Sahil Chopra
Psych 204/CS 428

Project Proposal: GANs for Causal Inference

Initially proposed by Ian Goodfellow in 2014, Generative Adversarial Networks (GANs) provide an unsupervised form of machine learning where first "generator network" produces candidate objects, e.g. images, and a second "discriminator network" evaluates the objects produced by the generator – classifying them as real or fake. The goal of a GAN is to produce realistic objects that "fool" the discriminator into classifying them as real.

Typically, GANs sample a latent variable vector, z, from a probability distribution, i.e. Gaussian, and transforms this vector into the desired object. This way one can generate new objects from the network using random latent space vectors. To produce realistic objects from random noise, GANs must learn some relationship between the inputted vector and the desired object space. This hints at the fact that there might be some causal structure captured by the GAN. Leveraging this intuition, I hope to explore whether GANs may learn to encode causal models. In terms of prior work, there have been recently published papers on Causal GANs by Kocaoglu et. al. that I hope to explore for potential insights during the project.

To investigate the idea of a GAN encoding causal structure, I first need a dataset that demonstrates causal structure. I plan on selecting a simple Bayesian network that will act as the causal model that we hope to learn. I'll configure this network in two ways – once as a series of Booleans, i.e. factors that have probabilities of either 0 or 1, and secondly as a series of continuous probabilities to construct two learning problems of different difficulty. I will then code this up in WebPPL and run these models forward to generate a dataset of true objects from the underlying distribution. Next, I will then train a GAN on this dataset and experiment with the dimensionality of the latent space vector, z, to see how that impacts generation of outputs.

Finally, I need a metric by which to evaluate the GAN's ability to learn the causal model. For this, I turn towards the work of Lucas and Kemp in the space of counterfactual reasoning. Humans often reason about outcomes in "reverse", i.e. they think about "what could have been" if a different world existed. For example, one might say "If I were tall, I might play basketball". Here, we consider a state that might fall out, given a different world, i.e. "being tall". Thinking about these "what ifs" is the essence of counterfactual reasoning. By perturbing the existing world, we can pose a series of conditions. Observing the results of these conditions, provides insight into the distribution of possibilities, given alternate worlds.

I hope to leverage this idea from counterfactual reasoning, of conditioning on perturbed worlds and observing their outcomes, to test whether the trained GAN has learnt the underlying causal structure of the data. I'll first present some latent space vector, z, to the GAN and observe the output. Z provides the true "world". I shall then perturb this noise vector and observe the outputs that are generated to examine the distributions of counterfactual statements that can made from the initial z. I can then convert the Bayesian network utilized to generate the dataset to an exogenous form, e.g. a flexible control model (FCM), and reason over these counterfactuals independently – to compare the distribution provided by the GAN to the "true" distribution on counterfactuals.