Continual Causality: AAAI 2023 Bridge Program

Memory, Invariance and Reasoning

Pillars of the Causal-Continual Bridge

7 Feb 2023

Vineeth N Balasubramanian

Visiting Faculty Fellow, Carnegie Mellon University

Department of Computer Science and Engineering/Artificial Intelligence

Indian Institute of Technology, Hyderabad

आई आई टी हैदराबाद

IIT Hyderabad

Causality and Continual Learning

Talk Focus

Key pillars of the bridge

Memory:

A key component of successful CL methods, where does causal come in?

Invariance:

Often an implicit need of CL, not modeled explicitly. Causal principles naturally well-suited.

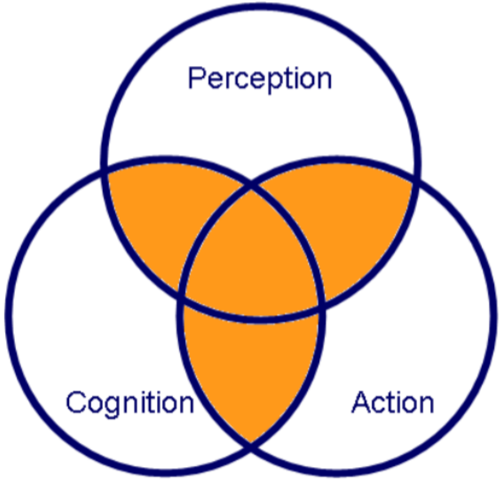
Reasoning:

Essential for long-term large-scale CL. How to bring causal perspectives?

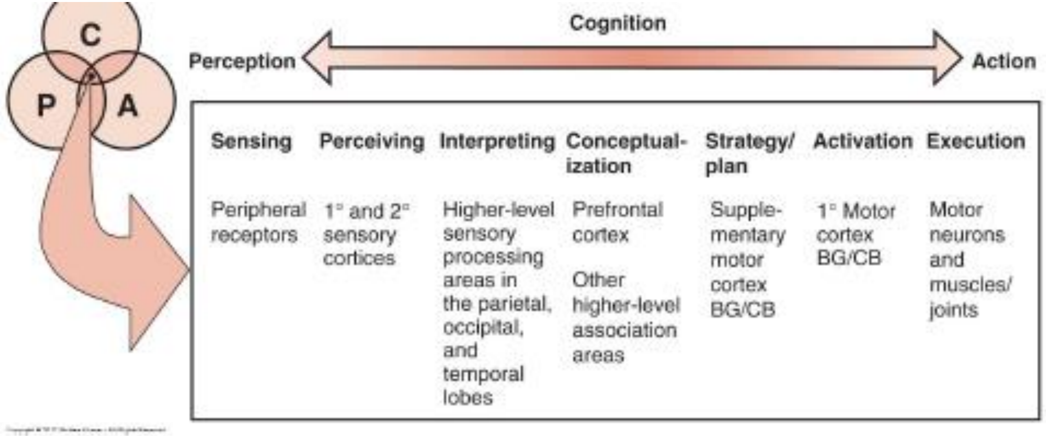
PCA

Information Processing in the Human Brain

Embodied view of the mind



Physiological View



Hurtienne, Cognition In HCI: An Ongoing Story, 2009

Going Beyond Statistical Learning

The Need

Machine Learning in the 1990s	Machine learning now

Machine Learning in the 1990s

- Training set carefully curated to cover all cases of interest
- Actual deployments (e.g. ATT-Lucent-NCR check reading machines with CNNs)

Machine learning now

- Datasets are too big to be carefully curated
- Data collection biases, confounding biases, feedback loops, …
- Machine learning algorithms recklessly take advantage of spurious correlations

Need for reasoning, robustness, Type 2 systems, causality

Leon Bottou, ICLR 2019 Keynote

Spurious Correlations

- Susceptibility to adversarial attacks?
- Lack of human-relatable explanations of model predictions
- Poor out-of-distribution generalization
- Bias in the model
- …

→ Spurious Correlations

Need to capture causal relationships

Causality and Continual Learning

Our Work

Causality in Deep Learning

- Matching Learned Causal Effects of Neural Networks with Domain Priors, **ICML 2022**
- On Causally Disentangled Representations, **AAAI 2022**
- Evaluating and Mitigating Bias in Image Classifiers: A Causal Perspective Using Counterfactuals, **WACV 2022**
- Neural Network Attributions: A Causal Perspective, **ICML 2019**

Talk Focus

Continual Learning

- Energy-based Latent Aligner for Incremental Learning, **CVPR 2022**
- Unseen Classes at a Later Time? No Problem, **CVPR 2022**
- Novel Class Discovery without Forgetting, **ECCV 2022**
- Incremental Object Detection via Meta-Learning, **TPAMI 2021**
- Towards Open-World Object Detection, **CVPR 2021**
- Meta-consolidation for Continual Learning, **NeurIPS 2020**

Causality and Continual Learning

Talk Focus

Key pillars of the bridge

Memory:

A key component of successful CL methods, where does causal come in?

Replay buffers. Can they be causal?

Invariance:

Often an implicit need of CL, not modeled explicitly. Causal principles naturally well-suited.

Stability-plasticity trade-off

Reasoning:

Essential for long-term large-scale CL. How to bring causal perspectives?

Prediction-by-reasoning

The “Invariance” Pillar

- Fundamental premise of causality
  - The more invariant a relationship between two variables, the more the relationship should be considered causal
- Implications for CL
  - From task-discriminative to task-invariant representation learning
  - Separating domain-invariant from domain-specific features in domain-incremental learning
  - Learning task/domain-agnostic (or even task/domain-specific) independent mechanisms
  - Core issue: *Disentanglement*

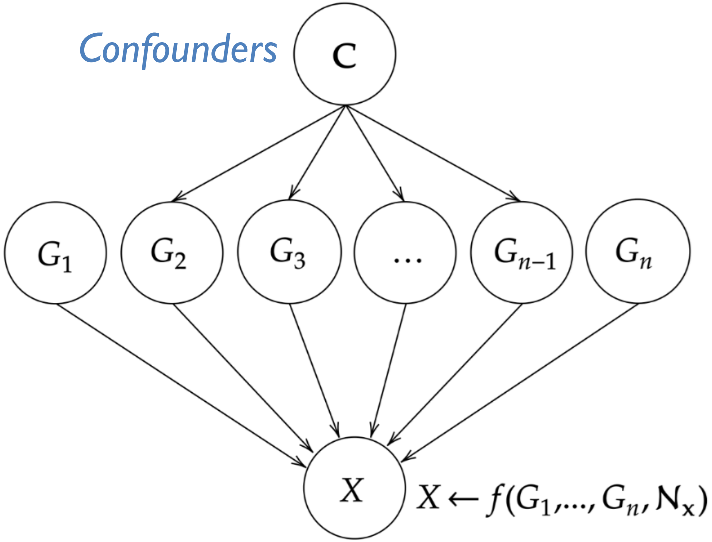
Causal Disentanglement

Our Work

Disentangled Causal Process

Confounders

Generative Factors



Causal model for X is disentangled *(iff)* it can be described by the SCM:

$$C_j \leftarrow \mathcal{N}_{C_j}; j \in \{1,\ldots,l\}$$
$$G_i \leftarrow g_i(PA_i^C, \mathcal{N}_{G_i}); i \in \{1,\ldots,n\}$$
$$X \leftarrow f(G_1,\ldots,G_n,\mathcal{N}_x)$$

f, g_i are independent causal mechanisms

Reddy, Godfrey, Balasubramanian, On Causally Disentangled Representations, AAAI 2022
Suter et al, Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness, ICML 2019

Evaluating Causal Disentanglement

Can Latent Variable Models (LVMs) learn to *causally* disentangle?

Metric 1: Unconfoundedness

- Encoder e of a LVM $\mathcal{M}$ (e, g, p_X) should learn the *mapping* from G_i to $\mathbf{Z}_I$ without any influence from C.

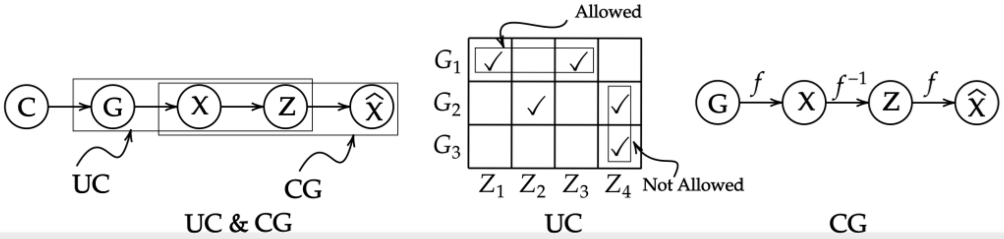
$$UC := 1 - \mathbb{E}_{x \sim p_X}\left[\frac{1}{S}\sum_{I,J}\frac{|\mathbf{Z}_I^x \cap \mathbf{Z}_J^x|}{|\mathbf{Z}_I^x \cup \mathbf{Z}_J^x|}\right]$$

Metric 2: Counterfactual Generativeness

- If $\mathbf{Z}$ is unconfounded, the counterfactual of x w.r.t. G_i, x_I^{cf} can be generated by intervening on $\mathbf{Z}_I^x$.
- Any change in $\mathbf{Z}_{\setminus I}^x$, should have no influence on x_I^{cf} w.r.t. G_i.

$$CG := \mathbb{E}_I\left[\left|ACE_{\mathbf{Z}_I^X}^{X_I^{cf}} - ACE_{\mathbf{Z}_{\setminus I}^X}^{X_{\setminus I}^{cf}}\right|\right]$$

ACE = Average Causal Effect



Learning Independent Mechanisms

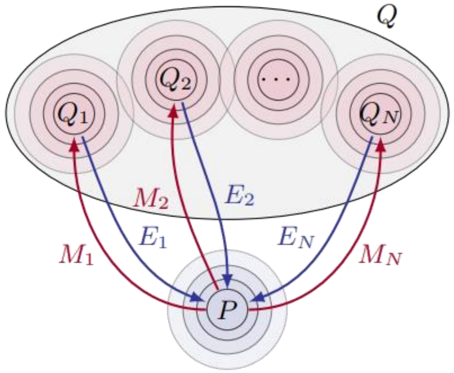


Figure 1. An overview of the problem setup. Given a sample from a canonical distribution P, and one from a mixture of transformed distributions Q_i obtained by mechanisms M_i on P, we want to learn inverse mechanisms E_i as independent modules. Modules (or *experts*) compete amongst each other for data points, encouraging specialization.

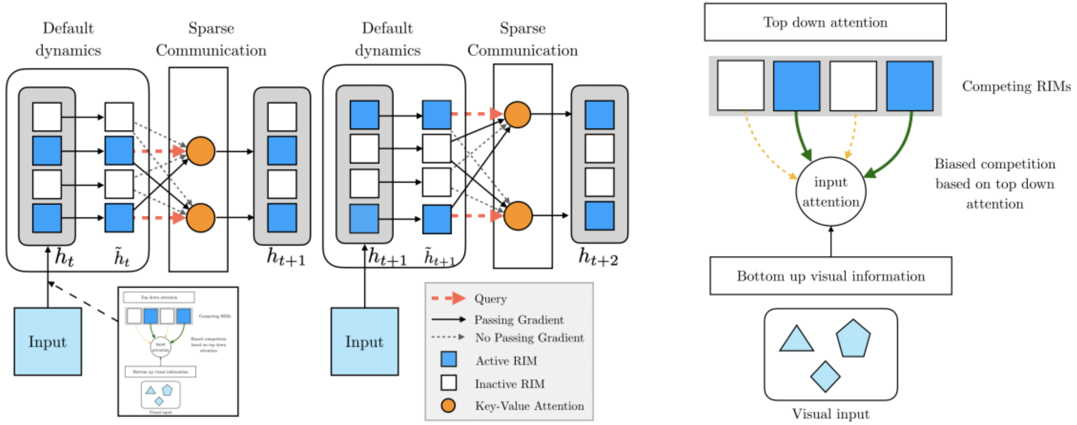


Figure 1: **Illustration of Recurrent Independent Mechanisms (RIMs).** A single step under the proposed model occurs in four stages (left figure shows two steps). In the first stage, individual RIMs produce a query which is used to read from the current input. In the second stage, an attention based competition mechanism is used to select which RIMs to activate (right figure) based on encoded visual input (blue RIMs are active, based on an attention score, white RIMs remain inactive). In the third stage, individual activated RIMs follow their own default transition dynamics while non-activated RIMs remain unchanged. In the fourth stage, the RIMs sparsely communicate information between themselves, also using key-value attention.

Parascandolo et al, Learning Independent Causal Mechanisms, ICML 2018

Goyal et al, Recurrent Independent Mechanisms, ICLR 2021

Generative Models for Independent Mechanisms

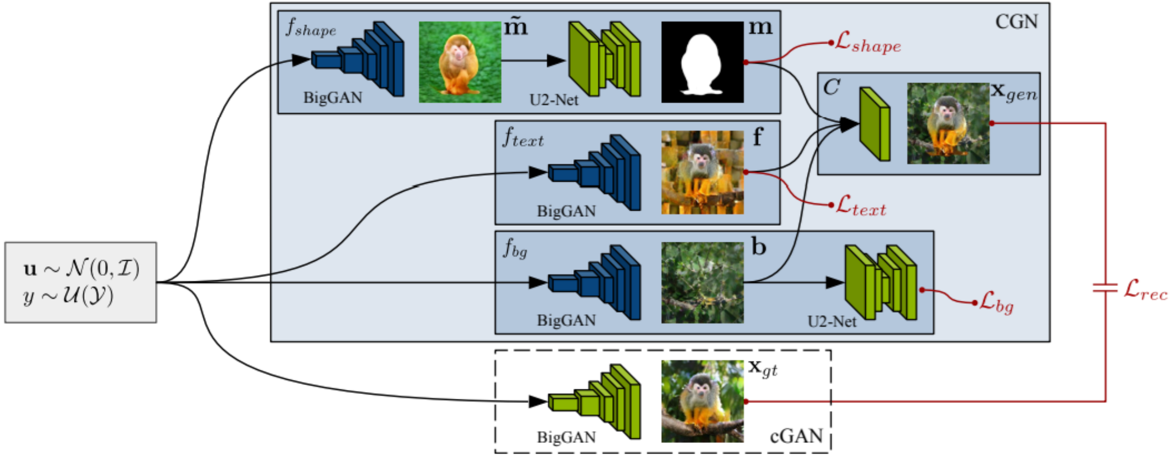
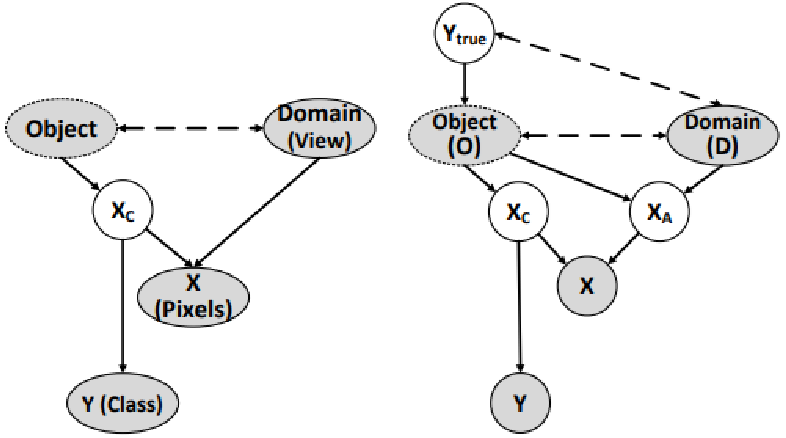


Figure 2: **Counterfactual Generative Network (CGN).** Here, we illustrate the architecture used for the ImageNet experiments. The CGN is split into four mechanisms, the shape mechanism f_{shape}, the texture mechanism f_{text}, the background mechanism f_{bg}, and the composer C. Components with trainable parameters are blue, components with fixed parameters are green. The primary supervision is provided by an unconstrained conditional GAN (cGAN) via the reconstruction loss $\mathcal{L}_{rec}$. The cGAN is only used for training, as indicated by the dotted lines. Each mechanism takes as input the noise vector $\mathbf{u}$ (sampled from a spherical Gaussian) and the label y (drawn uniformly from the set of possible labels $\mathcal{Y}$) and minimizes its respective loss ($\mathcal{L}_{shape}$, $\mathcal{L}_{text}$, and $\mathcal{L}_{bg}$). To generate a set of counterfactual images, we sample $\mathbf{u}$ and then independently sample y for each mechanism.

Sauer & Geiger, Counterfactual Generative Networks, ICLR 2021

Disentanglement in Domain Generalization

Causal View of DG



(a) Image classification. (b) General SCM.

Mahajan et al, Domain Generalization using Causal Matching, ICML 2021

Disentanglement of Domain-Invariant Features in VAEs

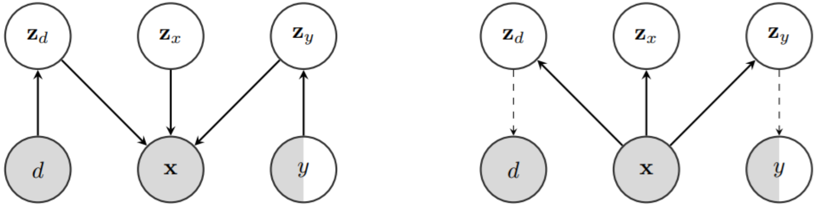


Figure 1: Left: Generative model. According to the graphical model we obtain $p(d, \mathbf{x}, y, \mathbf{z}_d, \mathbf{z}_x, \mathbf{z}_y) = p_\theta(\mathbf{x}|\mathbf{z}_d, \mathbf{z}_x, \mathbf{z}_y)p_{\theta_d}(\mathbf{z}_d|d)p(\mathbf{z}_x)p_{\theta_y}(\mathbf{z}_y|y)p(d)p(y)$. Right: Inference model. We propose to factorize the variational posterior as $q_{\phi_d}(\mathbf{z}_d|\mathbf{x})q_{\phi_x}(\mathbf{z}_x|\mathbf{x})q_{\phi_y}(\mathbf{z}_y|\mathbf{x})$. Dashed arrows represent the two auxiliary classifiers $q_{\omega_d}(d|\mathbf{z}_d)$ and $q_{\omega_y}(y|\mathbf{z}_y)$.

Ilse et al, DIVA: Domain Invariant Variational Autoencoders, MIDL 2020

The "Invariance" Pillar

A Few Takeaways

Invariance:

Often an implicit need of CL, not modeled explicitly. Causal principles naturally well-suited.

Stability-plasticity trade-off

- How to perform CL in terms of independent mechanisms?
- How does one disentangle independent mechanisms effectively?
- What kind of evaluation metrics do we need for such approaches?
- Do such approaches need fundamentally new approaches, or can they be embedded into existing CL methods?

Causality and Continual Learning

Talk Focus

Key pillars of the bridge

Memory:

A key component of successful CL methods, where does causal come in?

Replay buffers. Can they be causal?

Invariance:

Often an implicit need of CL, not modeled explicitly. Causal principles naturally well-suited.

Stability-plasticity trade-off

Reasoning:

Essential for long-term large-scale CL. How to bring causal perspectives?

Prediction-by-reasoning

The “Reasoning” Pillar

- Causality and reasoning
  - Tightly connected, as causal interpretations more important in practice
- Implications for CL
  - Reasoning a human solution for forgetting -- a core issue not been addressed significantly yet in CL as such
  - Concept-based/Ante hoc interpretable models for CL => More likely to generalize well to out-of-distribution samples, and be robust
  - Shift approach to predict-by-reasoning, rather than just discriminative
  - Reasoning in terms of latent variables (e.g. in vision) a challenge

Towards Explainable Deep Learning

Summary of our Efforts

Non-Causal

Post-hoc Explainability

* [WACV 2018] GradCAM++: Generic method for visual explanations for CNN models
* [IEEE Trans on Biometrics 2021] Canonical saliency maps for face recognition/processing models
* [AISTATS 2022] Submodular ensembles of attribution methods

Causal

* [ICML 2019] Causal attributions in neural networks

Explainability in Deep Learning

Intrinsic Interpretability

Non-Causal

* [CVPR 2022] Ante-hoc explainability via concepts
* [CVPR 2022] Transferring concepts in knowledge distillation tasks

Causal

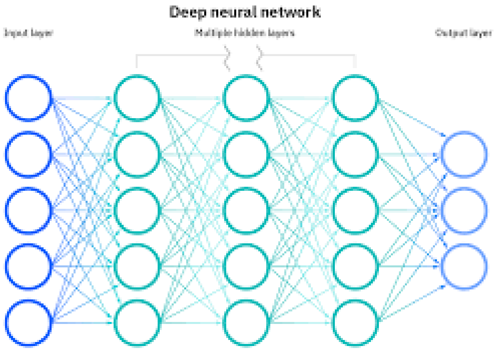
* [AAAI 2022] Causally disentangled representations
* [CVPR'W 2021] Dataset for causal representation learning
* [WACV 2022] Mitigating bias through causal perspectives
* [ICML 2022] Causal regularizers

Complementarity of explanations and robustness [AAAI 2021, NeurIPS 2021]

Causal Perspectives to Explanations in DNNs

Our Work

Consider a trained NN model. Did it learn causal relationships between input and output?



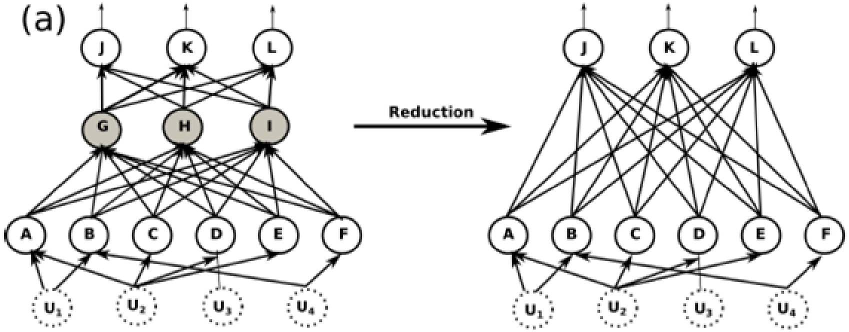
If we had access to prior causal relationships, can we bake them while training NN models?

Causal Attributions in Neural Networks
ICML 2019

Causal Regularization with Domain Priors
ICML 2022

Causal Attributions in DNNs

Neural Network as an SCM



$$M'([l_1, l_n], U, f', P_U)$$

- l_i – neurons in layer l
- f_i – corresponding causal functions

Sarkar et al, Causal Attributions in Neural Networks, ICML 2019

Compute Average Causal Effect of an input variable on output in terms of the NN SCM:

$$ACE^y_{do(x_i=\alpha)} = \mathbb{E}[y|do(x_i = \alpha)] - baseline_{x_i}$$

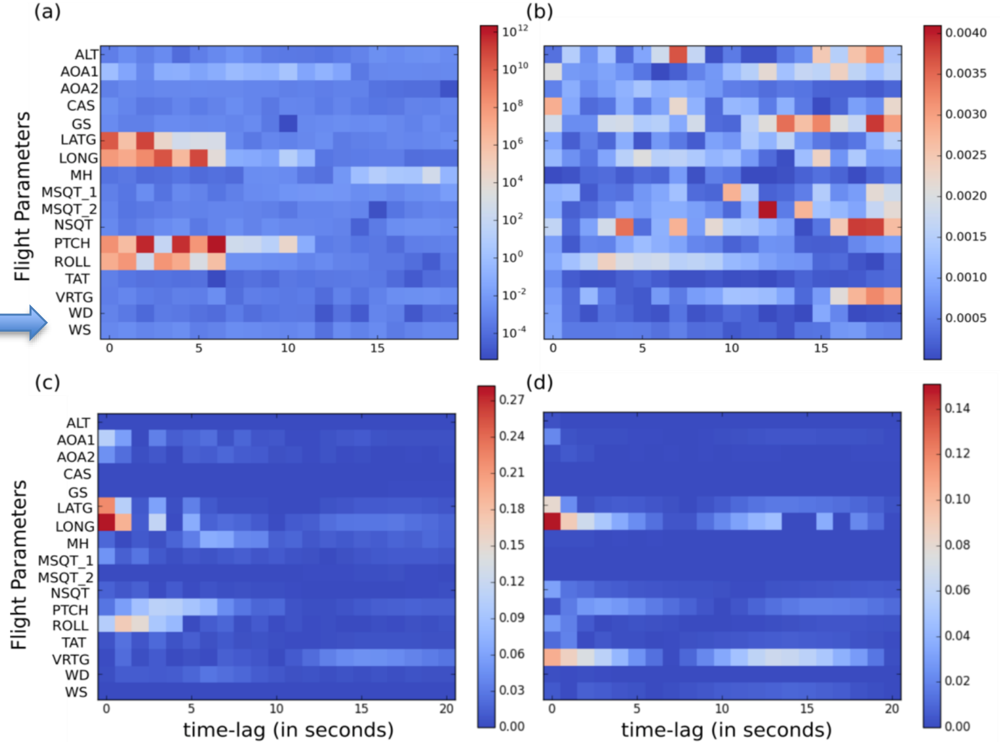
Interventional expectation: Challenging to compute

We propose an efficient solution using numerical linear algebra tricks

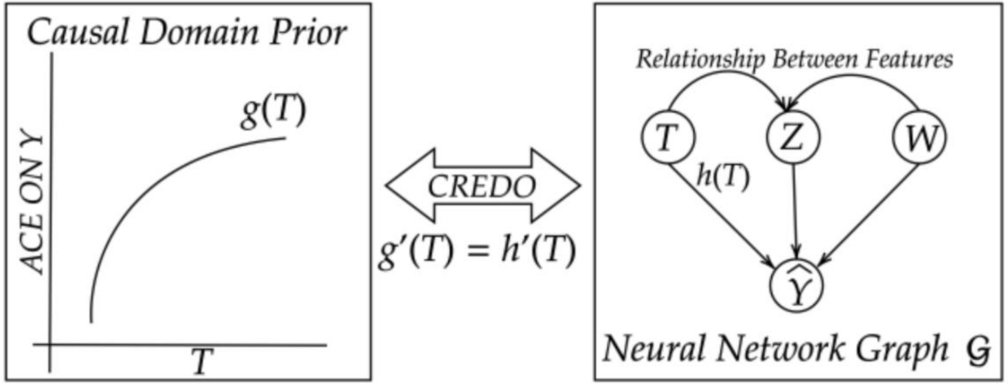
Results

Aircraft Data (NASA Dashlink Dataset)

FDR report: "….*due to slippery runway, the pilot could not apply timely brakes, resulting in a steep acceleration in the airplane post-touchdown*…"



Embedding Causal Knowledge in DNN Models



CREDO: Causal **RE**gularization with **DO**main Priors

We regularize for three kinds of causal effect in NN models:

- Controlled direct effect
- Natural direct effect
- Total causal effect

Reddy et al, Matching Learned Causal Effects of Neural Networks with Domain Priors, ICML 2022

Embedding Causal Knowledge in DNN Models

Proposition

(ACDE Identifiability in Neural Networks) For a neural network with output $\hat{Y}$, the ACDE of a feature T at t on $\hat{Y}$ is identifiable and give by $ACDE_t^{\hat{Y}} = \mathbb{E}_{PA^{\hat{Y}}}[\hat{Y}|t, PA^{\hat{Y}}] - \mathbb{E}_{PA^{\hat{Y}}}[\hat{Y}|t^*, PA^{\hat{Y}}]$.

Proposition

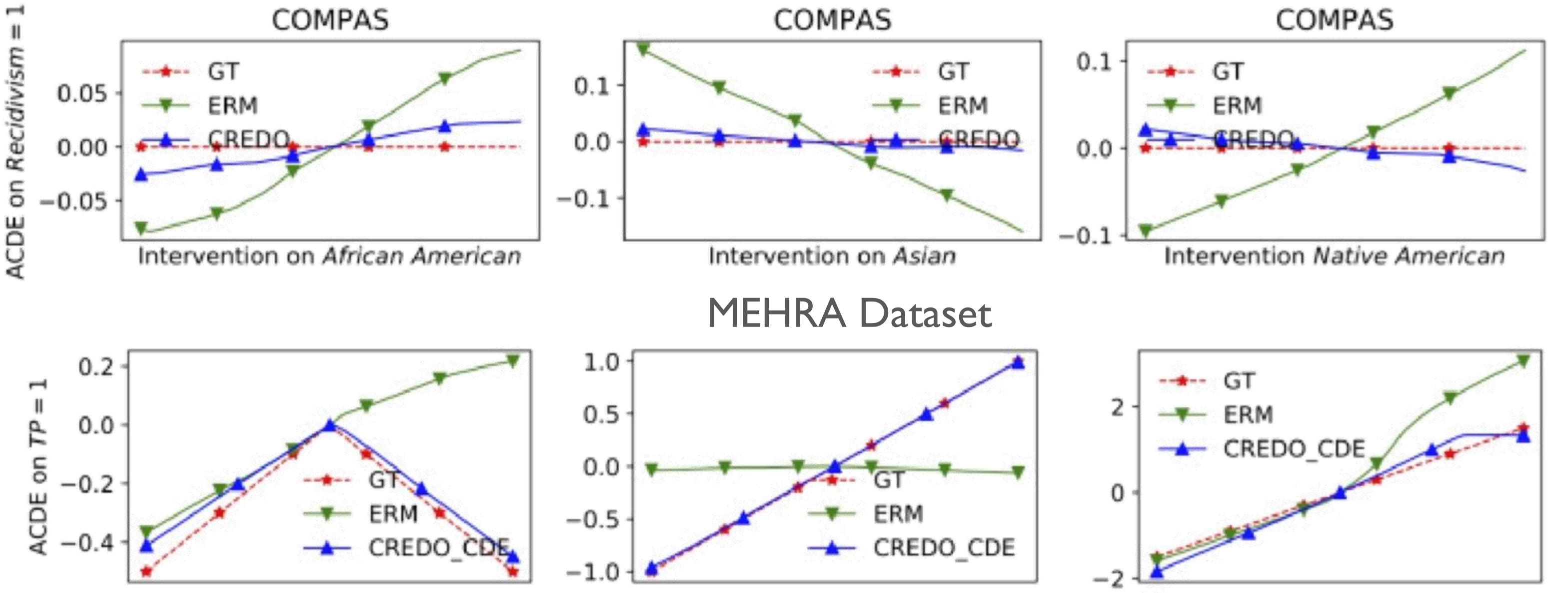
(ACDE Regularization in Neural Networks) The n^{th} partial derivative ACDE of T at t on $\hat{Y}$ is equal to the expected value of n^{th} partial derivative of $\hat{Y}$ w.r.t. T at t, that is: $\frac{\partial^n ACDE_t^{\hat{Y}}}{\partial t^n} = \mathbb{E}_{PA^{\hat{Y}}}\left[\frac{\partial^n [\hat{Y}(t, PA^{\hat{Y}})]}{\partial t^n}\right]$.

Reddy et al, Matching Learned Causal Effects of Neural Networks with Domain Priors, ICML 2022

Algorithm 1 CREDO Regularizer

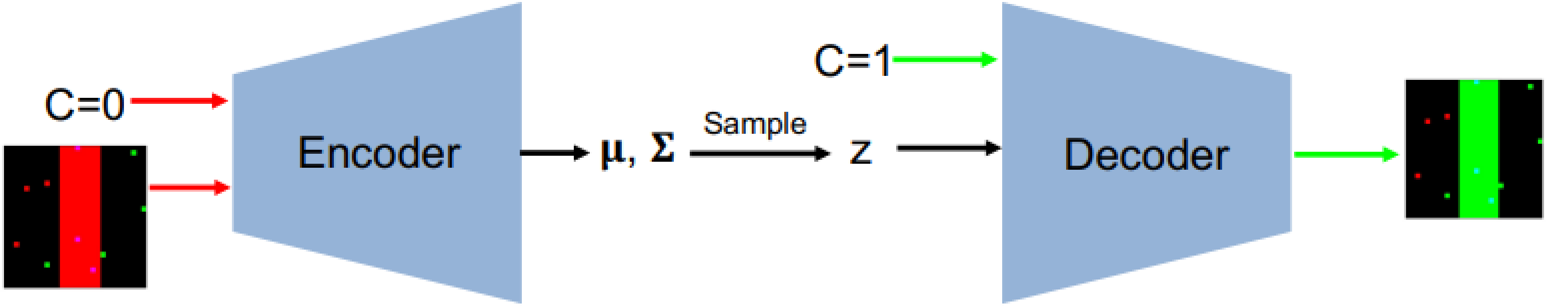
Result: Regularizers for ACDE, ANDE, ATCE in f.
Input: $\mathcal{D} = \{(x^j, y^j)\}_{j=1}^N$, $y^j \in \{0, 1, \ldots, C\}$, $x^j \sim X^j$; $\mathbb{Q} = \{i | \exists\, g_i^c \text{ } for \text{ } some \text{ } c\}$; $\mathbb{G} = \{g_i^c | g_i^c$ is prior for i^{th} feature w.r.t. class c$\}$; $\mathbb{F} = \{f^1, \ldots, f^K\}$ is the set of structural equations of the underlying causal model s.t f^i describes Z^i; ϵ is a hyperparameter
Initialize: $j = 1, \delta G^j = \mathbb{0}_{c \times d} \forall j = 1, \ldots, N$, $M = \mathbb{0}_{c \times d}$
while $j \leq N$ **do**
  foreach $i \in \mathbb{Q}$ **do**
    foreach $g_i^c \in \mathbb{G}$ **do**
      $\delta G^j[c, i] = \nabla g_i^c \big|_{x_i^j}$; $M[c, i] = 1$
      case *1: regularizing ACDE* **do**
        $\nabla_j f[c, i] = \frac{\partial \hat{Y}}{\partial x_i}|_{x^j}$
      case *2: regularizing ANDE* **do**
        `/* causal graph is known */`
        $t = x_i$
        $\nabla_j f[c, i] = \frac{\partial \hat{Y}}{\partial x_i}|_{(t^j, z_{t^*}^j, w^j)}$
      case *3: regularizing ATCE* **do**
        `/* causal graph is known */`
        $\nabla_j f[c, i] = \left[\frac{d\hat{Y}}{dx_i} + \sum_{l=1}^K \frac{\partial \hat{Y}}{\partial Z^l} \frac{df^l}{dx_i}\right]|_{x^j}$
    end
  end
  $j = j + 1$
end
return $\frac{1}{N} \sum_{j=1}^N max\{0, ||\nabla_j f \odot M - \delta G^j||_1 - \epsilon\}$

Sample Results



CREDO shows promising performance in matching causal domain priors with no significant impact on model accuracy/training time

Related Efforts



Goyal et al, Causal Concept Effect, arXiv:1907.07165

The "Reasoning" Pillar

A Few Takeaways

Reasoning:

Essential for long-term large-scale CL. How to bring causal perspectives?

Prediction-by-reasoning

- How to build DL models that inherently reason than discriminate? (Concept-based models, ante hoc interpretable models)
- Explaining/reasoning in terms of latent variables; how?
- What kind of evaluation metrics/benchmarks do we need for reasoning?
- What is the role of memory (esp from a CL perspective) in such a reasoning-based approach?
- Need to go multimodal

Causality and Continual Learning

Talk Focus

Key pillars of the bridge

Memory:

A key component of successful CL methods, where does causal come in?

Replay buffers. Can they be causal?

Invariance:

Often an implicit need of CL, not modeled explicitly. Causal principles naturally well-suited.

Stability-plasticity trade-off

Reasoning:

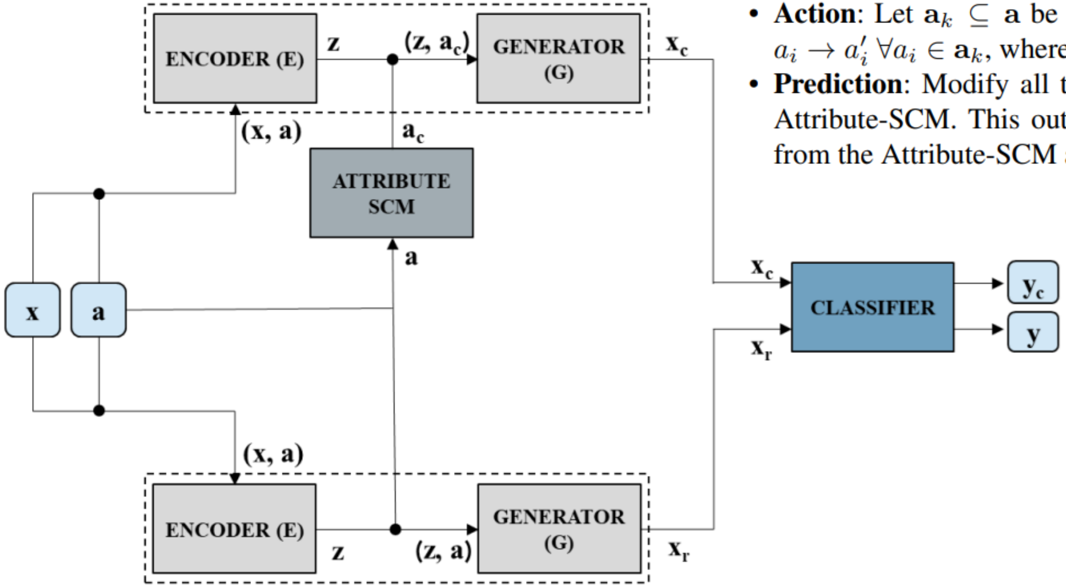
Essential for long-term large-scale CL. How to bring causal perspectives?

Prediction-by-reasoning

The “Memory” Pillar

- Causality and memory
  - Not a direct connection, at least in AI/ML
- Implications for CL
  - Memory very important component of CL methods – how do we make it represent the true causal graph?
  - Use of counterfactuals from a causal perspective in generative replay methods
  - Disentanglement of independent mechanisms is generative models used for CL

Counterfactual Generation



- **Abduction**: Infer latent $\mathbf{z}$ given the input $(\mathbf{x}, \mathbf{a})$ using the encoder.
- **Action**: Let $\mathbf{a}_k \subseteq \mathbf{a}$ be the set of k attributes that one wants to intervene on. Set attribute $a_i \rightarrow a'_i \; \forall a_i \in \mathbf{a}_k$, where $\mathbf{a}'_k = \{a'_i\}_{i=1}^{k}$.
- **Prediction**: Modify all the descendants of $\mathbf{a}_k$ according to the SCM equations learned by Attribute-SCM. This outputs $\mathbf{a}_c$, the intervened attributes. Use $\mathbf{z}$ from the encoder and $\mathbf{a}_c$ from the Attribute-SCM and input it to the generator to obtain the counterfactual x_c.

Dash et al, Evaluating and Mitigating Bias in Image Classifiers: A Causal Perspective Using Counterfactuals, WACV 2022

Counterfactual Generation

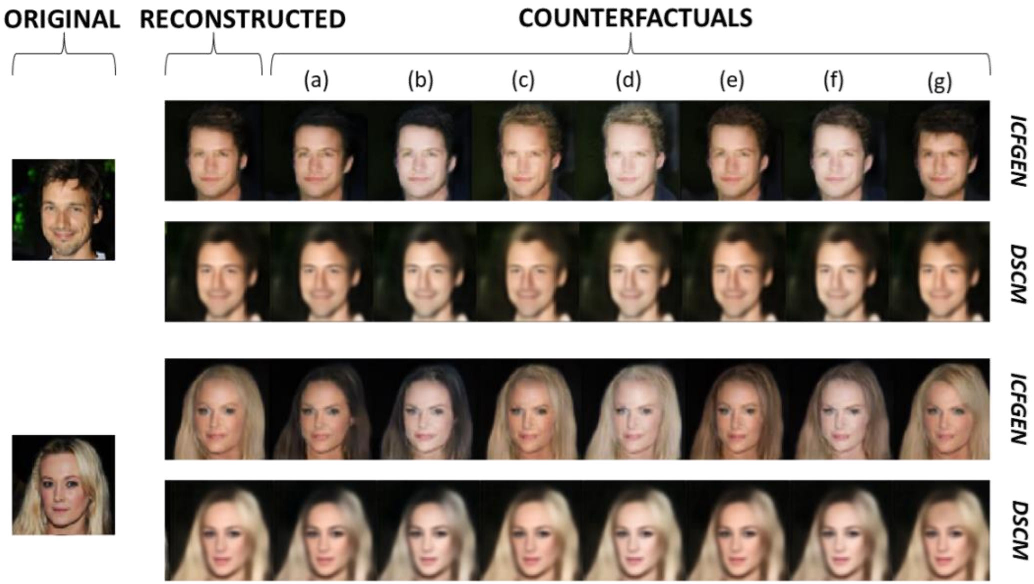
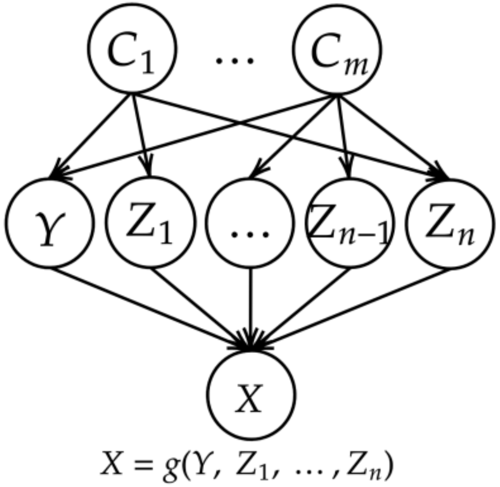


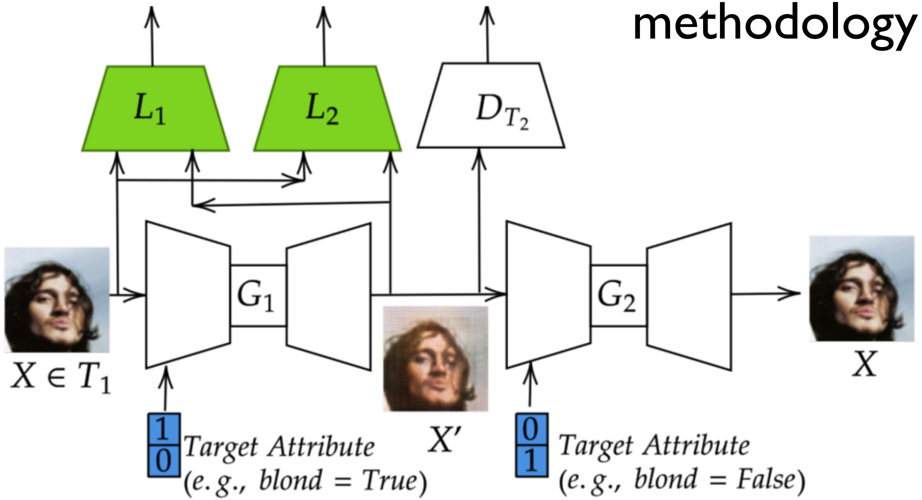
Figure 5: ***ImageCFGen* and DeepSCM Counterfactuals.** (a) denotes *do* (black hair = 1) and (b) denotes *do* (black hair = 1, pale =1). Similarly (c) denotes *do* (blond hair = 1); (d) denotes *do* (blond hair = 1, pale = 1); (e) denotes *do* (brown hair = 1); (hf denotes *do* (brown hair = 1, pale = 1); and (g) denotes *do* (bangs = 1).

Dash et al, Evaluating and Mitigating Bias in Image Classifiers: A Causal Perspective Using Counterfactuals, WACV 2022

Counterfactual Generation under Confounding



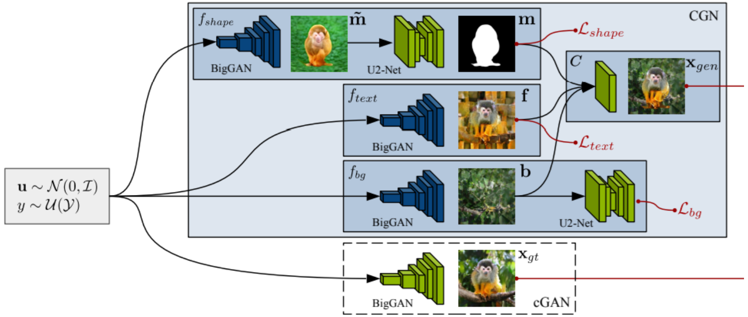
Cycle-GAN based learning methodology



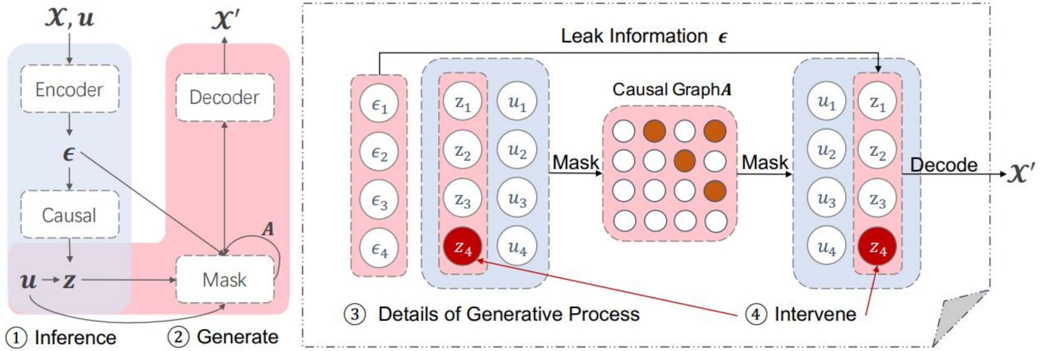
Reddy et al, Counterfactual Generation under Confounding, arXiv:2210.12368v2

Related Efforts

Counterfactual Generative Networks, ICLR 2021



CausalVAE, CVPR 2021



The "Memory" Pillar

A Few Takeaways

Memory:

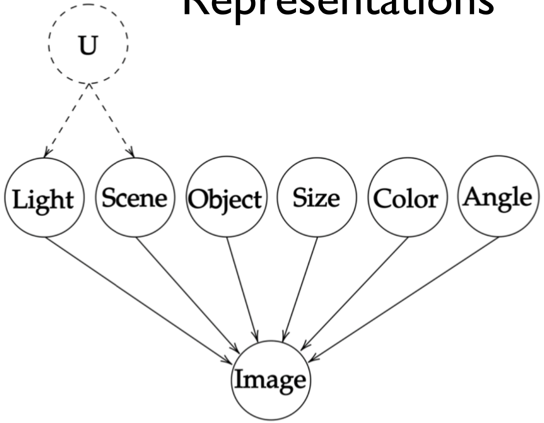
A key component of successful CL methods, where does causal come in?

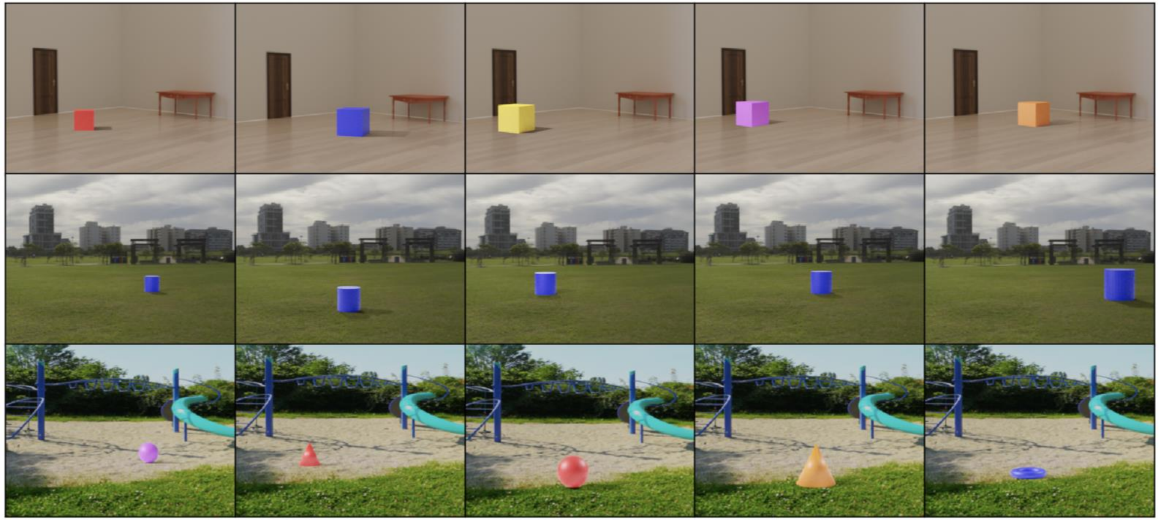
Replay buffers. Can they be causal?

- How to make replay buffers "causal"?
- How to leverage causal counterfactuals in feature-generative replay methods?
- Can memory go beyond data samples into causal domain knowledge? (e.g. our ICML 2022 work)

Need for Datasets/Benchmarks

CANDLE: An Image Dataset for Causal Analysis in Disentangled Representations





https://github.com/causal-disentanglement/CANDLE

Best Paper Award, CVPR 2021 Workshop on Causality in Vision

Context and Correlations

- Correlations have a life too!
- Dealing with context-based reasoning in causal models: An open question
- …

Takeaways

- Thinking/modeling in terms of independent causal mechanisms critical
- Disentanglement of causal mechanisms with real-world data non-trivial
- Need for (multimodal) datasets/benchmarks with causal ground truth
- Causal methods generally computationally intensive – how to cross this bridge?
- Maintain causal perspectives to counterfactuals in generative models
- Integration of causal domain knowledge into CL methods
- There is a place for correlation. What? Where?

Thank you!

Acknowledgements

…and to all students and collaborators

Questions?

vineethnb@cse.iith.ac.in

http://www.iith.ac.in/~vineethnb