CAUSAL GAN: LEARNING CAUSAL IMPLICIT GENERATIVE MODELS WITH ADVERSARIAL TRAINING

(Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis & Sriram Vishwanath, 2017)

Summer Term 2018

Created for the Seminar "Explainable Machine Learning"

Docent: PD Dr. rer. nat., Dipl. phys. Ulrich Köthe

Presenter: Stefan Radev

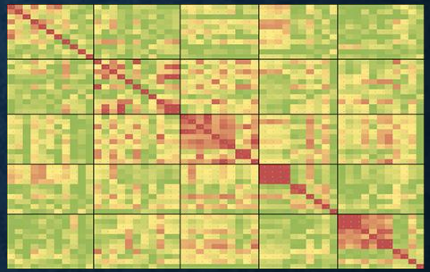
Presented on: 19.07.2018

OUTLINE

1. An introduction to causal inference
2. Causal implicit generative models (CIGMs)
3. Causal GAN: architecture and components
4. Results
5. Discussion

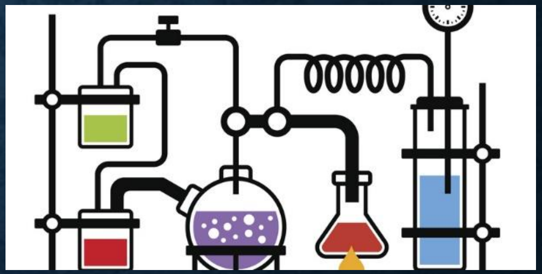
CAUSAL INFERENCE

- Association vs. Causation (Pearl, 2010)
  - **Standard statistical analysis** – infer parameters of a distribution from from finite samples; discover associations between parameters, for instance via:
    - Correlation
    - Regression
    - Conditional independence
    - **Virtually any method that relies on a joint distribution of observed variables**

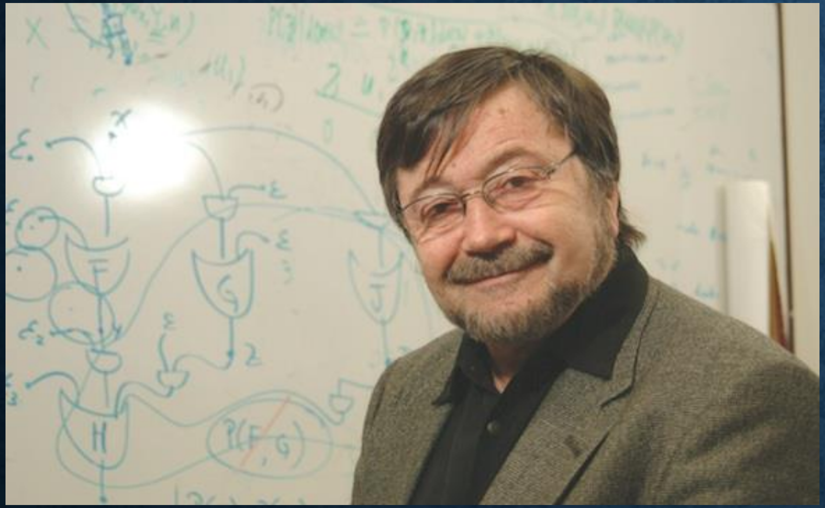


CAUSAL INFERENCE

- Association vs. Causation (Pearl, 2010)
  - **Causal analysis** – infer probabilities under changing conditions; discover changes in distributions due to external influences, for instance via:
    - Randomization
    - "Holding constant"
    - Intervention
    - **Virtually any method that *<u>does not</u>*** rely **on the distribution of observed variables alone**



CAUSAL INFERENCE



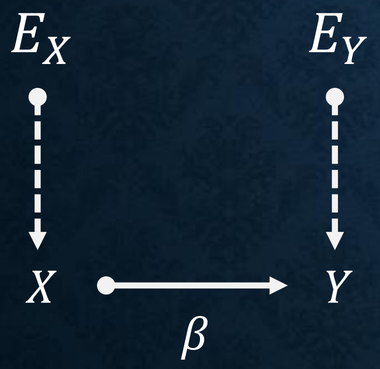
Judea Pearl (b. 1936)

"This distinction further implies that causal relations cannot be expressed in the language of probability and, hence, that any mathematical approach to causal analysis must acquire new notation – probability calculus is insufficient." (Pearl, 2010, p.2).

CAUSAL INFERENCE

- Representing linear causation (Pearl, 2010)

Path diagram



Linear structural equation

$$x = e_X$$

$$y = \beta x + e_y$$

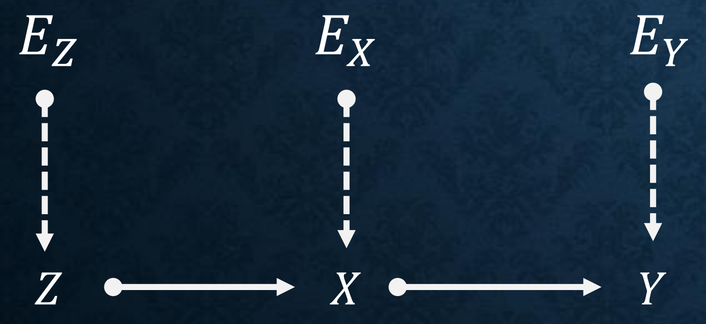
Interpretation

- $\{E_X, E_y\}$: *exogenous* variables or *errors* (unexplained factors)
- $\{X, Y\}$: *endogenous* variables (variables of interest)
- $X \to Y$: causal hypothesis, X (possibly) causes Y
- $\beta = \mathrm{Cov}(\mathrm{X}, \mathrm{Y})$: *path coefficient*, quantifies the causal effect of X on Y

CAUSAL INFERENCE

- Beyond linear models – redefining the notion of *effect* as a general way of transmitting changes between variables

Path diagram



Structural equation

$$z = f_Z(e_Z)$$

$$x = f_X(z, e_X)$$

$$y = f_Y(x, e_Y)$$

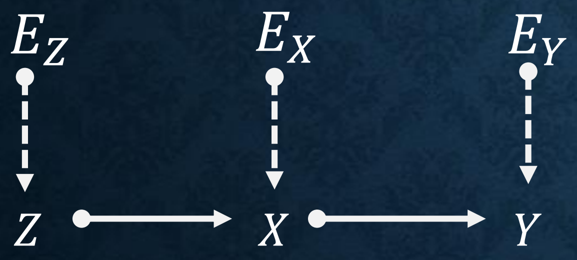
Interpretation

Each function represents a causal process that determines the value of the LHS variable (output) from the RHS variables (input).

CAUSAL INFERENCE

- Representing interventions via the $do(\cdot)$ operator

Path diagram



Structural equation

$$z = f_Z(e_Z)$$

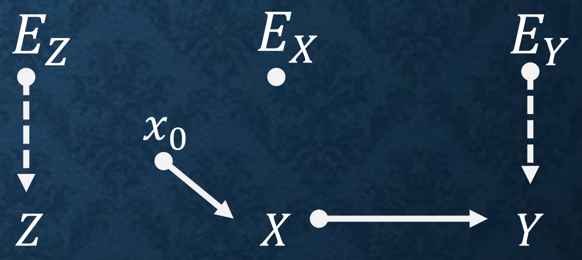
$$x = f_X(z, e_X)$$

$$y = f_Y(x, e_Y)$$

$\Rightarrow do(x_0) \Rightarrow$

(Delete a function from the model by replacing it with a constant)

Modified path diagram



Modified structural equation

$$z = f_Z(e_Z)$$

$$x = x_0$$

$$y = f_Y(x, e_Y)$$

CAUSAL INFERENCE

- Pre-intervention vs. post-intervention distribution
  - $P(x, y)$ – initial pre-intervention distribution
  - $P(y|do(x))$ – post-intervention distribution after modification of the original model
  - Central question in causal analysis (*Identifiability*): **Can the post-intervention distribution be estimated from data generated by the pre-intervention distribution?**

IMPLICIT GENERATIVE MODELS

- **Implicit generative models (IGM):** Implement a mechanism to sample from a (complicated) probability distribution without an explicit parametrization (e.g. GANs)
- **GAN**: Implement the sampling process via forward computation given random noise vectors
- **cGAN:** Extend GANs by feeding class labels to the generator along random noise vectors

CAUSAL IMPLICIT GENERATIVE MODELS

- Previous cGAN architectures **do not** capture dependencies between labels
- Idea of causal IGMs:
  I. Capture **dependencies** between labels
  II. Consider **causal effects** between labels
- Abstractly, model conditional generation as a causal process $L \rightarrow I$

CAUSAL IMPLICIT GENERATIVE MODELS

- Intervention vs. Conditioning (*Gender, Mustache, Image*)

(b) Top: Intervened on Mustache=1. Bottom: Conditioned on Mustache = 1. $Male \rightarrow Mustache$.

- Note: $P(Image, Gender|Mustache = true) \neq P(Image, Gender|do(Mustache = true))$

CAUSAL IMPLICIT GENERATIVE MODELS

- **Aside on assumptions:**

$$P_{data}(Gender = female | Mustache = true) \neq 0$$

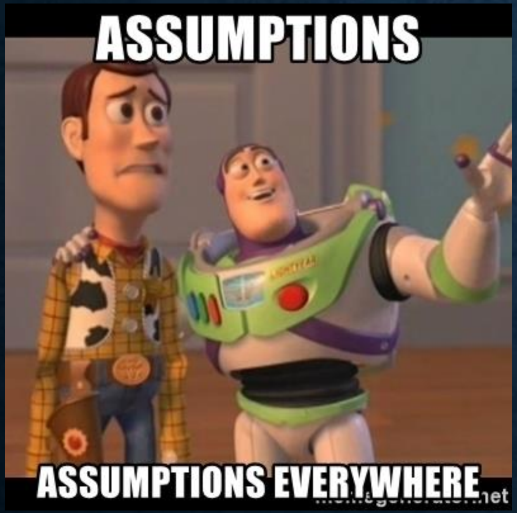
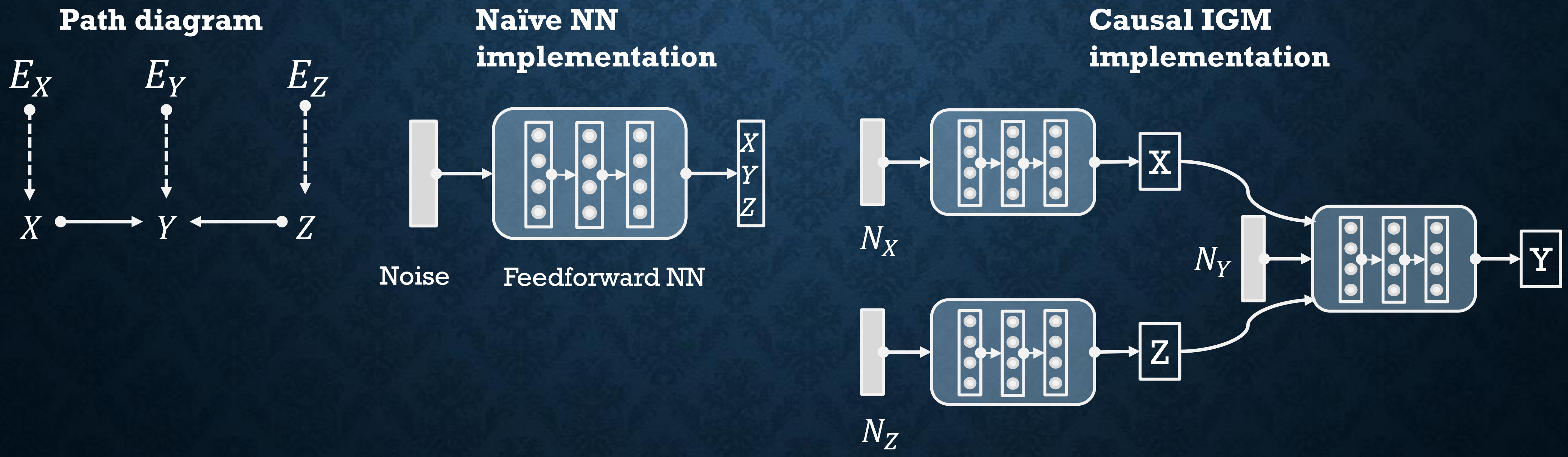


Image taken from: http://www.conchitawurst.com/

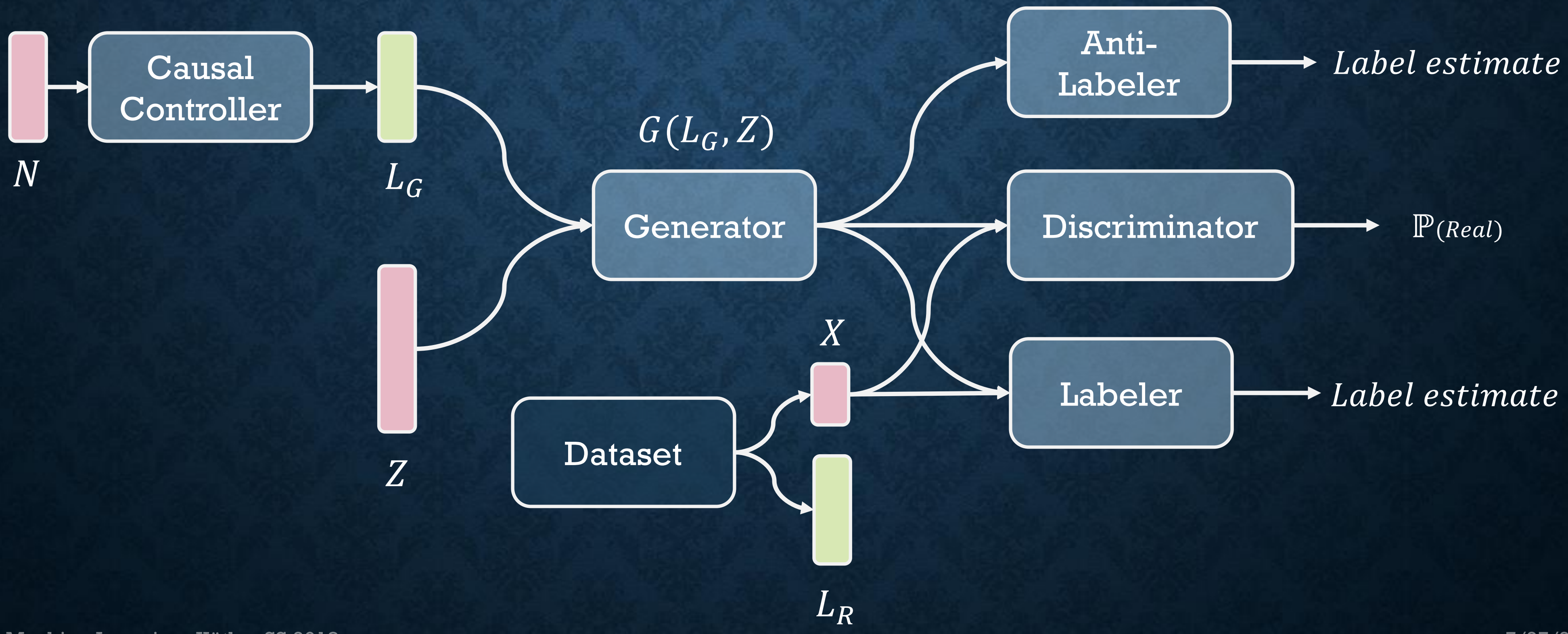
CAUSAL IMPLICIT GENERATIVE MODELS

- Representing causal structural equations via neural networks



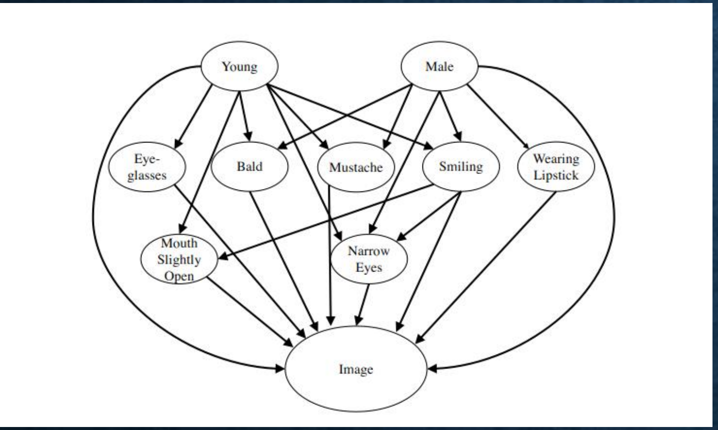
CAUSAL GAN ARCHITECTURE

- Full architecture



CAUSAL GAN ARCHITECTURE

1. Causal controller (WGAN)
   - Produces labels according to a causal graph



Note: Each mapping from parents to children is a neural network.

   - Controls which distribution the images will be sampled from (conditional or interventional)

CAUSAL GAN ARCHITECTURE

- Aside on Wasserstein GAN
  - No log in the loss
  - Clip the weight of *D*
  - Train *D* more than *G*
  - Use RMSProp instead of ADAM
  - Lower learning rate
- Aim: stabilize GAN training!

Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

Require: : α, the learning rate. c, the clipping parameter. m, the batch size. n_{critic}, the number of iterations of the critic per generator iteration.

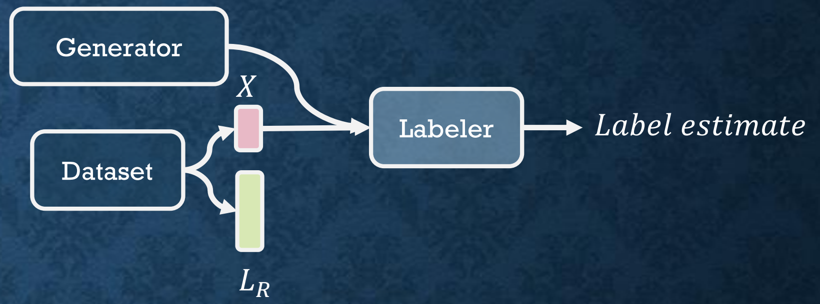
Require: : w_0, initial critic parameters. θ_0, initial generator's parameters.

while θ has not converged **do**
  for $t = 0, ..., n_{\text{critic}}$ **do**
    Sample $\{x^{(i)}\}_{i=1}^m \sim \mathbb{P}_r$ a batch from the real data.
    Sample $\{z^{(i)}\}_{i=1}^m \sim p(z)$ a batch of prior samples.
    $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^m f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^m f_w(g_\theta(z^{(i)}))\right]$
    $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$
    $w \leftarrow \text{clip}(w, -c, c)$
  end for
  Sample $\{z^{(i)}\}_{i=1}^m \sim p(z)$ a batch of prior samples.
  $g_\theta \leftarrow -\nabla_\theta \frac{1}{m}\sum_{i=1}^m f_w(g_\theta(z^{(i)}))$
  $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$
end while

Arjovsky et al., (2017), p. 8

CAUSAL GAN ARCHITECTURE

2. Labeler



- Trained to estimate the labels of images in the dataset
- Optimization criterion (single binary label l):

$$\max_{D_{LR}} \rho \mathbb{E}_{x \sim \mathbb{P}_r(x|l=1)} [\log (D_{LR}(x))] + (1-\rho) \mathbb{E}_{x \sim \mathbb{P}_r(x|l=0)} [\log(1 - D_{LR}(x))]$$

($\mathbb{P}_r$ - data distribution; ρ - label prior; $D_{LR}(x)$ - mapping due to labeler)

CAUSAL GAN ARCHITECTURE

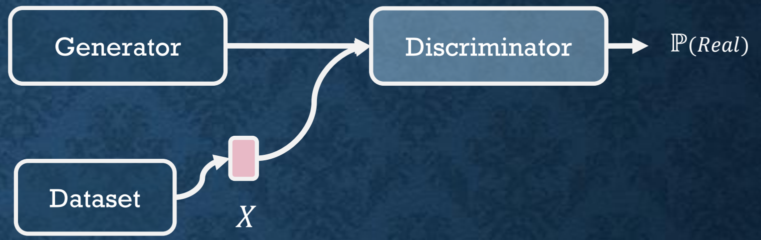
3. Anti-Labeler

- Trained to estimate the labels of images sampled from generator
- Optimization criterion (single binary label l):

$$\max_{D_{LR}} \rho \mathbb{E}_{x\sim\mathbb{P}_g(x|l=1)} [\log (D_{LG}(x))] + (1-\rho)\mathbb{E}_{x\sim\mathbb{P}_g(x|l=0)}[\log(1 \ - D_{LG}(x))]$$

($\mathbb{P}_g$ - generator induced distribution; ρ - label prior; $D_{LG}(x)$ - mapping due to anti-labeler)

CAUSAL GAN ARCHITECTURE



4. Discriminator

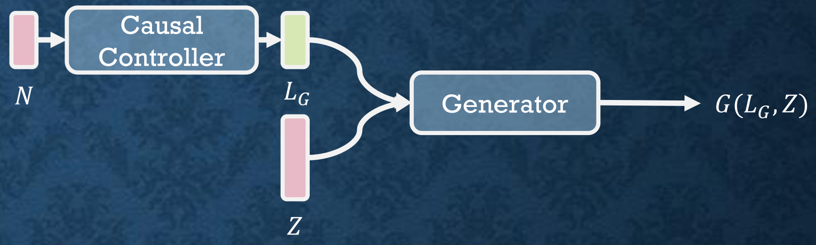
- Trained to discriminate between real and fake images
- Optimization criterion (single binary label l):

$$\max_D \mathbb{E}_{(l,x)\sim\mathbb{P}_r(l,x)} [\log (D(x))] + \mathbb{E}_{(l,x)\sim\mathbb{P}_g(l,x)}[\log(1 - D(x))]$$

($\mathbb{P}_r$ - data distribution; $\mathbb{P}_g$ - generator induced distribution; ρ - label prior; $D_{LG}(x)$ - mapping due to discriminator)

CAUSAL GAN ARCHITECTURE

5. Generator



- Optimization criterion (single binary label l):

$$\min_G \mathbb{E}_{(l,x)\sim\mathbb{P}_g(l,x)}\left[\log\left(\frac{1-D(x)}{D(x)}\right)\right]$$

Maximize discriminator loss

$$-\rho\mathbb{E}_{x\sim\mathbb{P}_g}[\log(D_{LR}(x))]-(1-\rho)\mathbb{E}_{x\sim\mathbb{P}_g}\left[\log\left(1-D_{LR}(x)\right)\right]$$

Minimize labeler loss

$$+\rho\mathbb{E}_{x\sim\mathbb{P}_g}[\log(D_{LG}(x))]+(1-\rho)\mathbb{E}_{x\sim\mathbb{P}_g}[\log(1-D_{LG}(x))]$$

Maximize anti-labeler loss

CAUSAL GAN ARCHITECTURE

6. Theoretical guarantees
   - *Given a perfect causal controller, as well as an optimal labeler, anti-labeler, and discriminator, the global minimum of the generator loss is achieved iff* $\mathbb{P}_r(l, x) = \mathbb{P}_g(l, x)$, *i.e. iff* $\mathbb{P}\big(G(l, z)\big) = \mathbb{P}_r(x|l))$
   - Proof sketch: *substitute expressions for optimal labelers and discriminator into generator objective and show that it yields* $KL(\mathbb{P}_g \,||\, \mathbb{P}_r)$

RESULTS

- Train causal GAN on the CelebA data set in a two stage procedure
  1. Train the **Causal Controller** on the image labels
  2. Train the **Causal GAN** on the images conditioned on the labels from the causal controller

RESULTS

1. Convergence of Causal Controller to the true marginal distributions of the labels

Label, L	$\mathbb{P}_{G1}(L=1)$	$\mathbb{P}_{cG1}(L=1)$	$\mathbb{P}_{D}(L=1)$
Bald	0.02244	0.02328	0.02244
Eyeglasses	0.06180	0.05801	0.06406
Male	0.38446	0.41938	0.41675
Mouth Slightly Open	0.49476	0.49413	0.48343
Mustache	0.04596	0.04231	0.04154
Narrow Eyes	0.12329	0.11458	0.11515
Smiling	0.48766	0.48730	0.48208
Wearing Lipstick	0.48111	0.46789	0.47243
Young	0.76737	0.77663	0.77362

Table 2: Marginal distribution of pretrained Causal Controller labels when Causal Controller is trained on CelebA Causal Graph (P_{G1}) and its completion(P_{cG1}), where $cG1$ is the (nonunique) largest DAG containing $G1$ (see appendix). The third column lists the actual marginal distributions in the dataset

RESULTS

2. CausalGAN: Sampling from the conditional/interventional dsitributions

Top: Intervene Mustache=1, Bottom: Condition Mustache=1

Figure 4: Intervening/Conditioning on Mustache label in CelebA Causal Graph with CausalGAN. Since $Male \rightarrow Mustache$ in CelebA Causal Graph, we do not expect $do(Mustache = 1)$ to affect the probability of $Male = 1$, i.e., $\mathbb{P}(Male = 1 | do(Mustache = 1)) = \mathbb{P}(Male = 1) = 0.42$. Accordingly, the top row shows both males and females with mustaches, even though the generator never sees the label combination $\{Male = 0, Mustache = 1\}$ during training. The bottom row of images sampled from the conditional distribution $\mathbb{P}(.|Mustache = 1)$ shows only male images.

RESULTS

2. CausalGAN: Sampling from the conditional/interventional distributions (2)

Top: Intervene Mouth Slightly Open=1, Bottom: Condition Mouth Slightly Open=1

Figure 5: Intervening/Conditioning on Mouth Slightly Open label in CelebA Causal Graph with CausalGAN. Since $Smiling \rightarrow MouthSlightlyOpen$ in CelebA Causal Graph, we do not expect $do(\textit{Mouth Slightly Open} = 1)$ to affect the probability of $Smiling = 1$, i.e., $\mathbb{P}(Smiling = 1|do(\textit{Mouth Slightly Open} = 1)) = \mathbb{P}(Smiling = 1) = 0.48$. However on the bottom row, conditioning on *Mouth Slightly Open* $= 1$ increases the proportion of smiling images (From 0.48 to 0.76 in the dataset), although 10 images may not be enough to show this difference statistically.

RESULTS

2. CausalGAN: Sampling from the conditional/interventional distributions (2)

Top: Intervene Mouth Slightly Open=1, Bottom: Condition Mouth Slightly Open=1

Figure 5: Intervening/Conditioning on Mouth Slightly Open label in CelebA Causal Graph with CausalGAN. Since $Smiling \rightarrow MouthSlightlyOpen$ in CelebA Causal Graph, we do not expect $do(\textit{Mouth Slightly Open} = 1)$ to affect the probability of $Smiling = 1$, i.e., $\mathbb{P}(Smiling = 1 | do(\textit{Mouth Slightly Open} = 1)) = \mathbb{P}(Smiling = 1) = 0.48$. However on the bottom row, conditioning on $\textit{Mouth Slightly Open} = 1$ increases the proportion of smiling images (From 0.48 to 0.76 in the dataset), although 10 images may not be enough to show this difference statistically.

RESULTS

2. CausalGAN: Sampling from the conditional/interventional distributions (3)

Intervening vs Conditioning on Wearing Lipstick, Top: Intervene Wearing Lipstick=1, Bottom: Condition Wearing Lipstick=1

Figure 12: Intervening/Conditioning on Wearing Lipstick label in CelebA Causal Graph. Since $Male \rightarrow WearingLipstick$ in CelebA Causal Graph, we do not expect $do(\textit{Wearing Lipstick} = 1)$ to affect the probability of $Male = 1$, i.e., $\mathbb{P}(Male = 1 | do(\textit{Wearing Lipstick} = 1)) = \mathbb{P}(Male = 1) = 0.42$. Accordingly, the top row shows both males and females who are wearing lipstick. However, the bottom row of images sampled from the conditional distribution $\mathbb{P}(.|\textit{Wearing Lipstick} = 1)$ shows only female images because in the dataset $\mathbb{P}(Male = 0 | \textit{Wearing Lipstick} = 1) \approx 1$.

RESULTS

2. CausalGAN: Diversity

Figure 17: Diversity of the proposed CausalGAN showcased with 256 samples.

SUMMARY AND DISCUSSION

1. Causal GANs allow us to obtain samples with desired properties that may not be present in the training set
2. Causal GANs assume the causal graph structure but **learn the functions of the structural equations**
3. What are the advantages of causal GANs over Bayesian networks?
4. Do causal GANs offer the possibility for simulating "real" science experiments?

THE END

Thank you!

REFERENCES

- Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.
- Kocaoglu, M., Snyder, C., Dimakis, A. G., & Vishwanath, S. (2017). CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. *arXiv preprint arXiv:1709.02023*.
- Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics surveys*, *3*, 96-146.
- Pearl, J. (2010). Causal inference. In *Causality: Objectives and Assessment* (pp. 39-58).
- Pearl J. (2010). An introduction to causal inference. *The international journal of biostatistics, 6*, 1-59.
- https://github.com/mkocaoglu/CausalGAN