Recovering Latent Causal Factor for Generalization to Distributional Shifts

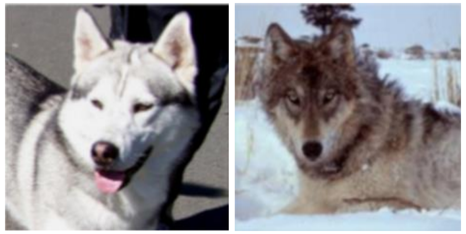
Xinwei Sun[1], Botong Wu[2], Xiangyu Zheng[2], Chang Liu[1],
Tao Qin[1], Wei Chen[1], Tie-Yan Liu[1].

[1] Microsoft Research Asia

[2] Peking University

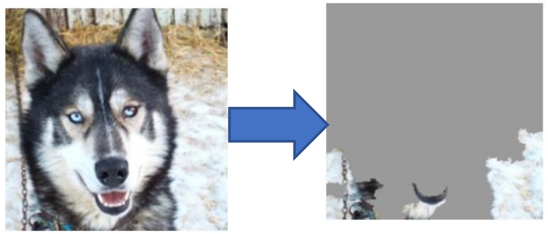
Recap: Causal Semantic Generative model (CSG)

Train:



“Husky” “Wolf”

Test:



“Husky”
(misleading to “Wolf”)

influential region
[Ribeiro’16]

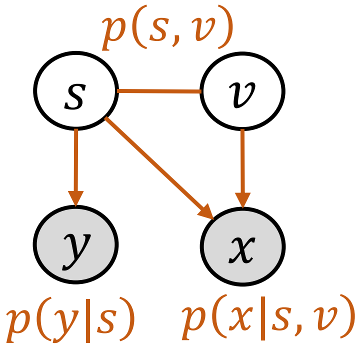
The problem:

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

Goal:

- Learn a **causal** representation on a **single supervised domain**, that distinguishes the *semantic factor* s (e.g., shape) and *variation factor* v (e.g., position, background).

CSG model:



- Only s causes y (changing background $v \nrightarrow$ label y).
- Spurious s-v correlation (husky-dark, but put it in snow does not turn it into a wolf).
- **Causal Invariance** principle: causal mechanisms $p(x|s,v)$, $p(y|s)$ are invariant, while the change of prior $p(s,v)$ leads to domain shift.

Liu, C., Sun, X., Wang, J., Tang, H., Li, T., Qin, T., Chen, W., Liu, T. Y. Learning Causal Semantic Representation for Out-of-Distribution Prediction. In *Advances in Neural Information Processing Systems*, 2021. [Slides]

Recap: Causal Semantic Generative model (CSG)

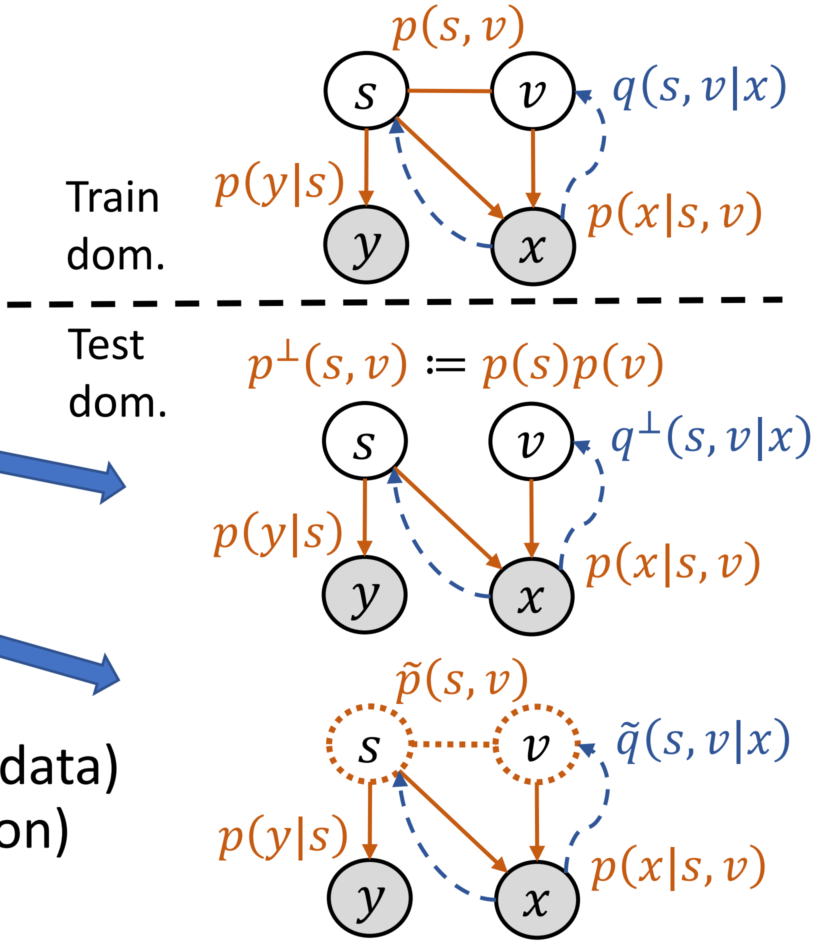
Method

- Variational Bayes using inference model $q(s,v|x)$, for both *tractable learning* (ELBO) and *easy prediction* ($\mathbb{E}_{q(s,v|x)}[p(y|s)]$).
- Predict on an unknown domain (*OOD generalization*):

  Use an independent prior (**CSG-ind**).
- Predict with unsupervised data (*domain adaptation*):

  Learn a new prior with the data (**CSG-DA**).
- To avoid two inference models:

  Express the training-domain $q(s,v|x)$ model (for fitting training data) with the test-domain $q^{\perp}(s,v|x)$ or $\tilde{q}(s,v|x)$ model (for prediction) via the relation between their targets.



Liu, C., Sun, X., Wang, J., Tang, H., Li, T., Qin, T., Chen, W., Liu, T. Y. Learning Causal Semantic Representation for Out-of-Distribution Prediction. In *Advances in Neural Information Processing Systems*, 2021. [Slides]

Recap: Causal Semantic Generative model (CSG)

Theory

- A CSG is **semantic-identified**, if there is a

  (1) *reparametrization* Φ to the ground-truth CSG, that

  (2) *does not mix* v *into* s, i.e. $s = \Phi^{\mathcal{S}}(s^*, v^*)$ is constant of v^*.

- $(s, v) = \Phi(s^*, v^*)$ s.t.: $\Phi_\#[p^*_{s,v}] = p_{s,v}$, $p^*(x|s^*, v^*) = p(x|\Phi(s^*, v^*))$, $p^*(y|s^*) = p(y|\Phi^{\mathcal{S}}(s^*, v^*))$.
- Describes the difference b/w CSGs with the same $p(x, y)$.

- **Thm (sem.-identifiability)**. A well-learned CSG (s.t. $p(x, y) = p^*(x, y)$) is semantic-identified, if:

  (a) Additive noise stru. whose **(b)** fn. are bijective, **(c)** $\log p_{s,v}$ is bounded, and
  (d) noise var. have vanishing variance σ_μ^2 or a.e. nonzero characteristic fn.

Excl. deterministic s-v relation

Benefits to OOD prediction:

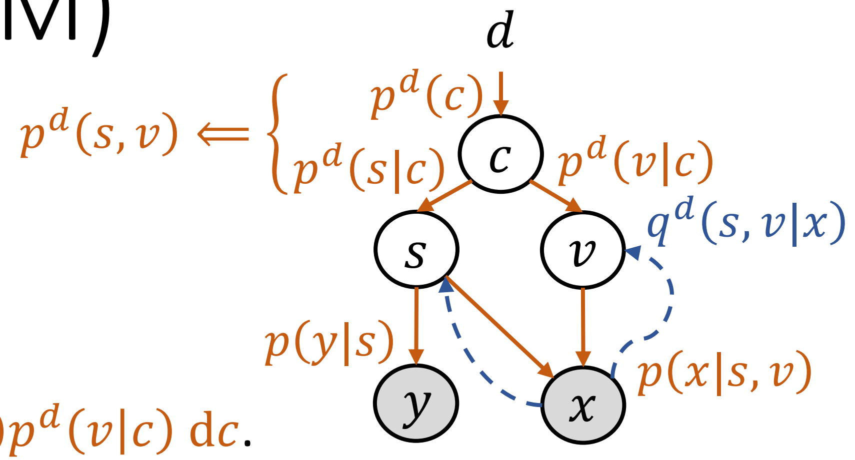
- **Thm (OOD gen).** Given sem.-identification, prediction error on an unknown domain is bounded:
  $\mathbb{E}_{\tilde{p}^*(x)} \left\| \mathbb{E}[y|x] - \tilde{\mathbb{E}}^*[y|x] \right\|_2^2 \leq C \sigma_\mu^4 \mathbb{E}_{\tilde{p}_{s,v}} \left\| \nabla \log(\tilde{p}_{s,v} / p_{s,v}) \right\|_2^2$.

CSG-ind makes a smaller bound

- **Thm (DA).** Given sem.-identification, a well-learned new prior $\tilde{p}_{s,v}$ (s.t. $\tilde{p}(x) = \tilde{p}^*(x)$) is a reparametrized ground-truth $\tilde{p}^*_{s,v}$, and makes an accurate prediction: $\tilde{\mathbb{E}}[y|x] = \tilde{\mathbb{E}}^*[y|x]$.

Liu, C., Sun, X., Wang, J., Tang, H., Li, T., Qin, T., Chen, W., Liu, T. Y. Learning Causal Semantic Representation for Out-of-Distribution Prediction. In *Advances in Neural Information Processing Systems*, 2021. [Slides]

Latent Causal Invariant Model (LaCIM)



Extending CSG for **multiple supervised domains** (i.e., *domain generalization*):

- Model the dependency on domain index d.
- Ascribe the spurious s-v correlation to a confounder c: integrating out c renders s-v correlated: $p^d(s,v) = \int p^d(c)p^d(s|c)p^d(v|c)\,\mathrm{d}c$.
- Causal invariance ➔ $p(x|s,v)$, $p(y|s)$ do not depend on d, while $p^d(s,v)$, $q^d(s,v|x)$ do.

Method

- **Training**: Apply **CSG training objective** to each domain, with the respective $p^d(s,v)$, $q^d(s,v|x)$:
$$\max_{p^d(s,v),p(x|s,v),p(y|s),q^d(s,v|x)} \sum_{d\in[D]} \mathbb{E}_{p^{*d}(x,y)}\left[\log q^d(y|x) + \frac{1}{q^d(y|x)} \mathbb{E}_{q^d(s,v|x)}\left[p(y|s) \log \frac{p^d(s,v)p(x|s,v)}{q^d(s,v|x)}\right]\right],$$
where $q^d(y|x) := \mathbb{E}_{q^d(s,v|x)}[p(y|s)]$.
- **Prediction** in an unseen test domain d':
Similar to **CSG-ind**, but by *direct optimization* (Max. A Posteriori est., not by inference model $q^\perp(s,v|x)$):
$p^{d'}(y|x) = p(y|s(x))$, where $\big(s(x), v(x)\big) := \arg\max_{s,v} p(x|s,v)p^\perp(s,v)^\lambda$.

Latent Causal Invariant Model (LaCIM)

Identifiability theory

K_S, K_V-dimensional

- Definition considering the dependence on d:

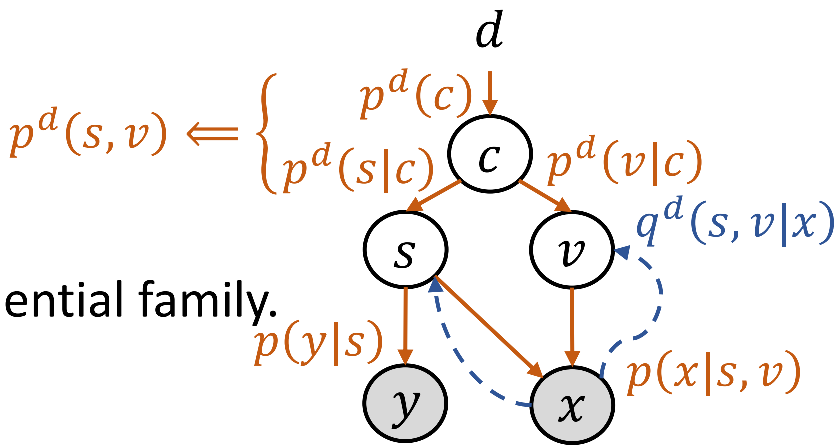
  Let $p^d(s|c) \propto \prod_{i\in[N_S]} \exp\left(\theta_i^d(c)^\top S_i(s_i) + A_i(s_i)\right)$

  and $p^d(v|c) \propto \prod_{j\in[N_V]} \exp\left(\phi_j^d(c)^\top V_j(v_j) + B_j(v_j)\right)$ be in exponential family.

  Then a learned LaCIM is **exp.-identified**, if there is a

  (1) *reparameterization* Φ to the gnd.-truth LaCIM, that

  (2) recovers S and V, up to a dimension permutation of each.



Stronger than **sem.-identification**:
v is also identified
(or, s and v are disentangled).

- **Theorem**. A *well-learned* LaCIM is **exp.-identified**, if:

  (a) Additive noise stru. whose **(b)** fn. are bijective and **(c)** noise var. have a.e. nonzero characteristic fn.

  (d) The K_S component fn. of S_i are *lin. indep.*, $\forall i \in [N_S]$ (sim. for each V_j). **(e)** $p^d(c) = \mathrm{Cat}(c|\boxed{\xi^d})$. $\in \Delta^C$

  (f) The D datasets are *diverse enough* s.t.: $\mathrm{rank}\{\xi^d\}_{d\in[D]} = C$, and

  $$\mathrm{rank}\left\{\mathrm{concat}\{\theta_i^d(c) - \theta_i^d(1)\}_{i\in[N_S]}\right\}_{c\in\{2,\dots,C\},d\in[D]} = N_S K_S \text{ (sim. for } \phi_j^d\text{)}.$$

  $$D \geq \max\left\{C, \frac{N_S K_S}{C-1}, \frac{N_V K_V}{C-1}\right\}$$

  - Stronger conclusion than CSG: due to *multi-domain info.*, and *exp. family structure*.
  - Stronger conclusion than iVAE [Khemakhem'20a]: s and v are separated; $p^d(s,v)$ allows a correlation.

Latent Causal Invariant Model (LaCIM)

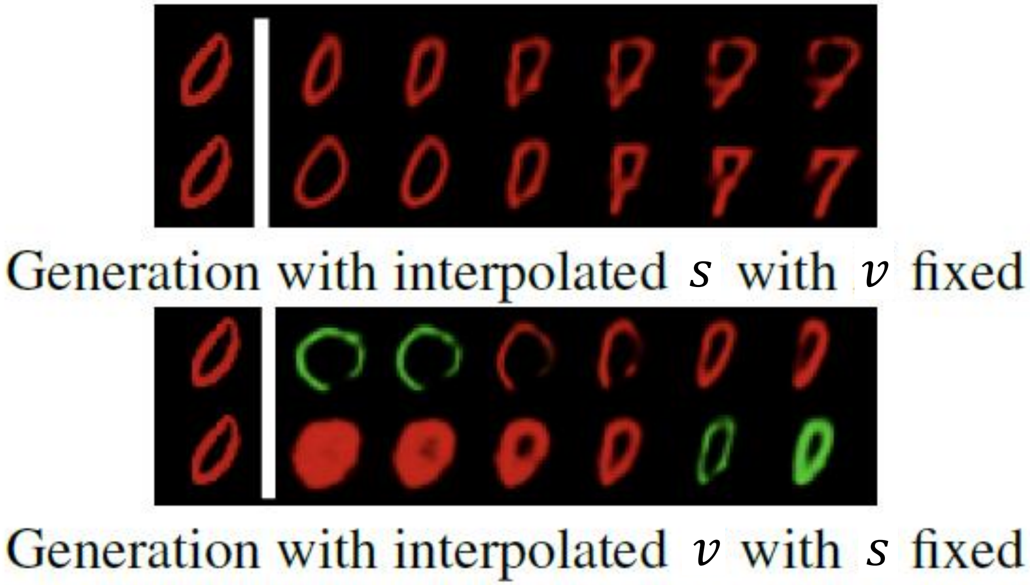
- Experiments: Test-domain accuracy (%)

Method \ Dataset	NICO				CMNIST		ADNI ($D=2$)		
	$D=8$		$D=14$		$D=2$		d: Age	d: TAU	# Params
	ACC	# Params	ACC	# Params	ACC	# Params	ACC	ACC	
CE	60.3 ± 2.8	18.08M	59.3 ± 2.1	18.08M	91.9 ± 0.9	1.12M	62.1 ± 3.2	64.3 ± 1.0	28.27M
DANN	58.9 ± 1.7	19.13M	60.1 ± 2.6	26.49M	84.8 ± 0.7	1.1M	61.0 ± 1.5	65.2 ± 1.1	30.21M
MMD-AAE	60.8 ± 3.4	19.70M	64.8 ± 7.7	19.70M	92.5 ± 0.8	1.23M	60.3 ± 2.2	65.2 ± 1.5	36.68M
DIVA	58.8 ± 3.4	14.86M	58.1 ± 1.4	14.87M	86.1 ± 1.0	1.69M	61.8 ± 1.8	64.8 ± 0.8	33.22M
IRM	61.4 ± 3.8	18.08M	62.8 ± 4.6	18.08M	92.9 ± 1.2	1.12M	62.2 ± 2.6	65.2 ± 1.1	28.27M
LaCIMz	60.4 ± 2.1	18.25M	64.3 ± 1.2	19.70M	93.6 ± 0.9	0.92M	62.7 ± 2.5	66.6 ± 0.8	37.78M
LaCIM (**Ours**)	$\mathbf{63.2 \pm 1.7}$	18.25M	$\mathbf{66.4 \pm 2.2}$	19.70M	$\mathbf{96.6 \pm 0.3}$	0.92M	$\mathbf{63.8 \pm 1.1}$	$\mathbf{67.3 \pm 0.9}$	37.78M

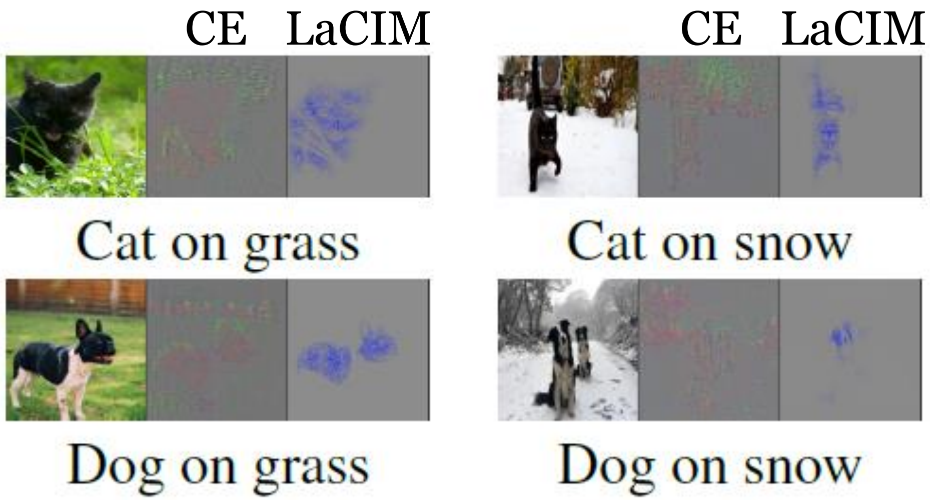
No s-v split

Latent Causal Invariant Model (LaCIM)

- Experiments: Visualization



(a) Visualization on CMNIST



(b) Visualization on NICO

Thanks!

https://arxiv.org/abs/2011.02203

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.
- [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.