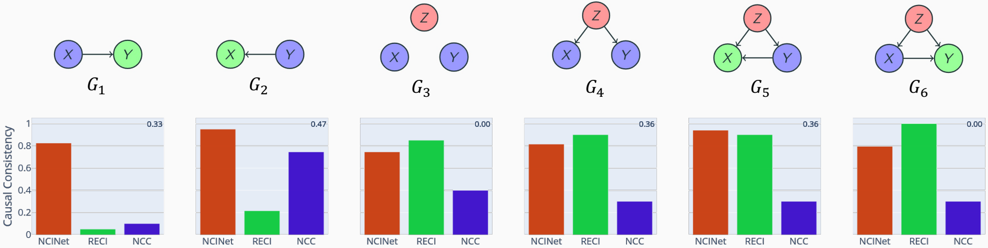
$$\text{Causal consistency} = \frac{\#\text{consistent subsets}}{\#\text{subsets}}$$

- Prevent outliers

Causal Inference with known relations between labels: 3D shape



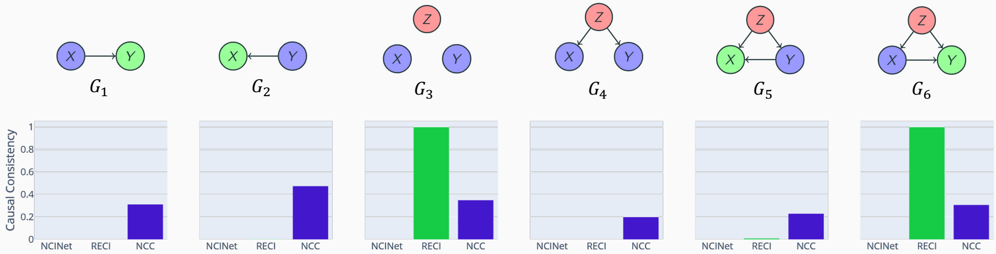
Causal Inference with known relations between labels: 3D shape



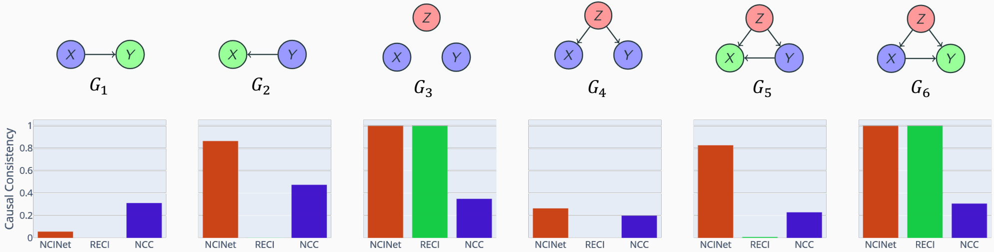
Takeaway

In controlled scenarios, learned attribute-specific representations indeed satisfy the same causal relations as the attributes.

Causal Inference with known relations between labels: CASIA-Webface



Causal Inference with known relations between labels: CASIA-Webface



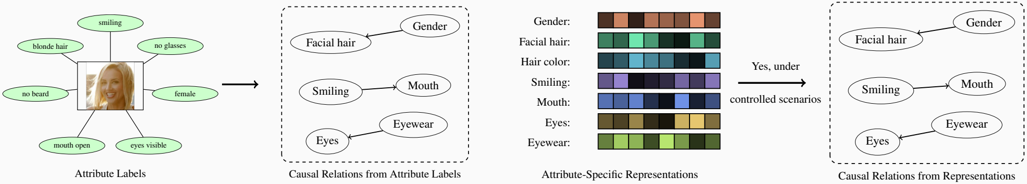
Takeaway

In more complex scenarios, the causal consistency decreases.

Analysis

- Generalization Ability
- **Question:** What is the relation between the causal consistency and training epochs?
- **Question:** What is the effect of overfitting on causal consistency of representations?
- **Question:** How does dimensionality of representation affect causal consistency?
- **Question:** How does network architecture affect causal consistency?

Conclusion



- Learned attribute-specific representations indeed satisfy the same causal relations between the corresponding attribute labels under controlled scenarios.
- Causal relations are positively correlated with predictive ability of representations.
- More investigation is needed for complex scenarios and data with weak causal relations.

https://github.com/human-analysis/causal-relations-between-representations