Essential Causal Representation learning
via Probability of Sufficient and Necessary Causes

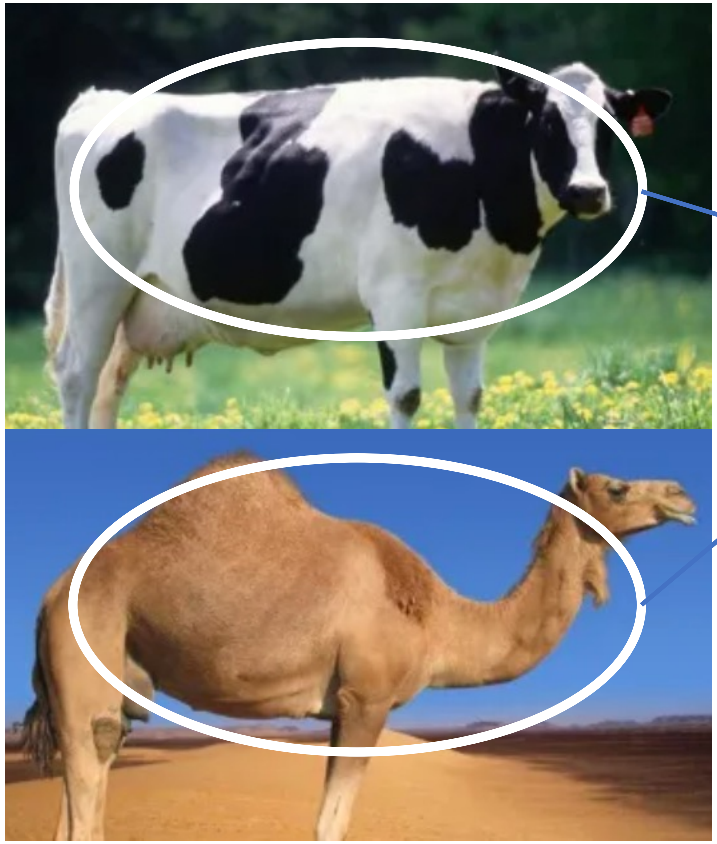
Mengyue Yang

University College London

Email: mengyue.yang.20@ucl.ac.uk

Causal representation

- Capture the causal features of prediction outcomes from high dimensional data.



Train

Avoid failure of generalization

Causal features of prediction outcomes

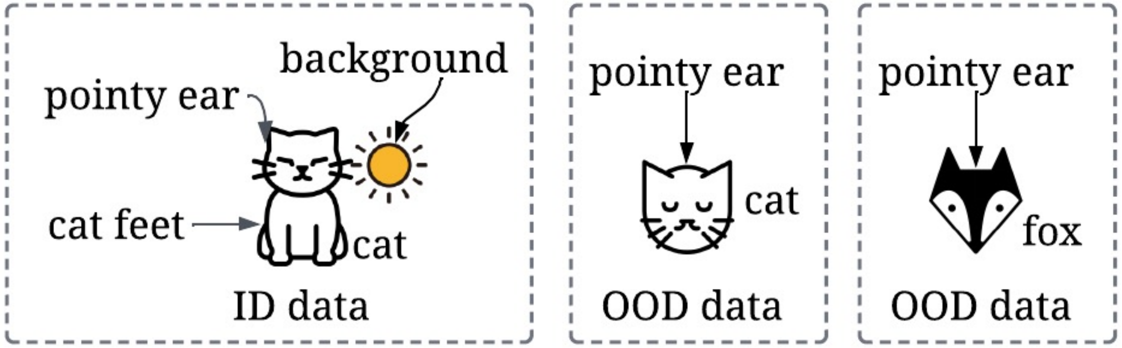
Test

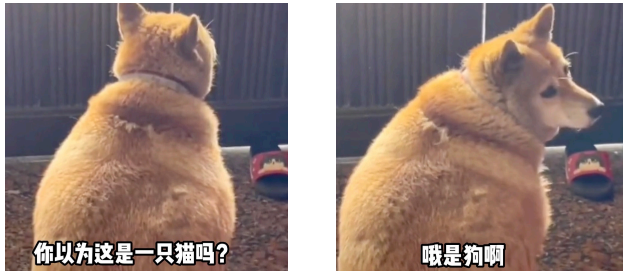
Causal representation

- Is causal feature enough for prediction?
- What kind of causal information is essential?

-- The sufficient and necessary causes!

'pointy ear' is necessary cause, 'cat feet' is sufficient cause
'background' is supurious correlation





Is that a cat? No!

Definition of Sufficiency and Necessity

- **A** is a Sufficient cause of **B** means when we know event **A**, the result **B** will happen.
- **A** is a Necessary cause of **B** means when the result **B** comes out, the event **A** must happened.

Pointy ear is necessary but insufficient

Cat feet is sufficient but unnecessary

Short mouth is sufficient and necessary

	ID train	OOD test	OOD test
pointy ear	True	True	True
cat feet	True	False	False
short mouth	True	False	True
label	Cat	Fox	Cat

Probability of Necessary and Sufficient

- Defining the sufficient and necessary causes.
  - Chapter 9 in book: Causality
  - Considering the counterfactual probability on variables C and Y

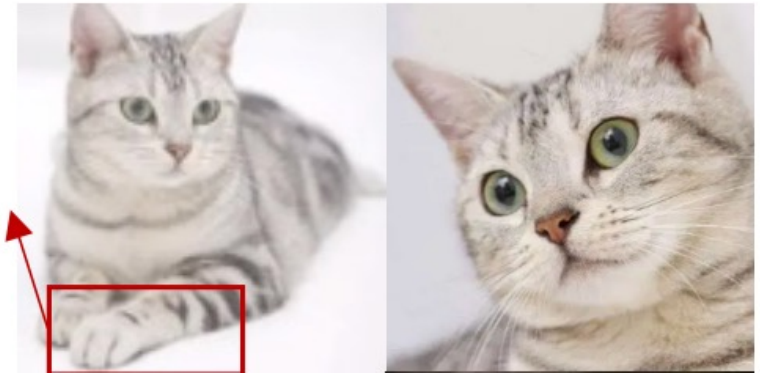
Definition 2.1 (Probability of Necessary and Sufficient (PNS) (Pearl, 2009)). Let the specific implementations of causal variable $\mathbf{C}$ as $\mathbf{c}$ and $\bar{\mathbf{c}}$, where $\bar{\mathbf{c}} \neq \mathbf{c}$. The probability that $\mathbf{C}$ is the necessary and sufficiency cause of Y on test domain $\mathcal{T}$ is

$$\begin{aligned}\mathrm{PNS}(\mathbf{c}, \bar{\mathbf{c}}) := & \underbrace{P_t(Y_{do(\mathbf{C}=\mathbf{c})} = y \mid \mathbf{C} = \bar{\mathbf{c}}, Y \neq y)}_{\text{sufficiency}} P_t(\mathbf{C} = \bar{\mathbf{c}}, Y \neq y) \\ & + \underbrace{P_t(Y_{do(\mathbf{C}=\bar{\mathbf{c}})} \neq y \mid \mathbf{C} = \mathbf{c}, Y = y)}_{\text{necessity}} P_t(\mathbf{C} = \mathbf{c}, Y = y).\end{aligned} \tag{2}$$

Understanding PNS

- Sufficiency

The ‘cat feet’ patch is sufficient but unnecessary



We assume $P(Y_{do(C=1)}=1)=1$ and $P(Y_{do(C=0)}=0)=0.5$, $P(Y=1)=0.75$, $P(C=1,Y=1)=0.5$, $P(C=0,Y=0)=0.25$, $P(C=0,Y=1)=0.25$.

Now, applying the concept of the probability of sufficiency and necessity, we obtain:

Probability of necessity: $P(Y_{do(C=0)}=0|Y=1,C=1)=\frac{P(Y=1)-P(Y_{do(C=0)}=1)}{P(Y=1,C=1)}=\frac{0.5-0.5}{P(Y=1,C=1)}=0$

Probability of sufficiency: $P(Y_{do(\mathbf{C}=1)}=1|Y=0,C=0)=\frac{P(Y_{do(C=1)}=1)-P(Y=1)}{P(Y=0,C=0)}=\frac{1-0.75}{P(Y=1,C=1)}=1$

Understanding PNS

- Necessity

The 'ear shape' patch
is necessary but
insufficient

we assume $P(Y_{do(C=1)} = 1) = 0.5$ and $P(Y_{do(C=0)} = 0) = 1$.

Now, applying the concept of the probability of sufficiency and necessity, we obtain:

Probability of necessity: $P(Y_{do(C=0)} = 0|Y = 1, X = 1) = 1$

Probability of sufficiency: $P(Y_{do(C=1)} = 1|Y = 0, X = 0) = 0.5$

In this example, we can state that variable C has a probability of being a necessary cause.

How to identify PNS from observational data

- Exogeneity : X is the cause of Y
- Monotonicity : Changes on X lead to monotonic changes on Y

Definition 9.2.9 (Exogeneity)

A variable X is said to be exogenous *relative to Y in model M if and only if*

$$\{Y_x, Y_{x'}\} \perp\!\!\!\perp X.$$

Definition 9.2.13 (Monotonicity)

A variable Y is said to be monotonic *relative to variable X in a causal model M if and only if the function $Y_x(u)$ is monotonic in x for all u. Equivalently, Y is monotonic relative to X if and only if*

$$y'_x \wedge y_{x'} = false. \tag{9.20}$$

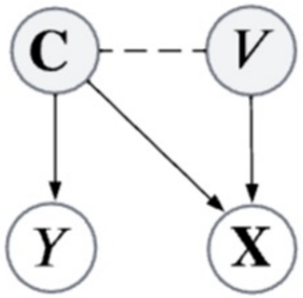
The identifiability results

- Exogeneity : X is the cause of Y
- Monotonicity : Changes on X lead to monotonically changes on Y

Lemma 2.4 (Pearl (2009))**.** *If* $\mathbf{C}$ *is exogenous relative to* Y*, and* Y *is monotonic relative to* $\mathbf{C}$*, then*

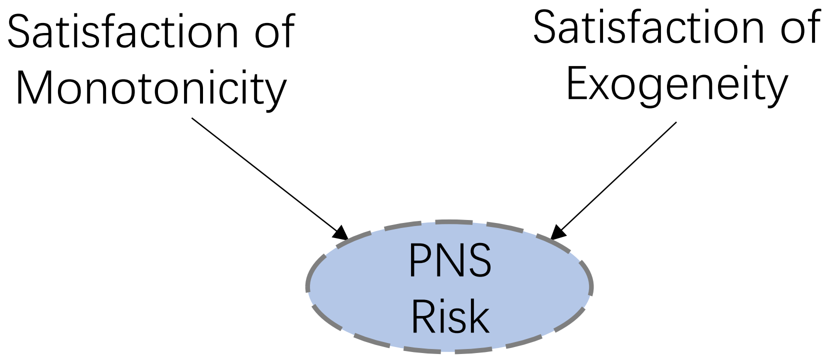
$$PNS(\mathbf{c}, \bar{\mathbf{c}}) = \underbrace{P_t(Y = y|\mathbf{C} = \mathbf{c})}_{\textit{sufficiency}} - \underbrace{P_t(Y = y|\mathbf{C} = \bar{\mathbf{c}})}_{\textit{necessity}} . \quad (3)$$

The PNS risk modeling



- Defining the PNS risk

$$R_t(\mathbf{w}, \phi, \xi) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{T}}\Big[\mathbb{E}_{\mathbf{c}\sim P_t(\mathbf{C}|\mathbf{X}=\mathbf{x})}\mathrm{I}[\mathrm{sign}(\mathbf{w}^\top \mathbf{c}) \neq y] + \mathbb{E}_{\bar{\mathbf{c}}\sim P_t(\bar{\mathbf{C}}|\mathbf{X}=\mathbf{x})}\mathrm{I}[\mathrm{sign}(\mathbf{w}^\top \bar{\mathbf{c}}) = y]\Big].$$



- Defining Monotonicity measurement.

$$M_t^{\mathbf{w}}(\phi, \xi) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{T}}\mathbb{E}_{\mathbf{c}\sim P_t^{\phi}(\mathbf{C}|\mathbf{X}=\mathbf{x})}\mathbb{E}_{\bar{\mathbf{c}}\sim P_t^{\xi}(\bar{\mathbf{C}}|\mathbf{X}=\mathbf{x})}\mathrm{I}[\mathrm{sign}(\mathbf{w}^\top \mathbf{c}) = \mathrm{sign}(\mathbf{w}^\top \bar{\mathbf{c}})],$$

then we have

$$R_t(\mathbf{w}, \phi, \xi) = M_t^{\mathbf{w}}(\phi, \xi) + 2SF_t(\mathbf{w}, \phi)NC_t(\mathbf{w}, \xi) \leq M_t^{\mathbf{w}}(\phi, \xi) + 2SF_t(\mathbf{w}, \phi).$$

Yang, Mengyue, et al. "Invariant Learning via Probability of Sufficient and Necessary Causes." NeurIPS2023 Spotlight.

Satisfaction of Monotonicity

- Connecting the Monotonicity measurement with PNS risk

$$M_t^{\mathbf{w}}(\phi,\xi) = SF_t(\mathbf{w},\phi)(1 - NC_t(\mathbf{w},\xi)) + (1 - SF_t(\mathbf{w},\phi))NC_t(\mathbf{w},\xi). \tag{14}$$

The following equation understands the above decomposition.

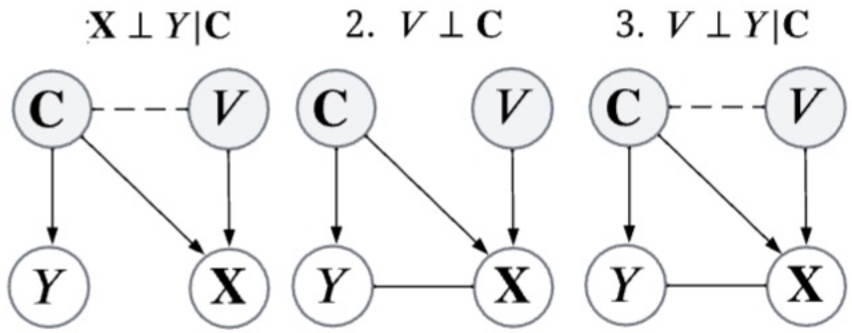
$$\begin{aligned} &P(\text{sign}(\mathbf{w}^\top \mathbf{c}) = \text{sign}(\mathbf{w}^\top \bar{\mathbf{c}})) \\ =&P(\text{sign}(\mathbf{w}^\top \mathbf{c}) = y)P(\text{sign}(\mathbf{w}^\top \bar{\mathbf{c}}) = y) + P(\text{sign}(\mathbf{w}^\top \mathbf{c}) \neq y)P(\text{sign}(\mathbf{w}^\top \bar{\mathbf{c}}) \neq y). \end{aligned} \tag{15}$$

We can further derive Eq.14 as follows.

$$\begin{aligned} M_t^{\mathbf{w}}(\phi,\xi) =& SF_t(\mathbf{w},\phi)(1 - NC_t(\mathbf{w},\xi)) + (1 - SF_t(\mathbf{w},\phi))NC_t(\mathbf{w},\xi) \\ =& \underbrace{SF_t(\mathbf{w},\phi) + NC_t(\mathbf{w},\xi)}_{R_t(\mathbf{w},\phi,T)} - 2SF_t(\mathbf{w},\phi)NC_t(\mathbf{w},\xi) \\ =& R_t(\mathbf{w},\phi,\xi) - 2SF_t(\mathbf{w},\phi)NC_t(\mathbf{w},\xi). \end{aligned} \tag{16}$$

Satisfaction of Exogeneity

- Exogeneity under different causal assumption
  - 1. C contain all information of Y in X
  - 2. There are no spurious correlation between causal information and domain knowledge
  - 3. C contain not all information of Y in X



Yang, Mengyue, et al. "Invariant Learning via Probability of Sufficient and Necessary Causes." NeurIPS2023 Spotlight.

Satisfaction of Exogeneity

- Exogeneity under different causal assumption
  - 1. PNS Risk can directly satisfies exogeneity
  - 2. Additional constraint of independency between V and C like MMD
  - 3. Additional constraint of conditional independence is required like IRM constraint.

Theorem 4.3. *The optimal solution of learned* $\mathbf{C}$ *is obtained by optimizing the following objective (the key part of the objective in Eq. (8))*

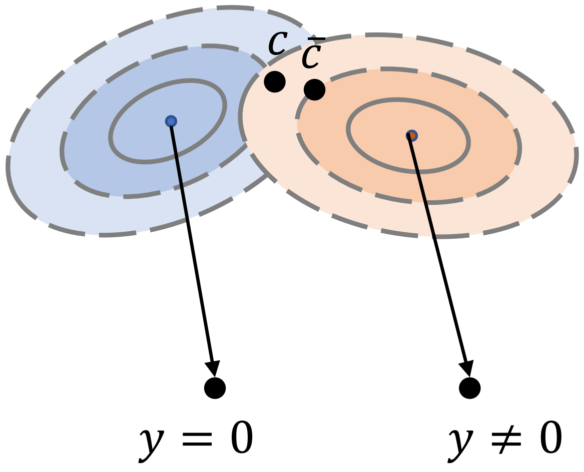
$$\min_{\phi,\mathbf{w}} \widehat{SF}_s(\mathbf{w},\phi) + \lambda \mathbb{E}_{\mathcal{S}^n} \mathrm{KL}(\hat{P}_s^{\phi}(\mathbf{C}|\mathbf{X}=\mathbf{x}) \| \pi_{\mathbf{C}})$$

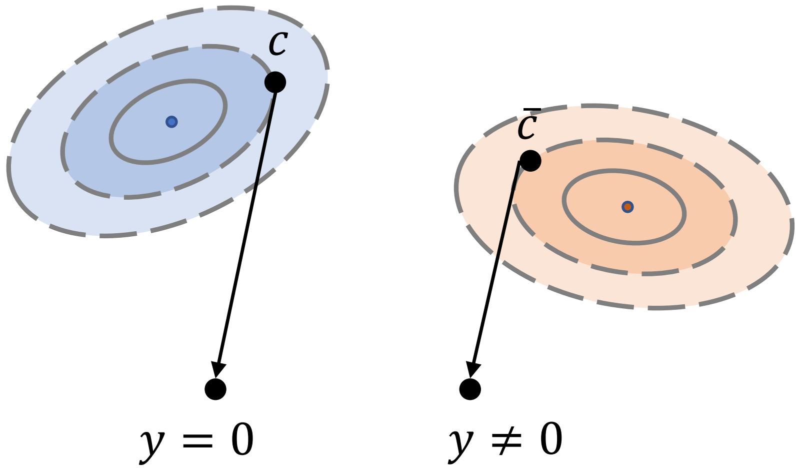
satisfies the conditional independence $\mathbf{X} \perp Y | \mathbf{C}$.

Yang, Mengyue, et al. "Invariant Learning via Probability of Sufficient and Necessary Causes." NeurIPS2023 Spotlight.

Failure case of learning PNS

- In continuous feature space. One problem is that we need to select two value of feature to determine the PNS value.
- A small perturbation on features induce changes on prediction.





Sematic separatable case

The changes of Y is because of the sufficiently changes of C

Failure case

Yang, Mengyue, et al. "Invariant Learning via Probability of Sufficient and Necessary Causes." NeurIPS2023 Spotlight.

Semantic Separability

- Under the case of Sematic separatable, the evaluation of PNS value is non-trivial on feature.

Assumption 4.1 (δ-Semantic Separability). For any domain index $d \in \{s, t\}$, the variable $\mathbf{C}$ is δ-semantic separable, if for any $\mathbf{c} \sim P_d(\mathbf{C}|Y = y)$ and $\bar{\mathbf{c}} \sim P_d(\mathbf{C}|Y \neq y)$, the following inequality holds almost surely: $\|\bar{\mathbf{c}} - \mathbf{c}\|_2 > \delta$.

- When sematic separatable satisfies in data, we add additional constraint on representation.

Yang, Mengyue, et al. "Invariant Learning via Probability of Sufficient and Necessary Causes." NeurIPS2023 Spotlight.

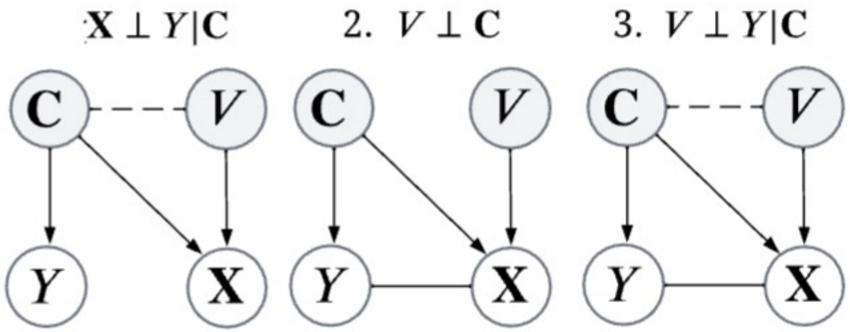
Main objective

- Final objective

Sematic separatable on representation space

$$\min_{\phi,\mathbf{w}} \max_{\xi} \quad \widehat{M_s^{\mathbf{w}}}(\phi,\xi) + \widehat{SF}_s(\mathbf{w},\phi) + \lambda L_{\mathrm{KL}}, \quad \text{subject to} \quad \|\mathbf{c}-\bar{\mathbf{c}}\|_2 > \delta,$$

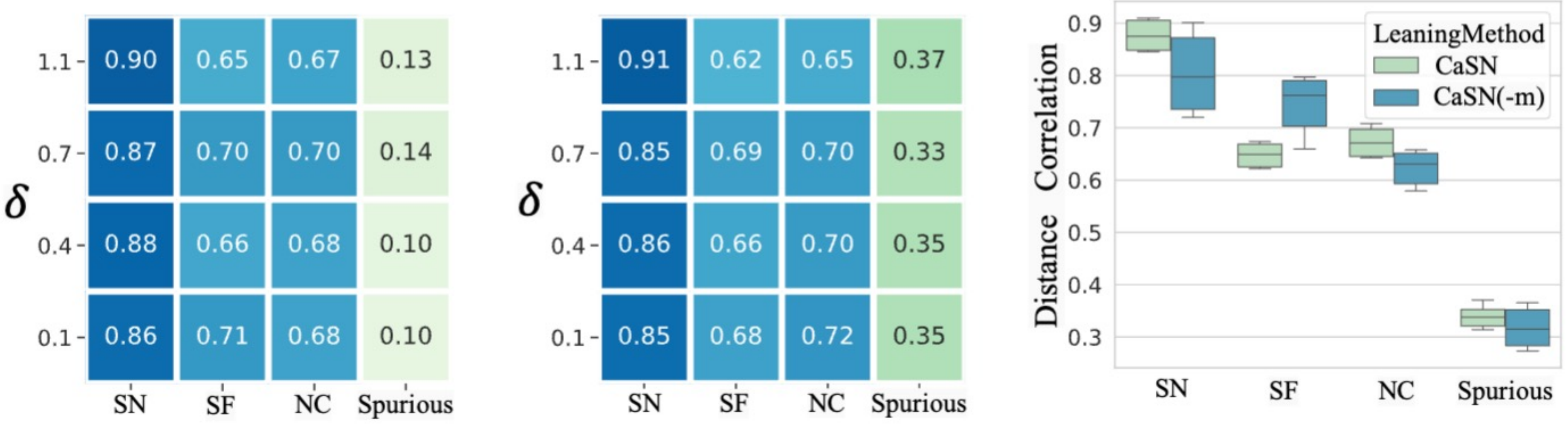
- For different causal assumption we need to add additional constraint



Yang, Mengyue, et al. "Invariant Learning via Probability of Sufficient and Necessary Causes." NeurIPS2023 Spotlight.

Experiment

- Can we learn the sufficient and necessary causes?



(a) Spurious degree $s = 0.1$ (b) Spurious degree $s = 0.7$ (c) Results of CaSN and the CaSN(-m)

Yang, Mengyue, et al. "Invariant Learning via Probability of Sufficient and Necessary Causes." NeurIPS2023 Spotlight.

Experiment

- The OOD generalization ability

Table 1: Results on PACS and VLCS dataset

Dataset	PACS						VLCS					
Algorithm	**A**	**C**	**P**	**S**	**Avg**	**Min**	**C**	**L**	**S**	**V**	**Avg**	**Min**
ERM	84.7 ± 0.4	80.8 ± 0.6	97.2 ± 0.3	79.3 ± 1.0	85.5	79.3	97.7 ± 0.4	64.3 ± 0.9	73.4 ± 0.5	74.6 ± 1.3	77.5	64.3
IRM	84.8 ± 1.3	76.4 ± 1.1	96.7 ± 0.6	76.1 ± 1.0	83.5	76.4	98.6 ± 0.1	64.9 ± 0.9	**73.4 ± 0.6**	**77.3 ± 0.9**	78.5	64.9
GroupDRO	83.5 ± 0.9	79.1 ± 0.6	96.7 ± 0.3	78.3 ± 2.0	84.4	79.1	97.3 ± 0.3	63.4 ± 0.9	69.5 ± 0.8	76.7 ± 0.7	76.7	63.4
Mixup	86.1 ± 0.5	78.9 ± 0.8	**97.6 ± 0.1**	75.8 ± 1.8	84.6	78.9	98.3 ± 0.6	64.8 ± 1.0	72.1 ± 0.5	74.3 ± 0.8	77.4	64.8
MLDG	86.4 ± 0.8	77.4 ± 0.8	97.3 ± 0.4	73.5 ± 2.3	83.6	77.4	97.4 ± 0.2	65.2 ± 0.7	71.0 ± 1.4	75.3 ± 1.0	77.2	65.2
MMD	86.1 ± 1.4	79.4 ± 0.9	96.6 ± 0.2	76.5 ± 0.5	84.6	79.4	97.7 ± 0.1	64.0 ± 1.1	72.8 ± 0.2	75.3 ± 3.3	77.5	64.0
DANN	86.4 ± 0.8	77.4 ± 0.8	97.3 ± 0.4	73.5 ± 2.3	83.6	77.4	**99.0** ± 0.3	65.1 ± 1.4	73.1 ± 0.3	77.2 ± 0.6	**78.6**	65.1
CDANN	84.6 ± 1.8	75.5 ± 0.9	96.8 ± 0.3	73.5 ± 0.6	82.6	75.5	97.1 ± 0.3	65.1 ± 1.2	70.7 ± 0.8	77.1 ± 1.5	77.5	65.1
CaSN (base)	**87.1 ± 0.6**	80.2 ± 0.6	96.2 ± 0.8	80.4 ± 0.2	**86.0**	80.2	97.5 ± 0.6	64.8 ± 1.9	70.2 ± 0.5	76.4 ± 1.7	77.2	64.8
CaSN (irm)	82.1 ± 0.3	77.9 ± 1.8	93.3 ± 0.8	**80.6 ± 1.0**	83.5	77.9	97.8 ± 0.3	65.7 ± 0.8	72.3 ± 0.4	77.0 ± 1.4	78.2	65.7
CaSN (mmd)	84.7 ± 0.1	**81.4 ± 1.2**	95.7 ± 0.2	80.2 ± 0.6	85.5	**81.4**	98.2 ± 0.7	**65.9 ± 0.6**	71.2 ± 0.3	76.9 ± 0.7	78.1	**65.9**

Yang, Mengyue, et al. "Invariant Learning via Probability of Sufficient and Necessary Causes." NeurIPS2023 Spotlight.

Application/Future work

- The scenario which need stable prediction.
  - Autonomous driving.
  - Adversarial attack.
  - Domain adaptation/generalization.

- Future work
  - More causal assumption
  - More general case

Yang, Mengyue, et al. "Invariant Learning via Probability of Sufficient and Necessary Causes." NeurIPS2023 Spotlight.