Learning Causal Semantic Representation for out-of-Distribution Prediction

Chang Liu[1], Xinwei Sun[1], Jindong Wang[1], Haoyue Tang[2], Tao Li[3],
Tao Qin[1], Wei Chen[1], Tie-Yan Liu[1].

[1] Microsoft Research Asia

[2] Tsinghua University

[3] Peking University

Introduction

The problem:

- Deep supervised learning lacks robustness to out-of-distribution (OOD) samples.

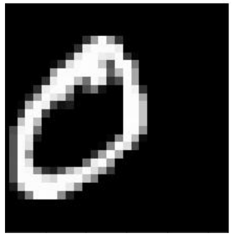
Reason behind:

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

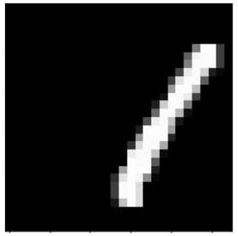
Goal:

- Learning the **causal** representation for OOD prediction.

Train:

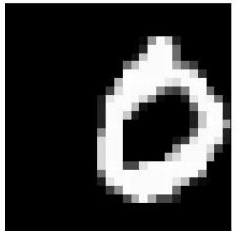


"0"



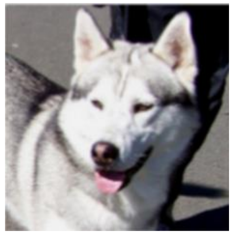
"1"

Test:



"0" (misleading to "1")

Train:



"Husky"

"Wolf"

Test:

"Husky" (misleading to "Wolf")

influential region [Ribeiro'16]

Introduction

In this work,

- Causal Semantic Generative model (CSG): describes latent causal structure.
- Methods for OOD prediction (OOD generalization and domain adaptation).
- Theory for identifying the semantic factor and the subsequent benefits for OOD prediction.

Related Work

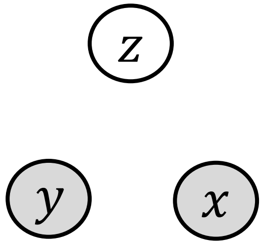
- Domain adaptation/generalization.
  - Observation-level causality: not suitable for general data like images.
  - Domain-invariant representation: inference invariance; insufficient to identify latent factors.
  - Latent generative models: inference invariance; semantic-variation independence; lack of identifiability guarantee.
- Learning disentangled representation.
  - Impossible in unsupervised learning, despite some empirical success.
  - With an auxiliary variable [Khemakhem'20a,b]: require sufficiently many different values of the variable (thus unsuitable for y); no description for domain change.

Related Work

- Generative supervised learning.
  - Few utilized the causal implications of the model.
  - Some aim at estimating causal/treatment effect: not suitable for OOD prediction.
- Causality with latent variables.
  - Most works still focus on the consequence on observation-level causality.
  - Works that identify latent variables do not have semantic-variation split.
- Causal discriminative learning.
  - Lack of identifiability guarantee and structure to capture causal relations.

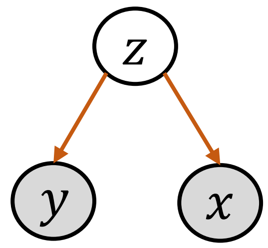
The Model

- Formal definition of causality:
  "*two variables have a* causal relation, *if intervening the cause (by changing external variables out of the considered system) may change the effect, but not* vice versa" [Pearl'09; Peters'17].



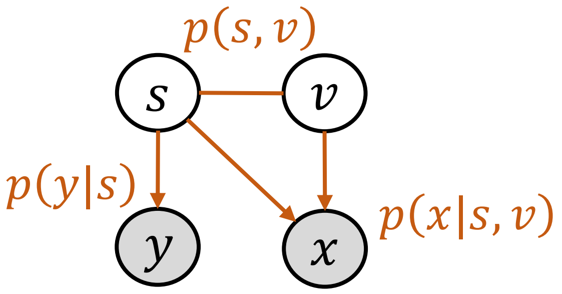
- Causal Semantic Generative (CSG) Model
  - The need of latent variable z:
    $x \nrightarrow y$ (breaking a camera sensor unit $x \nrightarrow$ label y), $y \nrightarrow x$ (labeling noise $y \nrightarrow$ image x).
    (For labeling process from image x: labelers are doing inference; preference may change from person to person.)

The Model



- Formal definition of causality:
  "*two variables have a* causal relation, *if intervening the* *cause* *(by changing external variables out of the considered system) may change the* *effect*, *but not* vice versa" [Pearl'09; Peters'17].
- Causal Semantic Generative (CSG) Model
  - The need of latent variable z:
    $x \nrightarrow y$ (breaking a camera sensor unit $x \nrightarrow$ label y), $y \nrightarrow x$ (labeling noise $y \nrightarrow$ image x).
    (For labeling process from image x: labelers are doing inference; preference may change from person to person.)
  - $z \rightarrow (x, y)$: changing object shape z in the scene $\rightarrow$ image x, label y;
    breaking sensor x or labeling noise $y \nrightarrow$ object shape z in the scene.
    (Particularly, different from works with $y \rightarrow s$: our y may be a noisy observation.)
  - No x-y edge: attribute all x-y relations to latent factors ("purely common cause", promotes identification)
    (breaking sensor x / labeling noise y while fixing all factors $z \nrightarrow$ label y / image x).
  -
  -
  -

The Model



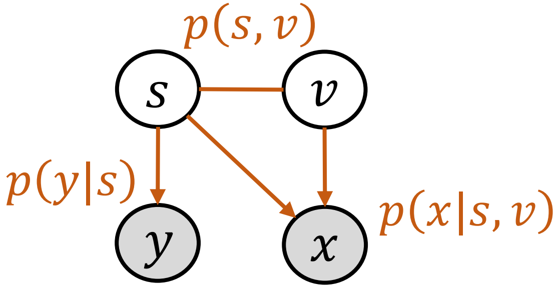
- Formal definition of causality:
  "two variables have a causal relation*, if intervening the cause (by changing external variables out of the considered system) may change the effect, but not* vice versa*"* [Pearl'09; Peters'17].
- Causal Semantic Generative (CSG) Model
  - The need of latent variable z:
    $x \nrightarrow y$ (breaking a camera sensor unit $x \nrightarrow$ label y), $y \nrightarrow x$ (labeling noise $y \nrightarrow$ image x).
    (For labeling process from image x: labelers are doing inference; preference may change from person to person.)
  - $z \to (x, y)$: changing object shape z in the scene $\to$ image x, label y;
    breaking sensor x or labeling noise $y \nrightarrow$ object shape z in the scene.
    (Particularly, different from works with $y \to s$: our y may be a noisy observation.)
  - No x-y edge: attribute all x-y relations to latent factors ("purely common cause", promotes identification)
    (breaking sensor x / labeling noise y while fixing all factors $z \nrightarrow$ label y / image x).
  - $z = (s, v)$: not all factors *cause* y (changing background $v \nrightarrow$ label y).
  - s-v has a relation, which is often spurious (desk ~ workspace, bed ~ bedroom, but putting a desk in bedroom does not turn it into a bed).
  - Denoted as $p := \langle p_{s,v}, p_{x|s,v}, p_{y|s} \rangle$.

The Model

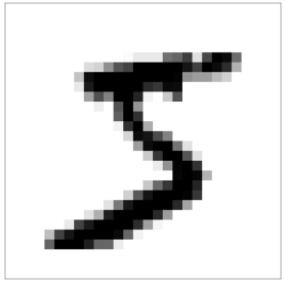
- The **causal invariance** principle:

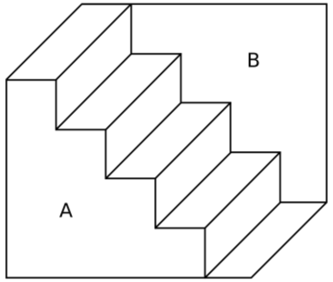
  Causal mechanisms $p(x|s,v)$ and $p(y|s)$ are domain-invariant, while the prior $p(s,v)$ is the source of domain shift.

  - Stems from the *Independent Causal Mechanisms* principle: intervening $p(s,v)$ does not affect $p(x|s,v)$, $p(y|s)$.



- Comparison to **inference invariance**: $p(s,v|x)$ is invariant.
  - Domain adapt./gen., invariant risk min.: use a *shared* encoder across domains.
  - Special case of causal invariance when generative mechanisms are almost deterministic and invertible (inferring object position from image, extracting F0 from audio).
  - When they are not, inference is ambiguous and rely on domain-specific prior.





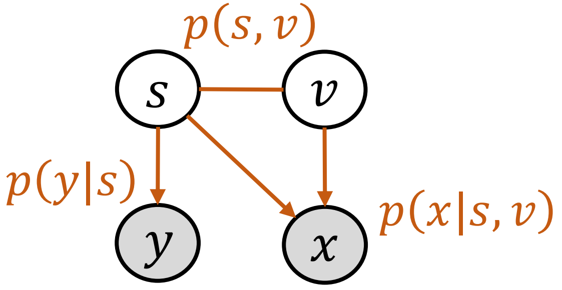
$$p(s,v|x) \propto \underbrace{p(s,v)}_{\text{domain-specific}} \underbrace{p(x|s,v)}_{\neq 0 \text{ for multiple } (s,v)}$$

Inference ambiguity in **Noisy (“5” or “3”?)** and **Degenerate (A or B nearer?)** generative mechanisms.

Method

true data distribution

$= \int p(s,v)p(x|s,v)p(y|s)\, \mathrm{d}s\mathrm{d}v$ is hard to evaluate.



- Direct MLE: $\max_p \mathbb{E}_{p^*(x,y)}[\log p(x,y)]$.
- Standard ELBO: using a tractable *inference model* $q(s,v|x,y)$,

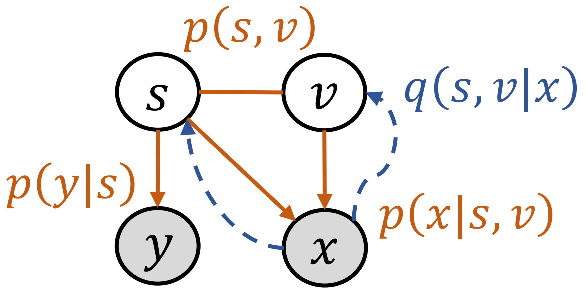
$$\mathcal{L}_{p,q}(x,y) \coloneqq \mathbb{E}_{q(s,v|x,y)}\left[\log \frac{p(s,v,x,y)}{q(s,v|x,y)}\right] \leq \log p(x,y).$$

  - $\max_q \mathcal{L}_{p,q}(x,y)$ makes $q(s,v|x,y) \to p(s,v|x,y)$ and $\mathcal{L}_{p,q}(x,y) \to \log p(x,y)$.
  - Prediction is still hard: hard to leverage $q(s,v|x,y)$.

Method

- Use a $q(s, v, y|x)$ model:
  - For prediction: ancestral sampling.
  - For learning: $\mathbb{E}_{p^*(x,y)}\left[\mathcal{L}_{p,\, q(s,v|x,y) = q(s,v,y|x) / \int q(s,v,y|x)\, \mathrm{d}s\mathrm{d}v}(x, y)\right]$

$$= \mathbb{E}_{p^*(x)}\left[\underbrace{\mathbb{E}_{p^*(y|x)}[\log q(y|x)]} + \underbrace{\mathbb{E}_{q(s,v,y|x)}\left[\frac{p^*(y|x)}{q(y|x)} \log \frac{p(s,v,x,y)}{q(s,v,y|x)}\right]}\right].$$



(negative) cross-entropy:
makes $q(y|x) \to p^*(y|x)$

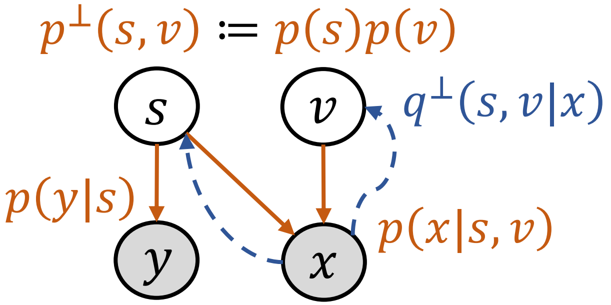
$= \mathcal{L}_{p,\, q(s,v,y|x)}(x)$ when $q(y|x) = p^*(y|x)$:
makes $q(s, v, y|x) \to p(s, v, y|x)$, $\mathcal{L}_{p,\, q(s,v,y|x)}(x) \to p(x)$

Since $p(s, v, y|x) = p(s, v|x) p(y|s)$,
let $q(s, v, y|x) = q(s, v|x) p(y|s)$

- Use a $q(s, v, y|x)$ model:

$$\mathcal{L}_{p,\, q(s,v|x,y) = [q(s,v|x), p]}(x, y) = \log q(y|x) + \frac{1}{q(y|x)} \mathbb{E}_{q(s,v|x)}\left[p(y|s) \log \frac{p(s,v) p(x|s,v)}{q(s,v|x)}\right].$$

Method



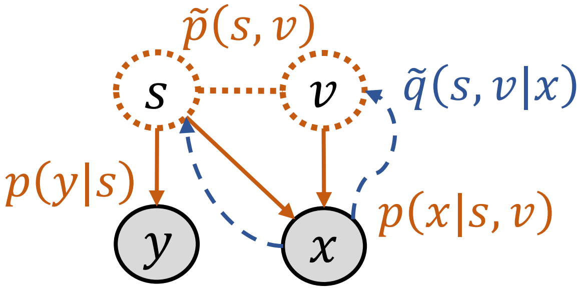
CSG-ind: for prediction in an *unknown* test domain (OOD gen.)

- Use an **ind**ependent prior $p^\perp(s,v) \coloneqq p(s)p(v)$:
  - Discard the spurious s-v correlation; *defensive* choice.
  - Larger entropy than $p(s,v)$: reduce training-domain-specific information.
  - Randomized experiment by independently soft-intervening s or v.
- On the test domain:
  - Prediction: $p^\perp(y|x) \approx \mathbb{E}_{q^\perp(s,v|x)}[p(y|s)]$. Different from $p(y|x)$ (inference invariance).
- On the training domain: avoid the $q(s,v|x)$ model.
  - Following the relation b/w their targets, let $q(s,v|x) = \frac{p(s,v)}{p^\perp(s,v)} \frac{p^\perp(x)}{p(x)} q^\perp(s,v|x)$:

$$\mathcal{L}_{p,\, q(s,v|x,y)=[q^\perp(s,v|x),p]}(x,y) = \log \pi(y|x) + \frac{1}{\pi(y|x)} \mathbb{E}_{q^\perp(s,v|x)}\left[\frac{p(s,v)}{p^\perp(s,v)} p(y|s) \log \frac{p^\perp(s,v)p(x|s,v)}{q^\perp(s,v|x)}\right],$$

where $\pi(y|x) \coloneqq \mathbb{E}_{q^\perp(s,v|x)}\left[\frac{p(s,v)}{p^\perp(s,v)} p(y|s)\right]$.

Method



CSG-DA: for prediction in a test domain with unsupervised data (domain adaptation)

- On the test domain:
  - Learn the test-domain prior $\tilde{p}(s,v)$ by fitting $\tilde{p}^*(x)$ using ELBO:
    $\mathcal{L}_{\tilde{p},\tilde{q}}(x) := \mathbb{E}_{\tilde{q}(s,v|x)}\left[\log \frac{\tilde{p}(s,v)p(x|s,v)}{\tilde{q}(s,v|x)}\right] \leq \log \tilde{p}(x).$
  - Prediction: $\tilde{p}(y|x) \approx \mathbb{E}_{\tilde{q}(s,v|x)}[p(y|s)]$. Different from $p(y|x)$ (inference invariance).
- On the training domain: avoid the $q(s,v|x)$ model.
  - Following the relation b/w their targets, let $q(s,v|x) = \frac{\tilde{p}(x)}{p(x)} \frac{p(s,v)}{\tilde{p}(s,v)} \tilde{q}(s,v|x)$:

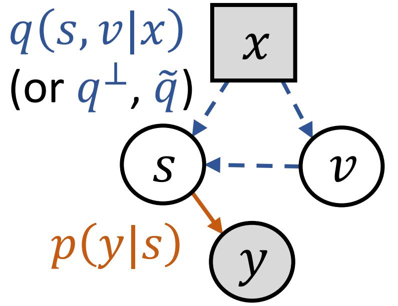
$$\mathcal{L}_{p,\, q(s,v|x,y)=[\tilde{q}(s,v|x),p]}(x,y) = \log \pi(y|x) + \frac{1}{\pi(y|x)} \mathbb{E}_{\tilde{q}(s,v|x)}\left[\frac{p(s,v)}{\tilde{p}(s,v)} p(y|s) \log \frac{\tilde{p}(s,v)p(x|s,v)}{\tilde{q}(s,v|x)}\right],$$

where $\pi(y|x) = \mathbb{E}_{\tilde{q}(s,v|x)}\left[\frac{p(s,v)}{\tilde{p}(s,v)} p(y|s)\right]$.

Method

Implementation details.

- Instantiating the model by parsing a general discriminative model:
  - In CSG, $y \perp (x, v) | s$, so no $v \to y$. We then have $p(y|s)$.
  - In CSG, $s \;!\!\perp\; v \mid x$, so let $v \to s$. We then have $q(s, v|x)$.
  - Use an additional model for $p(x|s, v)$.



(computation direction)

- Implementing the prior.
  - Multivariate Gaussian: $p(s, v) = \mathcal{N}\left(\begin{pmatrix} s \\ v \end{pmatrix} \middle| \begin{pmatrix} \mu_s \\ \mu_v \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{ss} & \Sigma_{sv} \\ \Sigma_{vs} & \Sigma_{vv} \end{pmatrix}\right)$ (no causal direction).
  - Parameterize $\Sigma = LL^{\top}, L = \begin{pmatrix} L_{ss} & 0 \\ M_{vs} & L_{vv} \end{pmatrix}$ (L_{ss}, L_{vv} are lower-triangular with positive diagonals).
  - $p(v|s) = \mathcal{N}\left(v | \mu_{v|s}, \Sigma_{v|s}\right)$, where $\mu_{v|s} = \mu_v + M_{vs} L_{ss}^{-1}(s - \mu_s)$, $\Sigma_{v|s} = L_{vv} L_{vv}^{\top}$.
- Model selection.
  - Use a validation set from the **training domain**.
  - For **CSG-ind/DA**, use $p(y|x) \propto \pi(y|x)$ ($\neq p^{\perp}(y|x)$ or $\tilde{p}(y|x)$) to evaluate validation accuracy.

Theory

Identifiability on the training domain.

- **Definition** (semantic-identification). A CSG p is said *semantic-identified*, if there exists a homeomorphism Φ on $\mathcal{S} \times \mathcal{V}$, s.t.: **(i)** $\Phi^{\mathcal{S}}(s^*, v^*)$ is constant of v^*, and **(ii)** Φ is a *reparameterization* of the ground-truth CSG p^*:
  $\Phi_{\#}\left[p^*_{s,v}\right] = p_{s,v}$, $p^*(x|s^*, v^*) = p(x|\Phi(s^*, v^*))$, $p^*(y|s^*) = p\left(y\middle|\Phi^{\mathcal{S}}(s^*)\right)$.
  - Reparameterization: describes the degree of freedom given $p(x, y) = p^*(x, y)$.
  - v-constancy: Φ is *semantic-preserving* (the learned s does not mix the ground-truth v^* into it).
  - **Proposition**: equivalent relation if $\mathcal{V}$ is connected and is either open or closed in $\mathbb{R}^{d_{\mathcal{V}}}$.
- Related concepts:
  - Neither sufficient nor necessary to **statistical independence**.
  - Weaker than **disentanglement**: the learned v can be entangled with ground-truth s^*.

Theory

Identifiability on the training domain.

- Assumptions.
  - (**A1**)[*additive noise*] There exist functions f and g with bounded derivatives up to 3rd-order, and indep. r.v.s μ and ν, s.t.:
    $p(x|s,v) = p_\mu\big(x - f(s,v)\big)$, and
    $p(y|s) = p_\nu\big(y - g(s)\big)$ for continuous y or $\mathrm{Cat}\big(y|g(s)\big)$ for categorical y.
    - Required to disable the anti-causal direction.
    - Excludes GAN, flow-based models.
  - (**A2**)[*bijectivity*] f is bijective and g is injective.
    - A common sufficient condition for the fundamental requirement of causal minimality.
    - Otherwise, s and v are allowed to have dummy dimensions.
    - The manifold hypothesis relaxes f to be injective, and allows $d_{\mathcal{S}} + d_{\mathcal{V}} < d_{\mathcal{X}}$.

Theory

Identifiability on the training domain.

- **Theorem** (semantic-identifiability). Assume **A1,A2**, bounded $\log p^*_{s,v}$ up to 2nd-order, and:
  (i) $\frac{1}{\sigma_\mu^2} \to \infty$, where $\sigma_\mu^2 \coloneqq \mathbb{E}[\mu^\top \mu]$, **or**
  (ii) p_μ has an a.e. non-zero characteristic function (e.g., a Gaussian distribution).

  Then a well-learned CSG (s.t. $p(x, y) = p^*(x, y)$) is *semantic-identified*.
  - (**Appropriate condition**) One cannot identify s in *extreme cases* (all "0"'s are on the left and all "1"'s are on the right): excluded by the condition on $\log p^*_{s,v}$.
  - (**Intuition**) In other cases, v for each s is noisy, so mixing s with v worsens training accuracy.
  - Condition (i) requires a *strong* causal mechanism: nearly deterministic and invertible. Condition (ii) covers more than inference invariance.
  - Does not contradict the impossibility result of disentanglement [Locatello'19]: only identify s as a whole; asymmetry from missing $v \to y$.

Theory

Benefit for OOD prediction.

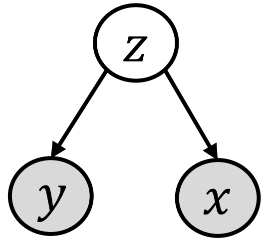
- The test-domain ground-truth CSG $\tilde{p}^* = \langle \tilde{p}^*_{s,v}, p^*_{x|s,v}, p^*_{y|s} \rangle$ (due to causal invariance).
- **Theorem** (OOD gen. error) With **A1**,**A2**, the test-domain prediction error of a *semantic-identified* CSG p is bounded ($B'_{f^{-1}}$, B'_g bounds the Jacobian 2-norms of f^{-1}, g, and $\tilde{p}_{s,v} \coloneqq \Phi_\#[\tilde{p}^*_{s,v}]$):
  $\mathbb{E}_{\tilde{p}^*(x)} \left\| \mathbb{E}[y|x] - \widetilde{\mathbb{E}}^*[y|x] \right\|_2^2 \leq \sigma_\mu^4 B'^4_{f^{-1}} B'^2_g \mathbb{E}_{\tilde{p}_{s,v}} \left\| \nabla \log\left(\tilde{p}_{s,v} / p_{s,v} \right) \right\|_2^2.$ (up to $O(\sigma_\mu^4)$)
  - For a *strong* causal mechanism $p(x|s,v)$, the bound is small.
  - $\mathbb{E}_{\tilde{p}_{s,v}} \left\| \nabla \log\left(\tilde{p}_{s,v} / p_{s,v} \right) \right\|_2^2$: $\mathrm{FisherDiv}\left(\tilde{p}_{s,v} \| p_{s,v} \right)$, "OODness" for prediction.
  - CSG-ind tends to have a smaller error bound:
    smaller $\mathrm{FisherDiv}\left(\tilde{p}_{s,v} \| \cdot \right) \Rightarrow$ distr. with a larger support, and $p^\perp_{s,v}$ has a larger support than $p_{s,v}$.
- **Theorem** (domain adaptation error) Assume the same for identifiability and the learned CSG p is *semantic-identified*. Then a well-learned (s.t. $\tilde{p}(x) = \tilde{p}^*(x)$) new prior
  (i) $\tilde{p}_{s,v} = \Phi_\# \left[\tilde{p}^*_{s,v} \right]$ is a reparametrized ground-truth $\tilde{p}^*_{s,v}$, and
  (ii) it leads to an accurate prediction: $\widetilde{\mathbb{E}}[y|x] = \widetilde{\mathbb{E}}^*[y|x], \forall x \in \mathrm{supp}(\tilde{p}^*_x)$.

Experiments

Baselines:

- For OOD generalization,
  - CE (cross entropy): standard supervised learning.
  - CNBB (ConvNet with Batch Balancing): a discriminative causal method.
- For domain adaptation,
  - DANN, DAN, CDAN, MDD, BNM: classical domain adaptation methods.
- For an ablation study,
  - CSGz / CSGz-DA: generative methods without separating z as s and v.

CSGz / CSGz-DA



Datasets:

- Shifted MNIST.
  - Training dataset: "0"s are horiz. shifted by $\delta_0 \sim \mathcal{N}(-5,1^2)$ px, "1"s by $\delta_1 \sim \mathcal{N}(5,1^2)$ px.
  - Two test datasets: (1) $\delta_0 = \delta_1 = 0$; (2) $\delta_0, \delta_1 \sim \mathcal{N}(0,2^2)$.
- ImageCLEF-DA.
- PACS, VLCS.

Experiments

- OOD prediction performance

OOD generalization

task		CE	CNBB	CSGz	**CSG**	**CSG-ind**
Shifted-MNIST	$\delta_0 = \delta_1 = 0$	42.9±3.1	54.7±3.3	53.0±6.7	81.4±7.4	**82.6**±4.0
	$\delta_0, \delta_1 \sim \mathcal{N}(0, 2^2)$	47.8±1.5	59.2±2.4	54.8±5.6	61.7±3.6	**62.3**±2.2
ImageCLEF-DA	**C**→**P**	65.5±0.3	72.7±1.1	73.3±1.0	73.6±0.6	**74.0**±1.3
	P→**C**	91.2±0.3	91.7±0.2	91.6±0.9	92.3±0.4	**92.7**±0.2
	I→**P**	74.8±0.3	75.4±0.6	77.0±0.2	76.9±0.3	**77.2**±0.2
	P→**I**	83.9±0.1	88.7±0.5	90.4±0.3	90.4±0.3	**90.9**±0.2
PACS	others→**P**	**97.8**±0.0	96.9±0.2	97.7±0.3	97.7±0.2	**97.8**±0.2
	others→**A**	88.1±0.1	73.1±0.3	87.3±0.8	**88.5**±0.6	**88.6**±0.6
	others→**C**	77.9±1.3	50.2±1.2	84.3±0.9	84.4±0.9	**84.6**±0.8
	others→**S**	79.1±0.9	43.3±1.2	80.6±1.4	80.7±1.0	**81.1**±1.2

Domain adaptation

task		DANN	DAN	CDAN	MDD	BNM	CSGz-DA	**CSG-DA**
Shifted-MNIST	$\delta_0 = \delta_1 = 0$	40.9±3.0	40.4±2.0	41.0±0.5	41.9±0.8	40.8±1.0	78.0±27.2	**97.6**±4.0
	$\delta_0, \delta_1 \sim \mathcal{N}(0, 2^2)$	46.2±0.7	45.6±0.7	46.3±0.6	45.8±0.3	45.7±1.0	68.1±17.4	**72.0**±9.2
ImageCLEF-DA	**C**→**P**	74.3±0.5	69.2±0.4	74.5±0.3	74.1±0.7	**75.2**±1.4	74.3±0.3	**75.1**±0.5
	P→**C**	91.5±0.6	89.8±0.4	**93.5**±0.4	92.1±0.6	**93.5**±2.8	92.7±0.4	**93.4**±0.3
	I→**P**	75.0±0.6	74.5±0.4	76.7±0.3	76.8±0.4	76.7±1.4	77.0±0.3	**77.4**±0.3
	P→**I**	86.0±0.3	82.2±0.2	90.6±0.3	90.2±1.1	**91.0**±0.8	90.6±0.4	**91.1**±0.5
PACS	others→**P**	97.6±0.2	97.6±0.4	97.0±0.4	97.6±0.3	87.6±4.2	97.6±0.4	**97.9**±0.2
	others→**A**	85.9±0.5	84.5±1.2	84.0±0.9	88.1±0.8	86.4±0.4	88.0±0.8	**88.8**±0.7
	others→**C**	79.9±1.4	81.9±1.9	78.5±1.5	83.2±1.1	83.6±1.7	**84.6**±0.9	**84.7**±0.8
	others→**S**	75.2±2.8	77.4±3.1	71.8±3.9	80.2±2.2	59.1±1.5	80.9±1.2	**81.4**±0.8

More suitable scenarios:
Solve the spurious correlation problem in cases with diverse v for each s (easier identification).

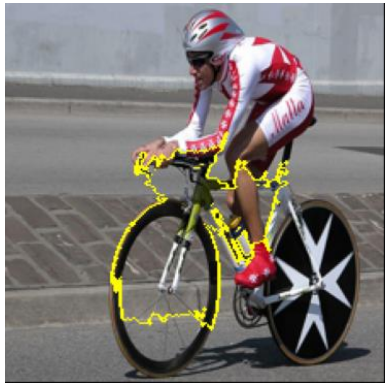
Experiments

- Visualization (using LIME [Ribeiro’16])

OOD generalization

CE

CSG-ind



Domain adaptation

MDD

CSG-DA

Thanks!

https://arxiv.org/abs/2011.01681

References

- [Ribeiro’16] M. T. Ribeiro, S. Singh, and C. Guestrin. “Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.
- [Khemakhem’20a] I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *the 23rd International Conference on Artificial Intelligence and Statistics, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217, 2020.
- [Khemakhem’20b] I. Khemakhem, R. P. Monti, D. P. Kingma, and A. Hyvärinen. ICE-BeeM: Identifiable conditional energy-based deep models. *arXiv preprint arXiv:2002.11537*, 2020.
- [Pearl’09] J. Pearl. *Causality*. Cambridge university press, 2009.
- [Peters’17] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.
- [Locatello’19] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124, Long Beach, California, USA, 09–15 Jun 2019. PMLR.