Causality in Modern Deep Generative Models

Humans reason and communicate about causality when learning about and interacting with the world. Most statistical models (including deep learning) are limited to measuring correlation, unable to make claims about causality between variables. In this project, we aim to augment a deep generative model such that it can represent causal relations in a human-interpretable way and guide its learning to be consistent with labeled causal relations.

We train our model on a dataset of MNIST images where the presence of one digit depends causally on the presence of another digit. We jointly train ALI and a linear classifier, which labels whether each image is an instance of a causal relation. This classifier takes as input the actual image as well as a “counterfactual” image. This counterfactual image is the result of inferring a value for the actual image in the latent space, applying a fixed set of transformations, and observing the corresponding “close” alternative images.

We aim to show whether this classifier, augmented with the latent space and counterfactual transform, can accurately distinguish between perceptually similar but causally different images.

Ideally, we would like to further show whether the model can learn to structure its generator such that the counterfactual transformation exposes the appropriate causal relations.

Data

Bayes Net where presence of one digit depends causally on another

MNIST digits with nonlinear transforms

Labels for whether an image is “causal” (1) or “noncausal” (0)

Experiments

Negative Control

Augmenting the GAN would be useless for improving the “perception” of the classifier if the classifier already has an easy time distinguishing the causal digit from the noncausal one. So we first show that we chose an appropriately difficult nonlinear transform for our dataset.

Classifier performance: {{CLASSIFIER PERFORMANCE ON ACTUAL TARGET IMAGES (low)}}

Positive Control

The GAN isn’t going to learn to represent the correct counterfactual images unless having the correct counterfactual images would be useful. So we generate an “oracle” set of augmented images with the correct counterfactuals, given our causal model of the data. We show that the linear classifier can use the correct counterfactuals to overcome its difficulty distinguishing the “causal” from “noncausal” target images.

Classifier performance: {{CLASSIFIER PERFORMANCE ON AUGMENTED TARGET IMAGES (high)}}

Regular GAN

We train a GAN and linear classifier together on the actual images and their labels.

We show that classifier performance is still low.

Classifier performance: {{CLASSIFIER PERFORMANCE ON ACTUAL TARGET IMAGES WHEN TRAINED WITH JOINT GAN+CLASSIFIER OBJECTIVE (low)}}

QUESTION: do the images look good?

QUESTION: inspect probabilities and CFs:

- What is P(3)? What is P(4)? What is P(3 and 4)? How do these compare to the correct probabilities?

    - CORRECT: P(3) = ; MODEL: P(3) = ;
    - CORRECT: P(4) = ; MODEL: P(4) = ;
    - CORRECT: P(3 AND 4) = ; MODEL: P(3 AND 4) = ;
- Inspect CFs (after adding CF transform):
    - (how) are the CFs different for “causal” and “noncausal” target images?
    - Given actual “causal” and “noncausal” target images, what is P_CF(3 | !4)? How does this compare to the correct CF probabilities?
        - CORRECT: P_CF(3 | !4) = ; MODEL: P(3 | !4) = ;

Causal GAN

We train ALI and a linear classifier together on the actual images and their labels. We augment the linear classifier’s input to include the CF transformed images.

We show that classifier performance improves (hopefully).

Classifier performance: {{CLASSIFIER PERFORMANCE ON ACTUAL TARGET IMAGES WHEN TRAINED WITH JOINT ALI+CFTRANSFORM+CLASSIFIER OBJECTIVE (high)}}

QUESTION: do the actual images look good?

QUESTION: inspect probabilities and CFs:

- What is P(3)? What is P(4)? What is P(3 and 4)? How do these compare to the correct probabilities?
    - CORRECT: P(3) = ; MODEL: P(3) = ;
    - CORRECT: P(4) = ; MODEL: P(4) = ;
    - CORRECT: P(3 AND 4) = ; MODEL: P(3 AND 4) = ;
- Inspect CFs (after adding CF transform):
    - (how) are the CFs different for “causal” and “noncausal” target images?
    - Given actual “causal” and “noncausal” target images, what is P_CF(3 | !4)? How does this compare to the correct CF probabilities?
        - CORRECT: P_CF(3 | !4) = ; MODEL: P(3 | !4) = ;

Related work

CausalGAN

Kocaoglu et al. (2018) present a method to separate the work of generating images from the work of learning a causal model. They assume that the true causal graph is given, and train one GAN to sample plausible configurations of features from that causal model and another GAN to generate images given different configurations of features. This separation allows them to sample implausible but imaginable images (e.g. women with mustaches) by *intervention* on a labeled variable (e.g. *mustache*) in the causal model, while still being able to sample only plausible images (e.g. men with mustaches) by *conditioning* on the same variable.

Background

Psychology

People use counterfactuals to understand “cause”

Gerstenberg et al. (2017) used eyetracking to more directly show that people spontaneously consider counterfactual outcomes while making causal judgements. They tracked eye movements while participants judged whether one billiard ball caused another to go through a gate. When asked for *causal* judgements like this, but not when asked about details of the *outcome* (how close the caused ball came to the edge of the gate), participants’ eye movements traced the counterfactual path that the caused ball would have taken had the causing ball been absent.

Expectations affect counterfactual and causal judgements

When researchers discuss the role of expectations in counterfactual and causal reasoning, they typically refer to two main categories of expectations: *statistical normality* and *prescriptive normality*. Statistical normality refers to the probabilities of what is likely to occur, whereas prescriptive normality refers to what “should” occur. Both seem to influence causal reasoning, and both concepts of normality seem to have similar effects.

People often imagine counterfactual changes that would undo a negative outcome, thinking things like “If only he had driven home a little later, he wouldn’t have gotten in a car crash.” Kahneman and Tversky (1981) show that people are sensitive to statistical normality when deciding on an “if only…” counterfactual. They presented stories where they varied whether a character Mr. Jones went home at the usual time via an unusual route or went home at an unusually early time via his usual route. Participants were much more likely to mention counterfactuals where Mr. Jones behaved more like he usually did, choosing the route to counterfactually change when the route had been unusual, and choosing the time to counterfactually change when the time had been unusual.

Faced with this case, participants tend to say that the professor caused the problem (Knobe and Fraser, 2008, Phillips et al., 2015).

Icard, Kominsky, and Knobe (2017) summarize the ways that statistical and prescriptive normality affect causal judgements into three main categories:

- **Abnormal inflation**: When there are multiple necessary causal factors for an outcome, people will be more likely to call something a cause if it is abnormal.
- **Supersession**: When there are multiple necessary causal factors for an outcome, people will be more likely to call something a cause if the *other* factors are normal.
- **No supersession with disjunction**: When there are multiple *sufficient* causal factors for an outcome (i.e. any one of the factors could have produced the outcome independently), people's judgements of how likely something is to be a cause does *not* depend on the normality of the other factors.

They show how a model of causal reasoning based on *probabilistic sampling* of counterfactual situations would result in these observed phenomena. They further present an additional principle, predicted by their model:

- **Abnormal deflation**: When there are multiple *sufficient* (but not necessary) causal factors for an outcome, people will be more likely to call something a cause if it is normal.

They present an experiment with animations simulating billiard balls colliding, representing either conjunctive (multiple necessary factors) or disjunctive (multiple sufficient factors) causal scenarios. As predicted, they demonstrate abnormal inflation in the conjunctive scenario and abnormal deflation in the disjunctive scenario.

References

Gerstenberg, Tobias, Matthew F. Peterson, Noah D. Goodman, David A. Lagnado, and Joshua B. Tenenbaum. 2017. "Eye-Tracking Causality." *Psychological Science* 28 (12): 1731–44. https://doi.org/10.1177/0956797617713053.

Icard, Thomas F., Jonathan F. Kominsky, and Joshua Knobe. 2017. "Normality and Actual Causal Strength." *Cognition* 161 (April): 80–93. https://doi.org/10.1016/j.cognition.2017.01.010.

Kahneman, Daniel, and Amos Tversky. 1981. "The Simulation Heuristic." TR-5. STANFORD UNIV CA DEPT OF PSYCHOLOGY. https://apps.dtic.mil/docs/citations/ADA099504.

Kocaoglu, Murat, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. 2018. "CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training." In *International Conference on Learning Representations*. https://openreview.net/forum?id=BJE-4xW0W.