GANs for Causal Inference: Using Vanilla and Counterfactually-Driven GANs to Learn Causal Structure

Sahil Chopra (schopra8@cs.stanford.edu)
Department of Computer Science, Stanford University

Abstract

Humans often try to use causal reasoning to derive insights about the world, but it is often difficult to accurately construct these models of causality. Here, we explore whether it is possible to train GANs to encode causal structure of underlying data in both unsupervised and semi-supervised settings. First, we propose a toy experiment, where data has been generated by a known Bayesian Network. Then, we train a GAN to generate potential worlds from experiment, examining whether the worlds as well as their counterfactuals follow the constraints imposed by the underlying causal network. Finally, we propose and train a new GAN Variant, the CFGAN, which leverages counterfactual statements about a given world to help a GAN learn remaining causal links in the underlying data. Results are still forthcoming.

Keywords: GANs; unsupervised learning; semi-supervised learning; Bayesian Models; causal structures

Introduction

Humans are constantly trying to reason about the world around them. It is our understanding of causality that allows us to properly interact with our worlds, both physically and socially. Thus in order to develop a better understanding of human cognition, we must develop better insight as to 1) How humans develop beliefs about causality and 2) What causal structures actually exist in the worlds around us.

Over the last two decades, cognitive scientists have been pursuing the first question – coupling human experiments with the framework provided by Bayesian Networks to construct models of human cognition in service of deriving a broader understanding of human reasoning (Gopnik et al., 2004). However, there hasn't been much progress towards solving the latter question.

As the computational power of GPUs have increased over the past decade, deep neural networks have become tractable to train. More recently, there have been an abundance of new neural network models that have been proposed for unsupervised learning. Specifically, a new breed of generative models called Generative Adversarial Networks (GANs) have shown significant promise on tasks such as face generation (Goodfellow et al., 2014; Radford, Metz, & Chintala, 2015).

A successful generator mimics true worlds. For example, a GAN that is successful at face generation will produce faces of men with mustaches and without mustaches, but won't produce faces of women with mustaches – as this unexpected according to the distribution of training images. It can thus be hypothesized that the GAN is intuiting an underlying causal structure of facial attributes.

In this paper, we try to explore the question – Can GANs learn causal structures in a set of given worlds? A reasonable question that follows is, how to test whether a GAN has learned a causal structure. There is a rich literature in cognitive science, positing that both children and adults often reason about causality in the world via counterfactual thinking. Counterfactuals force us to consider alternative outcomes within the world, given some aspect of the world has been perturbed (Harris, German, & Mills, 1996). If one can correctly reason about worlds that they have observed in addition to implausible and impossible worlds, one's mental model must contain the world's causal structure (Mackie, 1980). Thus, counterfactuals can be utilized to test the validity of a proposed causal structure, e.g. the internal machinery of the GAN.

Just as counterfactuals might be utilized to verify the presence of accurate causal structure within a generative model, they could also be used to help guide causal learning, as they do in humans. Thus, we posit a second question as well – Can we guide GANs towards learning additional causal structures over a set of worlds, by seeding the GAN with a small set of perceived causal links and and forcing it to reason over counterfactual statements?

Background and Related Work

Counterfactuals

Counterfactuals are often phrased in terms of a conditional proposition, where the existing world has been perturbed and an resultant observation is stated (Roese, 1997). As an example, let's say that our world is defined by a man Bob who is tall and plays basketball. A counterfactual of this world might be, "If Bob was short, he would not play professional basketball". This counterfactual makes intuitive sense, as we know that basketball players are generally tall and in order to compete with tall opponents, Bob must be tall himself.

Though simple, this example demonstrates the power of counterfactuals. Correctly reasoning about a counterfactual demonstrates an understanding of the causal structure of the world. If a person can intuit that a short person won't play professional basketball, we have evidence that they understand the underlying causal structure of world, i.e. the state of being tall has a causal connection to playing basketball. Alternatively, knowing the counterfactual as fact, i.e. that a short person won't play professional basketball, might help one learn the causal structure underlying this world itself (Mackie, 1980). Leveraging both of these implications, we hope to utilize counterfactuals in verifying the output of our Vanilla GAN and guiding our proposed CFGAN.

Generative Networks

In the subfield of generative networks, there are three primary approaches: 1) Boltzmann Machines (BMs), 2) Autoencoders (AEs), and 3) Generative Adversarial Networks (GANs).

Boltzmann Machines Boltzmann Machines consist of symmetrically connected neurons that make stochastic decisions as to their activation in order to approximate unknown probability distributions of input data. The learning algorithm for Boltzmann Machines is traditionally very slow but can be sped up by restricting the connections that are possible within these networks (Ackley, Hinton, & Sejnowski, 1985). These Restricted Boltzmann Machines (RBMs) consist of a single layer of visible units and a single layer of hidden units, with no visible-visible or hidden-hidden connections - thus restricting the structure of the RBM to that of a bipartite graph (Hu, Gao, & Ma, 2016).

Several deep architectures have been developed on top of RBMs. Specifically, Deep Belief Networks (DBNs) and Deep Boltzmann Machines (DBMs) have been shown to be successful. DBNs maintain an undirected RBM at its top layer, while leveraging previous layers as directed sigmoid belief networks (Hinton, 2009). On the other hand, DBMs are somewhat more general in that they allow for connections between hidden units across layers, while still restricting intra-layer connections to maintain the bipartite graph structure for faster learning (Salakhutdinov & Hinton, 2009). The issue presented by Boltzmann Machines is that their gradients are computationally intractable in most cases, so Markov Chain Monte Carlo (MCMC) methods are often used to approximate these derivatives (Goodfellow et al., 2014).

Autoencoders Autoencoders (AEs) are neural networks that are trained to produce outputs that are equal to their inputs (Bourlard & Kamp, 1988). Within the field of generative networks, there are two popular variants - Variational Autoencoders (VAEs) and Adversarial Autoencoders (AAEs). Variational Autoencoders consist of two halves, an encoder and a decoder. The goal of the encoder is to produce a low dimensional representation of an input image. The decoder then uses strided convolutions to produce an output image from this embedding, while minimizing some distance between the output and input images, i.e. reconstruction loss. VAEs leverage this encoder-decoder structure to introduce a second loss term, which computes the KL Divergence between the distribution of embedding vectors produced by the encoder and a unit normal distribution. The KL Divergence loss term is included so that the embeddings will fit some target distribution that can then be sampled and decoded to create new outputs (Kingma & Welling, 2013). Meanwhile, AAEs explicitly train a discriminator network to force the embedding vector to the target distribution (Makhzani, Shlens, Jaitly, Goodfellow, & Frey, 2015)

GANs Generative Adversarial Networks (GANs) rely on optimization over network outputs rather than latent variables. Specifically, GANs consist of a generator network (G) and discriminator network (D). Here D's objective is to correctly discriminate between between true images and generated images, while G's objective is to maximize D's error, i.e. fool the discriminator into believing that the generated images are real (Goodfellow et al., 2014). Mathematically, this is formulated as a Minimax game between G and D:

$$\min_G \max_D V(G,D)$$

$$V(G,D) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim p_z(z)}[log(1-D(G(z)))]$$

While training GANs it is often common to let the generator update once for every n epochs of training the discriminator. Similarly, in practice V may not provide sufficient gradient for G to learn well, as D can reject samples with high confidence during the early stages of training G. As a result, $log(1-D(G(z)))$ saturates, so instead we often maximize $log(D(G(z))$ to get stronger gradients early on during training (Goodfellow et al., 2014).

While RBMs and AEs may be interesting to explore for the development of causal models, we focus on GANs because of our interest in exploring learning guided by counterfactual reasoning. GANs produce their outputs from seed vectors z, which are made of independently and identically distributed random variables. This is important, as it means that all randomness is exogenized from the causal structure that a GAN might learn. It follows that if a GAN learned causal relationships, these would persist even if we perturbed a given z when counterfactualizing.

CasualGAN Recently, (Kocaoglu, Snyder, Dimakis, & Vishwanath, 2017) published a paper about a new architecture which they call CausalGAN. Specifically, they train two GANs in accordance with one another. They leverage the Celebrity Faces Dataset and a set of labels regarding different facial components, e.g. mustaches, baldness, etc. First they train a GAN on a subset of these labels to generate incomplete worlds over the possible facial features. They then utilize these generated worlds to train a second GAN, which is conditioned on the world. Out of this two-GAN structure they demonstrate that the network is able to infer the remaining causal links among labels, which were not included while the training the initial discrete-label-GAN.

(Kocaoglu et al., 2017) seem to bootstrap their model with a few known causal links, implicit in the subset of variables selected to train the first GAN. We hope to take this intuition forward by conditioning upon counterfactuals, with the belief that the framework for counterfactual reasoning should make it easier to learn causal structure and potentially allow the GAN to learn a more correct causal structure.

A Toy Problem

While real datasets, e.g. LFW, Imagenet, CIFAR-10, present the opportunity to learn on "natural photos", they also introduce a large amount of complexity and noise. Furthermore, there is no explicit causal relationship in these datasets. While

the ultimate goal is the ability to construct causal models in an unsupervised or semi-supervised fashion from arbitrary data, we first want to establish whether or not it is possible for a GAN to learn simple causal relationships that are known to exist within a dataset. Thus, we constructed a toy problem inspired by Gopnik's work with blickets (Fig.1) (Gopnik et al., 2004).

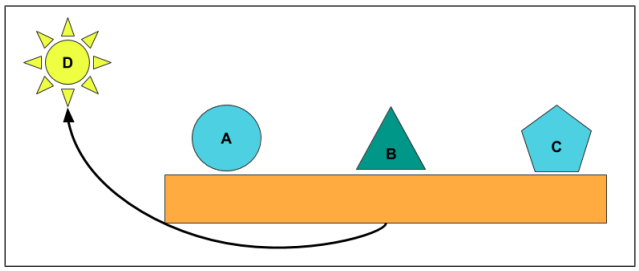

Figure 1: The toy problem consists of four objects. A,C (blue) are causally linked to the light D (yellow). B (green) has no association with A,C,D.

In our toy problem, there are four objects - A, B, C, and D. A and C are causally linked to D, i.e. placing these two objects in a certain configuration on the platform allows the light, D, to shine. Meanwhile B, the remaining object, has no causal relationship to the other presented variables. The presence of B poses an interesting question to the GAN, i.e. whether it can discern that B has no role in the game's causal structure (Fig. 2). Given this "blicket-like" setup, we then imposed constraints on A and C, under which D would be triggered. Any constraint might suffice for the purposes of this toy problem. We chose one where A and C must both have a "weight" greater than a certain threshold in order to trigger D.

$$A, B, C \sim Uni(0, 1)$$

$$D = 1.0, \text{ where } A >= 0.4 \text{ and } C >= 0.4$$

We then simulated the game within WebPPL to generated two datasets (10,000 and 100,000 samples) and proceeded to train our generative networks [1].

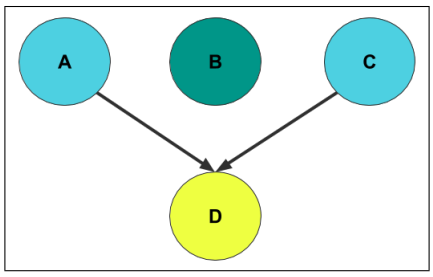

Figure 2: Causal structure of the toy problem.

[1] Code: https://github.com/schopra8/psych204_cfgans

Models

Vanilla GAN

The architecture for of the Vanilla GAN consisted of a 3-Layer MLP for the Discriminator and a 3-Layer MLP for the Discriminator. We intentionally chose a latent space of size 5 for the generator so that it didn't exactly match our causal structure of 4 items. In a realistic situation, one would not necessarily have a good intuition as to the number of latent space variables needed to exactly map the causal structure of the underlying data.

Network	Layer Description	Input Sz	Output Sz
D (L1)	Linear + Leaky ReLU	4	10
D (L2)	Linear + Leaky ReLU	10	10
D (L3)	Linear	10	2
G (L1)	Linear + Leaky ReLU	5	10
G (L2)	Linear + Leaky ReLU	10	10
G (L3)	Linear	10	4

Table 1: Vanilla GAN Architecture

CFGAN Architecture

We are still thinking about the ideal CFGAN Architecture and loss formulation. Currently, we have two primary ideas and would appreciate any and all feedback:

CFGAN Proposal #1 Train a generative network for some arbitrary number n epochs. If the loss is decreasing, i.e. it is able to fool the discriminator somewhat effectively, this implies that it is potentially learning some causal structure about the worlds. At this point, "inject" prior knowledge about the world, i.e. perceived causal linkages. In our toy example, we might present the link of $A \Rightarrow D$. Forward sample the GAN to produce some number of worlds c consistent with this linkage, e.g. worlds where A is high and D is 1.0. Now perturb the z latent vectors that produced these c worlds such that A is low and D is 0.0, i.e. the counterfactual of the perceived causal link. Finally utilize the perturbed z vectors $\hat{z}$ that produced the counterfactual worlds as additional training samples for either both the discriminator and generator, or just the generator alone.

CFGAN Proposal #2 Utilize the CasualGAN structure of training an "implicit generative model" on a subset of your labels. Then use the perceived causal relationships implicit to the worlds generated by this GAN to both condition the second GAN and propose counterfactuals that are utilized in training the second GAN.

Results

Vanilla GAN

Training the Vanilla GAN required somewhat extensive hyperparameter tuning - in the number of epochs to train the discriminator before starting the generator's training, number

of minibatches given to the discriminator versus the generator, learning rate, and learning rate decay. Additionally, the results were mixed across the 10,000 and 100,000 datasets.

For the 10,000 sample dataset, we trained the discriminator for 5 epochs before starting the generator, provided 2 minibatches to the discriminator for every minibatch provided to the generator, held learning rates .001 for both networks, and decayed by 0.1 every 20 epochs.

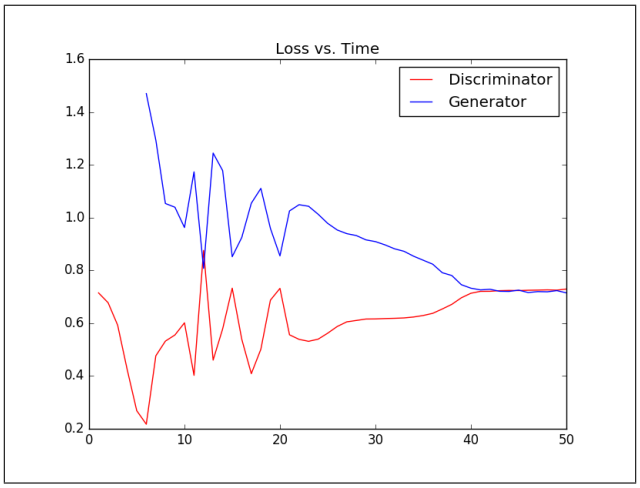

Figure 3: Training Loss for 10,000 sample game.

Eventually both the discriminator and generator converged (Fig. 3) for the 10,000 dataset. After training, we randomly sampled 50 vectors from a Gaussian to produce 50 outputs, 32 of which were in accordance with the underlying casual structure (64%). 31 of these 32 has $b < 0.5$, indicating that the GAN may have learned a spurious causal relationship between the value associated with B and the outcome D. Additionally, there as high variability in the values produced for A in these worlds but most values for C, were just slightly above 0.5. This may be do an imbalance in the underlying training data that could be phased out, given a larger dataset.

For the 100,000 sample dataset, we trained the discriminator for 0 epochs before starting the generator, provided 2 minibatches to the discriminator for every minibatch provided to the generator, held learning rates .001 for both networks, and decayed by 0.1 every 20 epochs.

The discriminator and generator did not converge originally in 50 epochs so we reran for 100 epochs, after which it converged on the 100,000 dataset (Fig. 4). After training, we randomly sampled 50 vectors from a Gaussian to produce 50 outputs, 46 of which were in accordance with the underlying casual structure (92%).This time around there was a greater spread in variability in both A and C values. Additionally, the bias towards small values for B in correct worlds disappeared when training on the larger dataset.

Together these initial results seem promising towards the hypothesis that GANs may accurately infer causal structure by themselves. Given a smaller batch data, the GAN was able to learn some structure, presenting a significant number of correct worlds – though the structure may not have been entirely right. Given the larger dataset, the GAN performed better – generating a larger number of accurate worlds. It will be interesting to perform counterfactual analysis on both models to determine how strong these underlying causal models actually are.

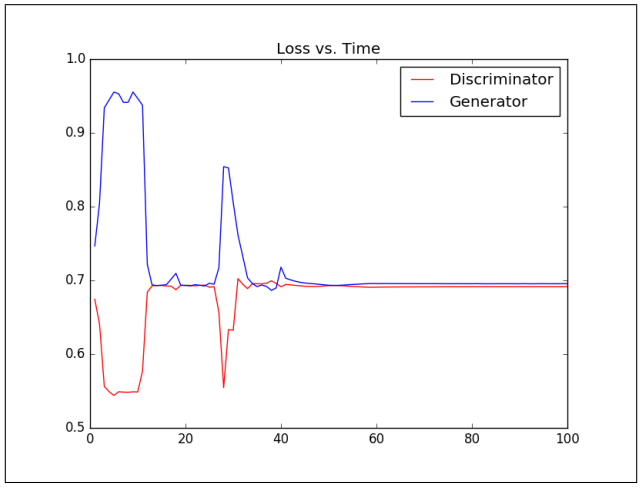

Figure 4: Training Loss for 100,000 sample game.

Next Steps

1. Perform counterfactual analysis on the Vanilla GAN
2. Try out one or both of the proposed CFGAN models.
3. Reason about better loss formulations for this task. Perhaps a separate loss for counterfactual statements?
4. If time permits, move beyond the toy problem and try on a larger dataset, e.g. Celebrity Faces.

Acknowledgments

Special thanks to Erin Bennett for all her help in developing the ideas presented above.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines*. *Cognitive Science*, *9*(1), 147–169.

Bourlard, H., & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, *59*(4), 291–294.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). *Generative adversarial networks.*

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, *111*(1), 3.

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, *61*(3), 233–259.

Hinton, G. E. (2009). Deep belief networks. *Scholarpedia*, *4*(5), 5947.

Hu, H., Gao, L., & Ma, Q. (2016). *Deep restricted boltzmann networks.*

Kingma, D. P., & Welling, M. (2013). *Auto-encoding variational bayes.*

Kocaoglu, M., Snyder, C., Dimakis, A. G., & Vishwanath, S. (2017). Causalgan: Learning causal implicit generative models with adversarial training. *CoRR*, *abs/1709.02023*.

Mackie, J. L. (1980). *The cement of the universe: a study of causation*. Clarendon.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). *Adversarial autoencoders.*

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, *abs/1511.06434*.

Roese, N. J. (1997). Counterfactual thinking. *Psychological bulletin*, *121*(1), 133.

Salakhutdinov, R., & Hinton, G. (2009). Deep boltzmann machines. In *Artificial intelligence and statistics* (pp. 448–455).