Learning Causal Semantic Representation for Out-of-Distribution Prediction

Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, Tie-Yan Liu
changliu@microsoft.com

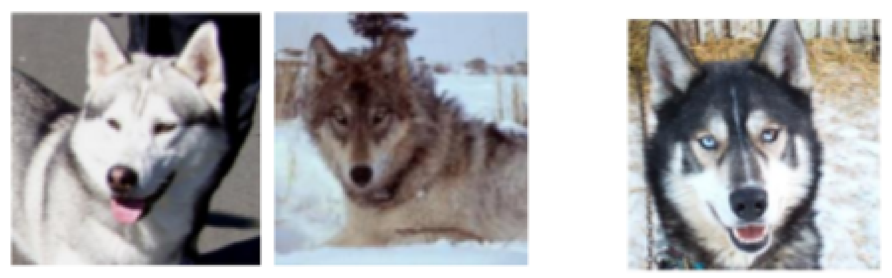
Introduction

Deep supervised learning lacks robustness to OOD samples.

Reason behind:

- The learned representation mixes **semantic factor** s (*e.g.*, shape) and **variation factor** v (*e.g.*, background), since both are correlated to y,
- but **only** s **causes** y: intervening v does not change y.

Train [Ribeiro'16]: Test:



"Husky" "Wolf" "Husky" (misleading to "Wolf")

This work: learn the **causal representation** for OOD prediction.

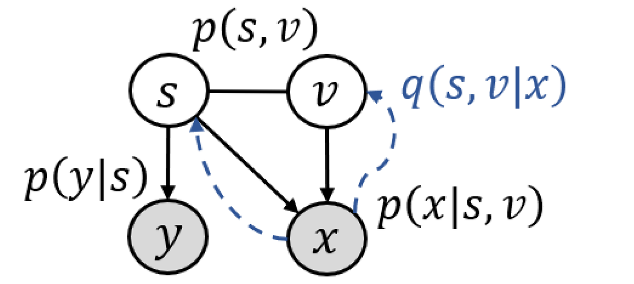
- Model: Causal Semantic Generative model (CSG) for latent causal structure.
- Method: **OOD generalization** and **domain adaptation** (single training domain).
- Theory: identification of the semantic factor and the subsequent benefits for OOD prediction.

Causal Semantic Generative Model (CSG)

Causality: intervening the cause may change the effect, but not vice versa.

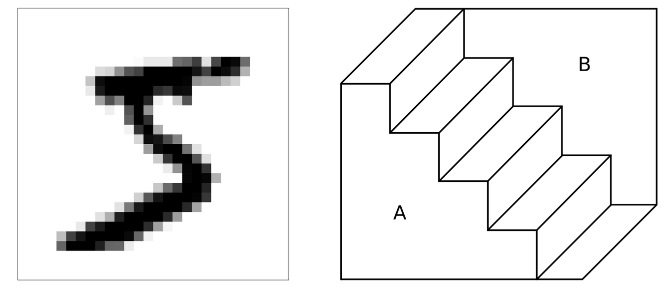
- Need latent variable z: breaking camera $x \nrightarrow y$, disturbing labeler $y \nrightarrow x$.
- $z \to (x, y)$: changing shape $z \to (x, y)$, breaking camera $x \nrightarrow z$.
- $z = (s, v)$: not all of z causes y (background $v \nrightarrow y$).
- s-v has a **spurious correlation** ("Wolf"-snow, but putting a "Wolf" in dark does not turn the background to snow).

CSG $p := \langle p_{s,v}, p_{x|s,v}, p_{y|s} \rangle$



Causal Invariance principle:

- Causal mechanisms $p(x|s, v)$ and $p(y|s)$ are domain-invariant, while the prior $p(s, v)$ is domain-specific.
- More general than **inference invariance**: $p(s, v|x)$ depends on $p(s, v)$ when $p(x|s, v)$ is noisy ("5" or "3"?) or degenerate (A or B is nearer).



Method

Training domain: fit data distribution $p^*(x, y)$.

max. likelihood $\overset{p(x,y)\text{ intractable}}{\Longrightarrow}$ max. ELBO $\mathcal{L}_{p,\, q_{s,v|x,y}}(x, y) := \mathbb{E}_{q(s,v|x,y)}[\log \frac{p(s,v,x,y)}{q(s,v|x,y)}] \leqslant \log p(x, y)$

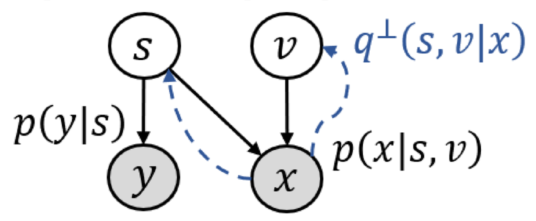
$\overset{q(s,v|x,y)\text{ does not help prediction}}{\Longrightarrow}$ use $q(s, v, y|x)$ and max. $\mathcal{L}_{p,\, q(s,v|x,y) = q(s,v,y|x) / \int q(s,v,y|x)\, \mathrm{d}s\mathrm{d}v}(x, y)$

$\overset{q(s,v,y|x)\text{ targets } p(s,v,y|x) = p(s,v|x)p(y|s)}{\Longrightarrow}$ approx. minimal intractable part $p(s, v|x)$ with $q(s, v|x)$ and $\max_{p, q_{s,v|x}} \mathbb{E}_{p^*(x,y)}[\mathcal{L}_{p,\, q(s,v|x,y) = q(s,v|x)p(y|s) / \int q(s,v|x)p(y|s)\, \mathrm{d}s\mathrm{d}v}(x, y)]$.

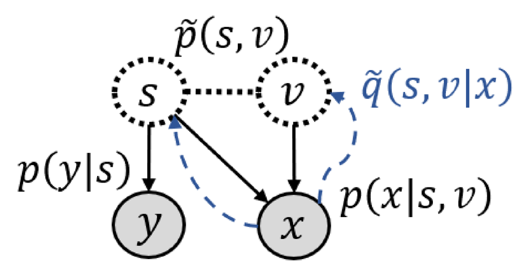
Test domain: same $p_{s,v|x}$, $p_{y|s}$, different prior $p_{s,v}$.

- **CSG-ind**: for OOD generalization, use the **ind**ependent prior $p^{\perp\!\!\!\perp}(s, v) := p(s)p(v)$ for the test domain.
- **CSG-DA**: for domain adaptation, learn the test-domain prior $\tilde{p}(s, v)$ using unsupervised data.
- **Avoid two q models**: use test-dom. q to express train-dom. q, *e.g.*, $q(s, v|x) = \frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)} \frac{p^{\perp\!\!\!\perp}(x)}{p(x)} q^{\perp\!\!\!\perp}(s, v|x)$.

CSG-ind



CSG-DA



Theory

Semantic identification: The learned s does not change with the ground-truth v.

- $\not\Rightarrow\!\not\Leftarrow$ s-v independence. $\not\Rightarrow\!\Leftarrow$ s-v disentanglement.

Theorem (semantic identifiability) A **well-learned** CSG achieves sem. identification, if: **(a)** $p_{x|s,v}$ is an additive noise μ model **(b)** with invertible mean function, and **(c)** $\log p(s, v)$ is bounded, and **(d1)** $\sigma_\mu^2 \to 0$, or **(d2)** p_μ has an a.e. non-zero characteristic function.

- **(c)** excludes deterministic s-v (all "Husky" in dark, all "Wolf" in snow): identification is impossible.
- Probabilistic s-v makes mixing ground-truth v into the learned s worsen training accuracy.

Theorem (OOD generalization) Prediction error of a sem. identified CSG on an unknown test domain is bounded: $\mathbb{E}_{\tilde{p}^*(x)} \|\mathbb{E}[y|x] - \tilde{\mathbb{E}}^*[y|x]\|_2^2 \leqslant C\sigma_\mu^4 D_{\text{Fisher}}(\tilde{p}_{s,v} \| p_{s,v})$.

- $p_{s,v}^{\perp\!\!\!\perp}$ has larger support than $p_{s,v}$ $\Longrightarrow$ $p_{s,v}^{\perp\!\!\!\perp}$ has smaller D_{Fisher} $\Longrightarrow$ **CSG-ind has a smaller error bound**.

Theorem (domain adaptation) Given a sem. identified CSG, a well-learned new prior is a reparameterization of the ground-truth, and gives accurate prediction: $\tilde{\mathbb{E}}[y|x] = \tilde{\mathbb{E}}^*[y|x]$.

Experiments

Datasets:

- **Shifted MNIST**: Train: horizontally move "0"s by $\delta_0 \sim \mathcal{N}(-5, 1^2)$ pixels and "1"s by $\delta_1 \sim \mathcal{N}(5, 1^2)$ pixels.
  Test: **(1)** $\delta_0 = \delta_1 = 0$, **(2)** $\delta_0, \delta_1 \sim \mathcal{N}(0, 2^2)$.
- **ImageCLEF-DA**, **PACS**: real-world images from multiple domains.

Baselines:

- **OOD gen.**: CE (conventional Cross-Entropy), CNBB (discriminative method with causal consideration).
- **Domain adaptation**: inference-invariance-based methods.

Test accuracy (%)		OOD generalization				domain adaptation				
Dataset	task	CE	CNBB	**CSG**	**CSG-ind**	DANN	DAN	CDAN	MDD	**CSG-DA**
Shifted MNIST	$\delta_0 = \delta_1 = 0$	42.9±3.1	54.7±3.3	81.4±7.4	**82.6±4.0**	40.9±3.0	40.4±2.0	41.0±0.5	41.9±0.8	**97.6±4.0**
	$\delta_0, \delta_1 \sim \mathcal{N}(0, 2^2)$	47.8±1.5	59.2±2.4	61.7±3.6	**62.3±2.2**	46.2±0.7	45.6±0.7	46.3±0.6	45.8±0.3	**72.0±9.2**
ImageCLEF-DA	**C→P**	65.5±0.3	72.7±1.1	73.6±0.6	**74.0±1.3**	74.3±0.5	69.2±0.4	74.5±0.3	74.1±0.7	**75.1±0.5**
	P→C	91.2±0.3	91.7±0.2	92.3±0.4	**92.7±0.2**	91.5±0.6	89.8±0.4	**93.5±0.4**	92.1±0.6	**93.4±0.3**
	I→P	74.8±0.3	75.4±0.6	76.9±0.3	**77.2±0.2**	75.0±0.6	74.5±0.4	76.7±0.3	76.8±0.4	**77.4±0.3**
	P→I	83.9±0.1	88.7±0.5	90.4±0.3	**90.9±0.2**	86.0±0.3	82.2±0.2	90.6±0.3	90.2±1.1	**91.1±0.5**
PACS	others→**P**	**97.8±0.0**	96.9±0.2	97.7±0.2	**97.8±0.2**	97.6±0.2	97.6±0.4	97.0±0.4	97.6±0.3	**97.9±0.2**
	others→**A**	88.1±0.1	73.1±0.3	**88.5±0.6**	**88.6±0.6**	85.9±0.5	84.5±1.2	84.0±0.9	88.1±0.8	**88.8±0.7**
	others→**C**	77.9±1.3	50.2±1.2	84.4±0.9	**84.6±0.8**	79.9±1.4	81.9±1.9	78.5±1.5	83.2±1.1	**84.7±0.8**
	others→**S**	79.1±0.9	43.3±1.2	80.7±1.0	**81.1±1.2**	75.2±2.8	77.4±3.1	71.8±3.9	80.2±2.2	**81.4±0.8**

Visualization

CE

CSG-ind

MDD

CSG-DA