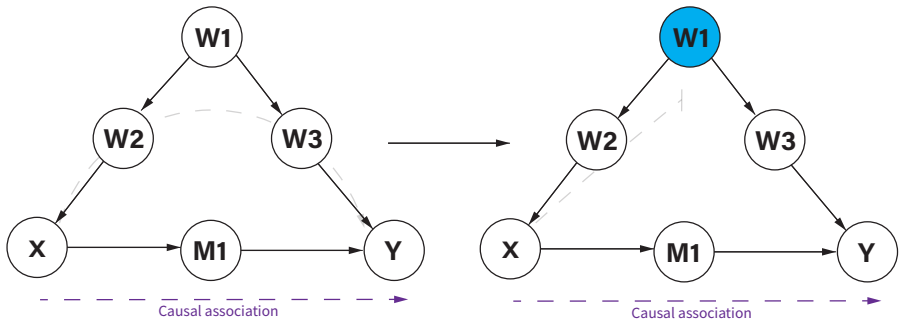
Causal Inference and Causal Representation Learning in Vision: A Survey

Changjie Lu

Wenzhou-Kean University



Motivation

- Dilemma on machine learning research
  - Catastrophe on predicting o.o.d data
  - Confounder in data
- Reflection on my previous research
  - Image deblurring: various blur
  - Image deraining: complex rain streak
  - UDA-VAE++: lesion detection
- Medical Image Analysis
  - The gap between research and clinical application
  - High risk tasks require the explainable algorithm

Content

- Concept of Causal Inference
- Causal Representation Learning
  - Theoretical Works
    - Invariant Risk Minimization
  - Practical Works
    - Feature Understanding
    - Transfer Learning
- Future Work on Medical Image Analysis
  - Attention Mechanism
  - Deconfounded Optimizer
  - Domain Adaptation/Generalization
  - Feature Disentanglement
  - Class Imbalance Learning

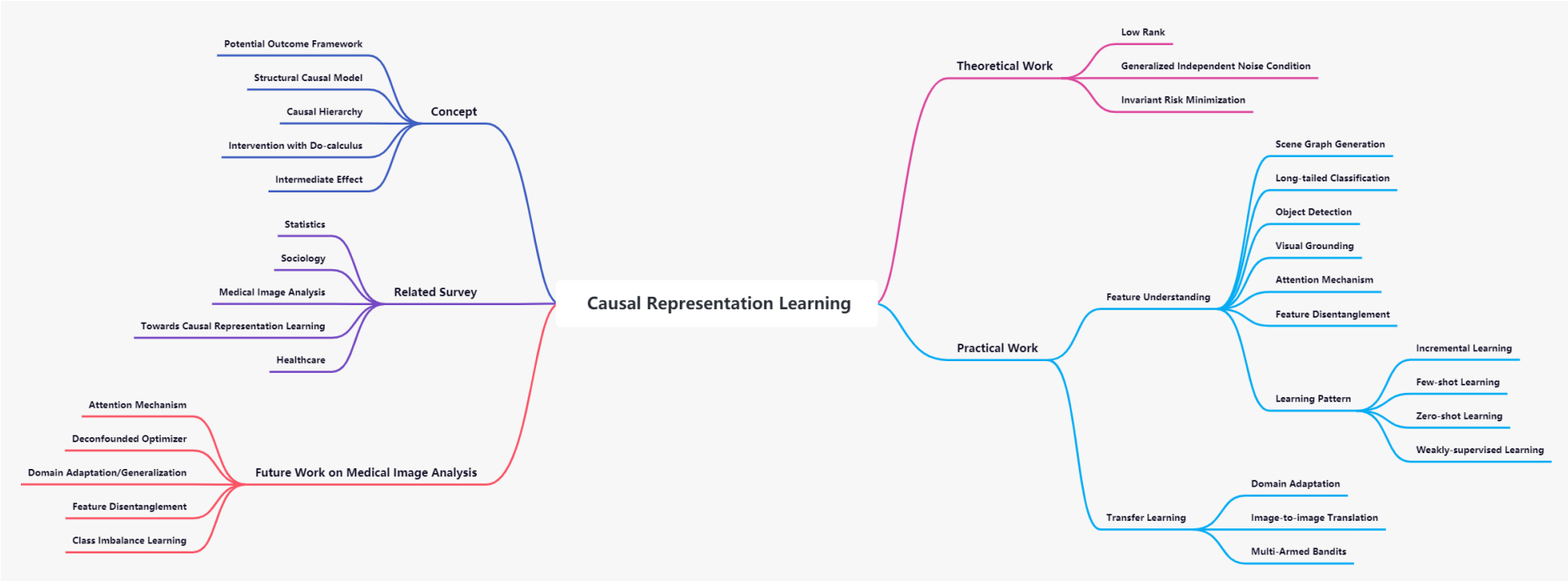


Figure 1: The mindmap of causal representation learning in vision

Introduction

The **correlation** does not imply **causation** [Geer Jr, 2011].

	Condition		
Treatment	Mild	Severe	Total
A	15% (210/1400)	30% (30/100)	**16% (240/1500)**
B	10% (5/50)	20% (100/500)	19% (105/550)

Table 1: The example of Simpson's paradox[Wagner, 1982]. Here are A/B treatments for patients. 16% donates the 16% death rate after the treatment.

Traditional Solution

- Randomized controlled trial (RCT) -> mechanism unclear
- Observational data -> limited in correction
- Physical modeling -> complex

Advanced Solution

Pearl introduced a Structural causal model (SCM) with the three-layer causal hierarchy.[Pearl, 2009]

- Structural equation models (SEM)
- Potential outcome framework
- Directed acyclic graphs (DAG)
- Do-calculus[Pearl, 1994]

Related works on causal inference

- Statistics[Pearl, 2009]
- Sociology[Gangl, 2010]
- Causal Representation Learning[Schölkopf et al., 2021]
- Healthcare[Zhang et al., 2021c]
- Deep Learning Causal Discovery[Chen et al., 2022a]
- Medical Image Analysis[Vlontzos et al., 2022]

Concept of causal inference

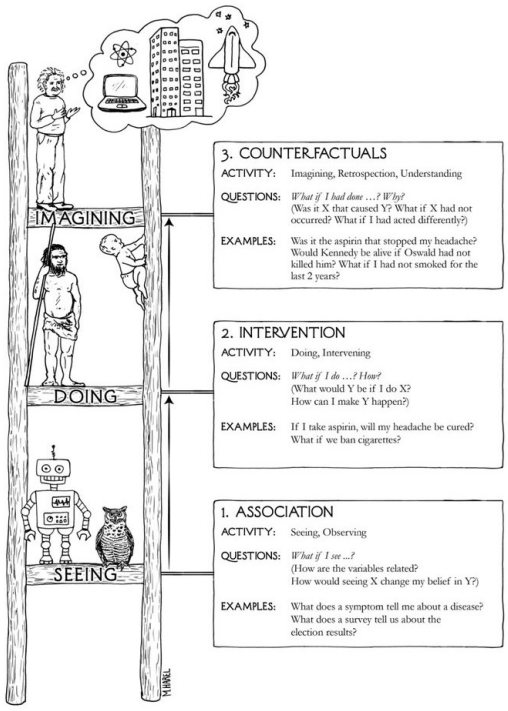


Figure 2: Pearl Causal Hierarchy. (Drawing by Maayan Harel[Harel, 2019])

Zero-shot Learning, Long-tailed Classification

Feature learning, Few-shot learning

Supervised learning, Unsupervised learning

Structural Causal Model

The structural causal model[Pearl, 2009] is a 4-tuple $\langle \mathrm{U}, \mathrm{V}, \mathcal{F}, \mathrm{P}(\mathrm{U}) \rangle$, where:

- U: background variables (exogenous variables)
- V: endogenous variables (determined by other variables within the model)
- $\mathcal{F}$: functions mapping from U to V.
- $\mathrm{P}(\mathrm{U})$ is a probability function defined over the domain of U.

Example

X: Treatment solution, Y: lung function of asthma patient, Z: Air pollution.
X and Y are endogenous variable, Z is exogenous variables.

$$
\begin{aligned}
&U = \{Z, U_x, U_y\}, V = \{X, Y\}, F = \{f_X, f_Y\} \\
&f_X : X \leftarrow f_X(U_x) \\
&f_Y : Y \leftarrow f_Y(X, Z, U_y)
\end{aligned}
\tag{1}
$$

Causal Hierarchy

- Seeing

$$P^{\mathcal{M}}(Y = y) = \sum_{\{u|Y(u)=y\}} P(u) \tag{2}$$

- Doing

$$P^{\mathcal{M}}(Y_x = y_x) = \sum_{\{u|Y_x(u)=y_x\}} P(u) \tag{3}$$

- Imaging

$$P\left(Y'_{x'} \mid X = x, Y = y\right) \tag{4}$$

Intervention with do-calculus
D-separation

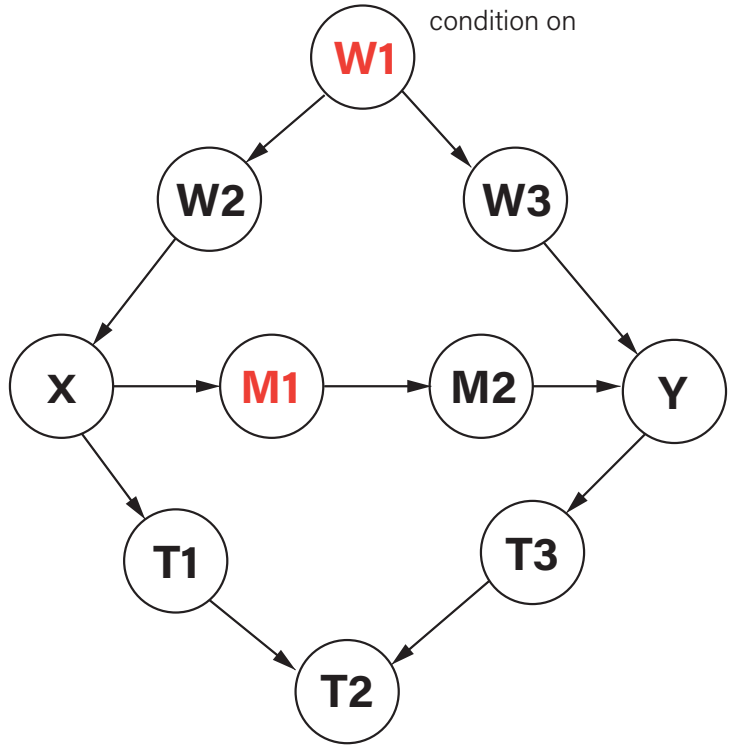


Figure 3: The demonstration for the d-separation. W1 and M1 are conditioned, which block all possible way from X $\rightarrow$ Y. Therefore, X,Y are d-separated.

Intervention

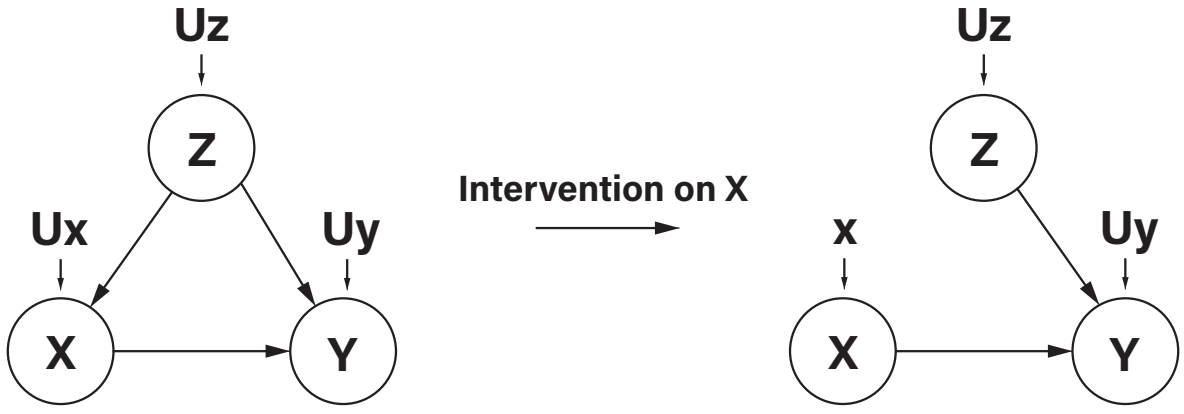


Figure 4: After the intervention on X = x, the edge from Z → X should be deleted.

Adjustment formula:

$$P(Y = y|do(X = x)) = \sum_{z} P(Y = y|X = x, Z = z)P(Z = z) \tag{5}$$

	Condition			
Treatment	Mild	Severe	Total	Causal
A	15% (210/1400)	30% (30/100)	**16% (240/1500)**	19.4%
B	10% (5/50)	20% (100/500)	19% (105/550)	**12.9%**

$X = A/B$ represents patients take the drug A/B. $Z = \text{Mild/Severe}$ represents the level of illness. Y represents the dead rate.
The effect of taking the drug A:

$$E[Y|do(X = A)] = \frac{1450}{2050}(0.15) + \frac{600}{2050}(0.30) \approx 0.194 \tag{6}$$

The effect of taking the drug B:

$$E[Y|do(X = B)] = \frac{1450}{2050}(0.10) + \frac{600}{2050}(0.20) \approx 0.129 \tag{7}$$

Backdoor adjustment

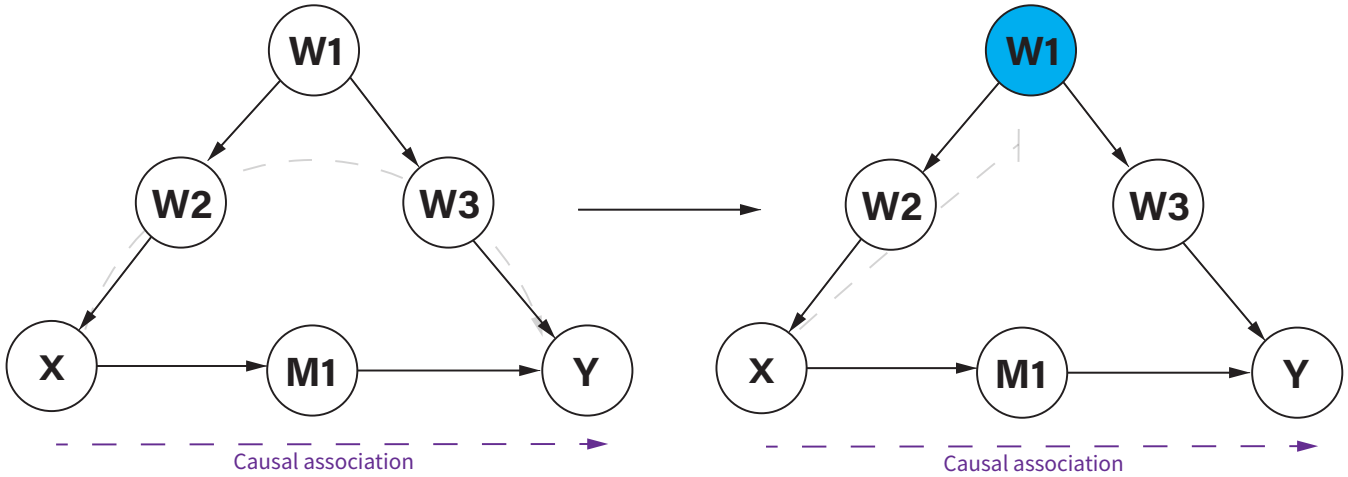


Figure 5: Second equation in Equ.8: W is a sufficient adjustment set, blocking all backdoor path, only reserving the causation $X \rightarrow Y$. Third equation: $do(X)$ blocks all T's parents.

$$
\begin{aligned}
P(y \mid do(X)) &= \sum_{w} P(y \mid do(X), w) P(w \mid do(X)) \\
&= \sum_{w} P(y \mid X, w) P(w \mid do(X)) \\
&= \sum_{w} P(y \mid X, w) P(w)
\end{aligned}
\tag{8}
$$

Frontdoor Adjustment

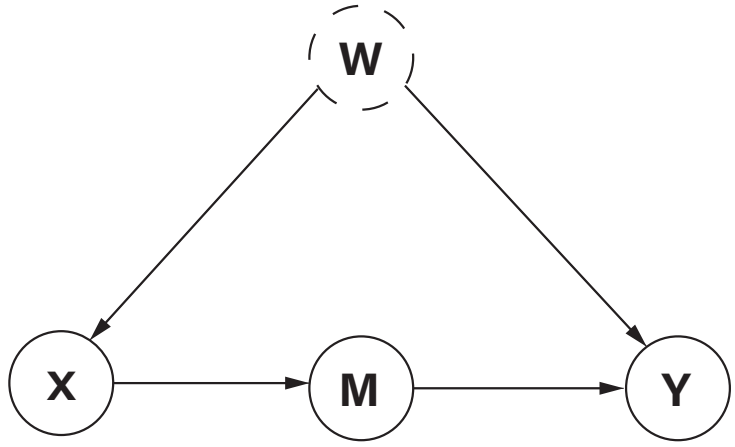


Figure 6: Typically, we do twice backdoor adjustment. Firstly, from $X \to M$, there is no backdoor path. Secondly, from $M \to Y$, X block the backdoor path $M \leftarrow X \leftarrow W \to Y$.

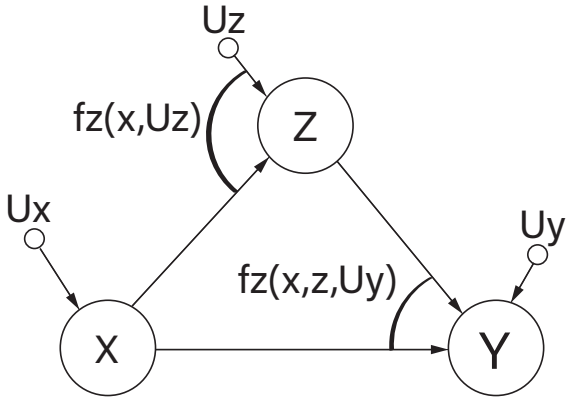
- X on M: $P(m \mid do(x)) = P(m \mid x)$
- M on Y: $P(y \mid do(m)) = \sum_x P(y \mid m, x)P(x)$
- X on Y: $P(y \mid do(x)) = \sum_m P(m \mid do(x))P(y \mid do(m)) = \sum_m P(m \mid x) \sum_{x'} P(y \mid m, x')P(x')$

Intermediate Effect

In a causal model, a classical intermediate problem can be defined as:

$$x = f_X(U_X), z = f_Z(x, U_Z), y = f_Y(x, z, U_Y) \tag{9}$$

where X donates treatment, Z donates mediator, Y donates outcome. f_X, f_Z, f_Y are any function. U_X, U_Z, U_Y are background variables.



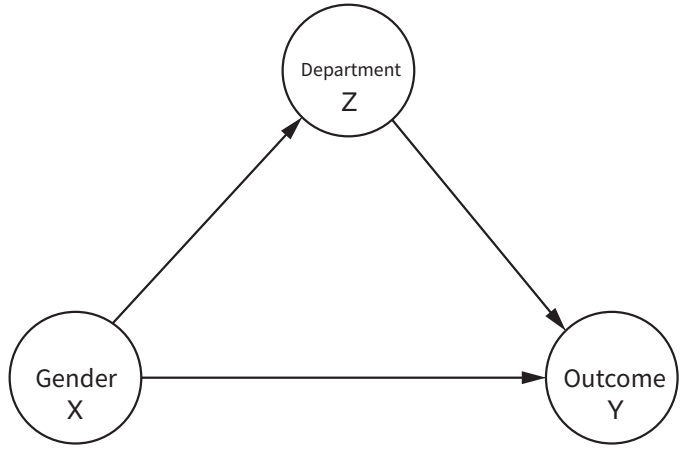


Figure 7: Example of intermediate effect

Treatment Effect

- Average Treatment Effect:

$$\begin{aligned} \mathrm{ATE} &= \mathrm{E}[Y_1] - \mathrm{E}[Y_0] \\ &= \mathrm{E}[Y \mid \mathrm{do}(X = 1)] - \mathrm{E}[Y \mid \mathrm{do}(X = 0)]] \end{aligned} \tag{10}$$

- Controlled Direct Effect:

$$\begin{aligned} \mathrm{CDE}(z) &= \mathrm{E}\left[Y_{1,z} - Y_{0,z}\right] \\ &= \mathrm{E}[Y \mid \mathrm{do}(X = 1, Z = z)] - \mathrm{E}[Y \mid \mathrm{do}(X = 0, Z = z)] \end{aligned} \tag{11}$$

- Natural Direct Effect:

$$\mathrm{NDE} = \mathrm{E}\left[Y_{1,Z_0} - Y_{0,Z_0}\right] \tag{12}$$

- Natural Indirect Effect:

$$\mathrm{NIE} = \mathrm{E}\left[Y_{0,Z_1} - Y_{0,Z_0}\right] \tag{13}$$

- Total Direct Effect:

$$\mathrm{TDE} = \mathrm{E}\left[Y_{1,Z_1} - Y_{0,Z_1}\right] \tag{14}$$

- Total Indirect Effect:

$$\mathrm{TIE} = \mathrm{E}\left[Y_{1,Z_1} - Y_{1,Z_0}\right] \tag{15}$$

Theoretical works

Invariant Risk Minimization[Arjovsky et al., 2019]
Consider a data representation $\Phi : \mathcal{X} \to \mathcal{H}$. we desire an invariant predictor across the environment $\mathrm{w} \circ \Phi$. If there exists a predictor $\mathrm{w} : \mathcal{H} \to \mathcal{Y}$ achieve optimal performance in every environment $\mathcal{E}_{\mathrm{tr}}$ such that $\mathrm{w} \in \underset{\bar{\mathrm{w}}:\mathcal{H}\to\mathcal{Y}}{\arg\min} \mathrm{R}^{\mathrm{e}}(\bar{\mathrm{w}} \circ \Phi)$.
Therefore, the constrained optimization problem is defined as:

$$\begin{array}{ll} \min\limits_{\substack{\Phi:\mathcal{X}\to\mathcal{H} \\ \mathrm{w}:\mathcal{H}\to\mathcal{Y}}} & \sum_{\mathrm{e}\in\mathcal{E}_{\mathrm{tr}}} \mathrm{R}^{\mathrm{e}}(\mathrm{w} \circ \Phi) \\ \text{subject to} & \mathrm{w} \in \underset{\bar{\mathrm{w}}:\mathcal{H}\to\mathcal{Y}}{\arg\min} \mathrm{R}^{\mathrm{e}}(\bar{\mathrm{w}} \circ \Phi), \text{ for all } \mathrm{e} \in \mathcal{E}_{\mathrm{tr}} \end{array} \tag{16}$$

For gradient descent method, the loss function is defined as:

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{tr}} R^e(\Phi) + \lambda \cdot \left\|\nabla_{w|w=1.0} R^e(w\cdot\Phi)\right\|^2 \tag{17}$$

where D measures the risk when changing the environment, $\lambda \in [0, \infty)$ is a hyper-parameter balancing the ERM and the IRM.

Algorithm	**Acc. train envs.**	**Acc. test env.**
ERM	87.4 ± 0.2	17.1 ± 0.6
IRM (ours)	70.8 ± 0.9	$\mathbf{66.9 \pm 2.5}$

```
import torch
from torch.autograd import grad

def compute_penalty(losses, dummy_w):
    g1 = grad(losses[0::2].mean(), dummy_w, create_graph=True)[0]
    g2 = grad(losses[1::2].mean(), dummy_w, create_graph=True)[0]
    return (g1 * g2).sum()

def example_1(n=10000, d=2, env=1):
    x = torch.randn(n, d) * env
    y = x + torch.randn(n, d) * env
    z = y + torch.randn(n, d)
    return torch.cat((x, z), 1), y.sum(1, keepdim=True)

phi = torch.nn.Parameter(torch.ones(4, 1))
dummy_w = torch.nn.Parameter(torch.Tensor([1.0]))

opt = torch.optim.SGD([phi], lr=1e-3)
mse = torch.nn.MSELoss(reduction="none")

environments = [example_1(env=0.1),
                example_1(env=1.0)]

for iteration in range(50000):
    error = 0
    penalty = 0
    for x_e, y_e in environments:
        p = torch.randperm(len(x_e))
        error_e = mse(x_e[p] @ phi * dummy_w, y_e[p])
        penalty += compute_penalty(error_e, dummy_w)
        error += error_e.mean()

    opt.zero_grad()
    (1e-5 * error + penalty).backward()
    opt.step()

    if iteration % 1000 == 0:
        print(phi)
```

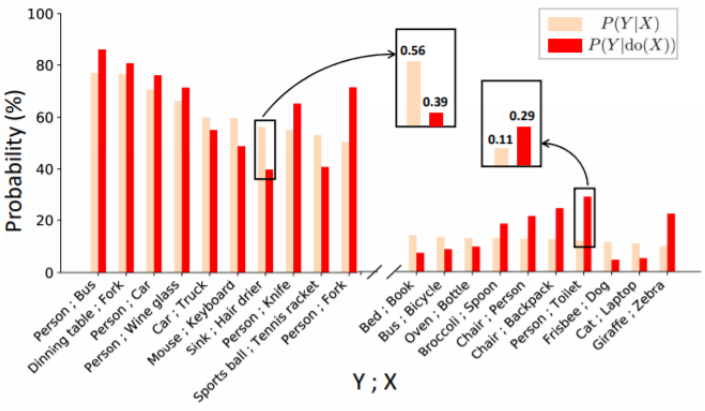
A series of works beyong IRM

- [Rosenfeld et al., 2020] proved that IRM can not find the optimal invariant predictor on most occasions and introduced and analyzed the **non-linear scene**.
- [Kamath et al., 2021] deeply discussed the limitation of the IRM and proved that the IRM prefers an invariant predictor with worse o.o.d generalization.
- [Lin et al., 2022] introduced **Bayesian inference** to diminish the over-fitting problem
- [Zhou et al., 2022] proposed **sparse** IRM prevent the spurious correlation leaking to sub models.
- [Liu et al., 2021a] proposed a Heterogeneous Risk Minimization (HRM) structure to address the mixture of multi-source data without a source label.
- [Liu et al., 2021b] extended to kernel space, enhancing the ability to deal with more complex data and invariant relationship.
- [Zhang et al., 2021a] tried to modify the parameters to make the model robust when domain shifting based on mate-learning structure.
- [Li et al., 2022] combined information bottleneck with IRM for domain adaptation.

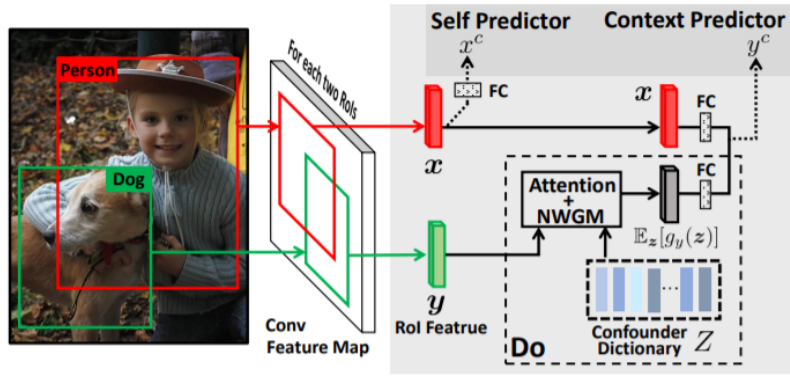
Practical works

Feature understanding

VC R-CNN[Wang et al., 2020]

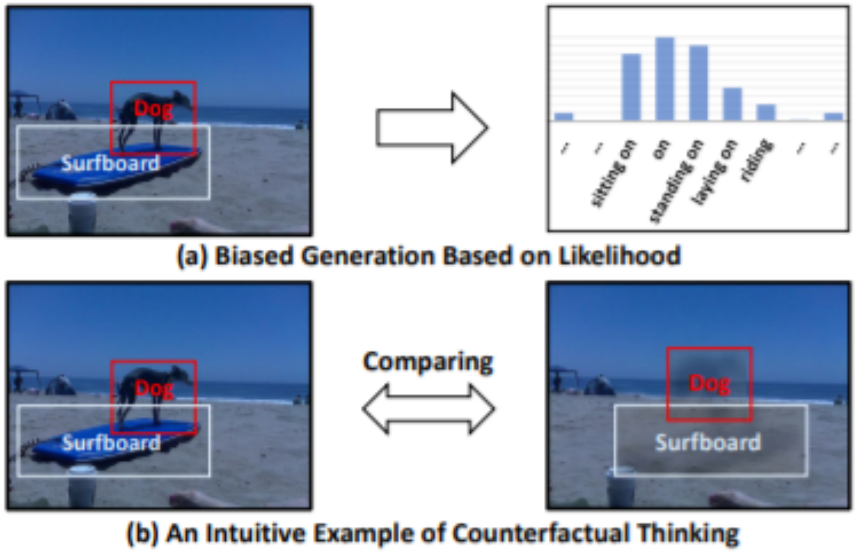


Observational bias made the model ignore the common causal relationship

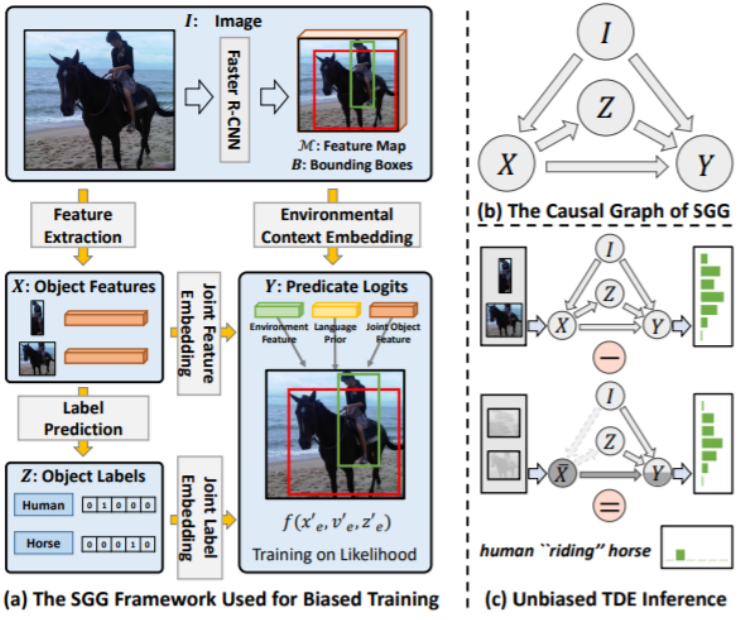


They introduced an intervention (confounder dictionary) to measure the true causal effect.

Scene graph generation[Tang et al., 2020b]

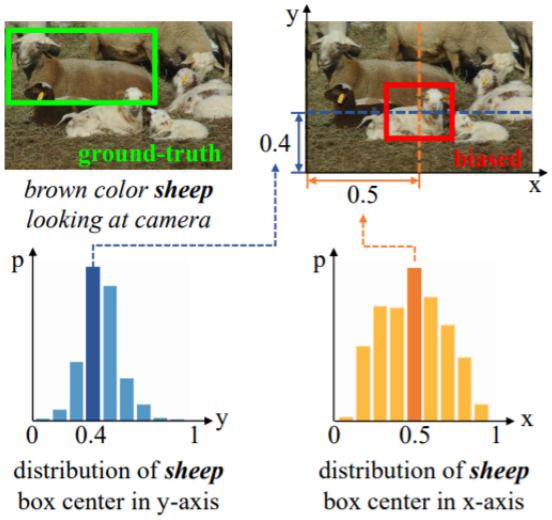


Compared the counterfactual scene and factual scene

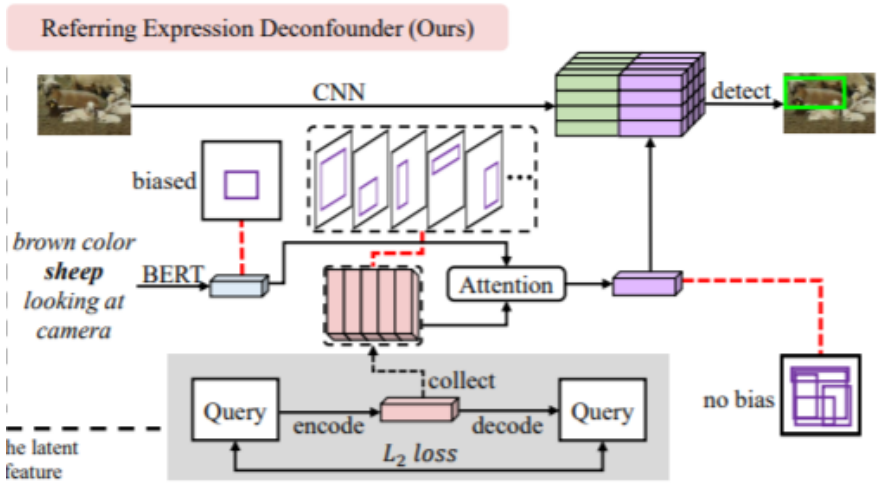


Total Direct Effect (TDE) analysis framework to remove the bias in training

Deconfounded visual grounding[Huang et al., 2022]

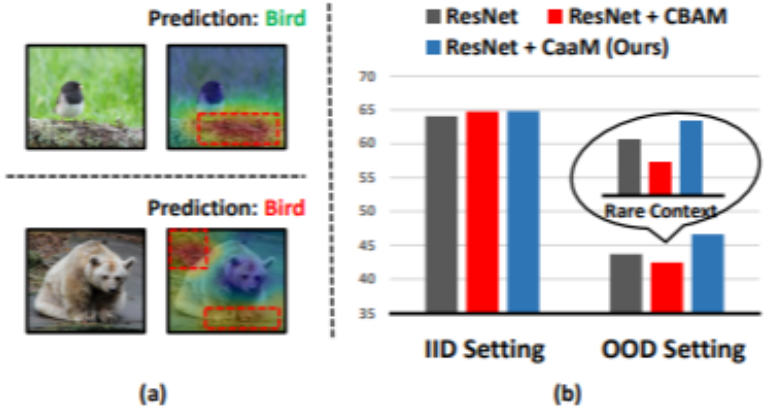


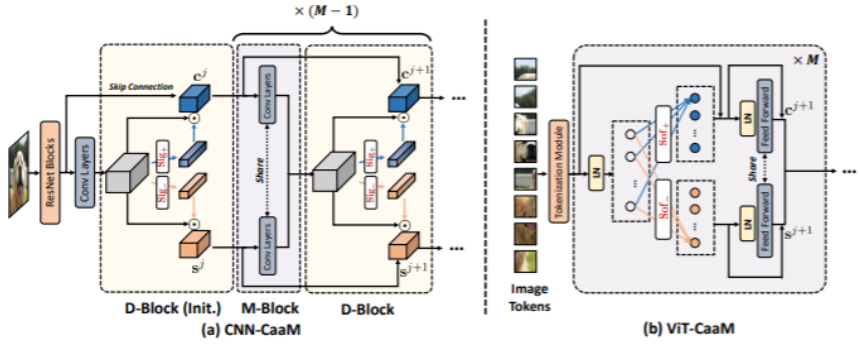
The location of the target bounding box highly depends on the query



A plug-and-play framework Referring Expression Deconfounder (RED), to make the backdoor adjustment to find the causal relationship between images and sentences.

Attention mechanism[Wang et al., 2021b]

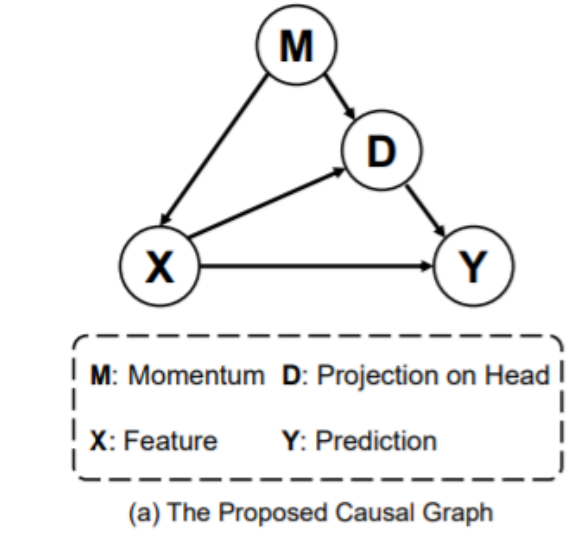




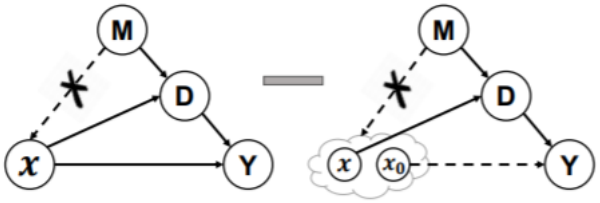
The o.o.d data degrades the performance when using attention.

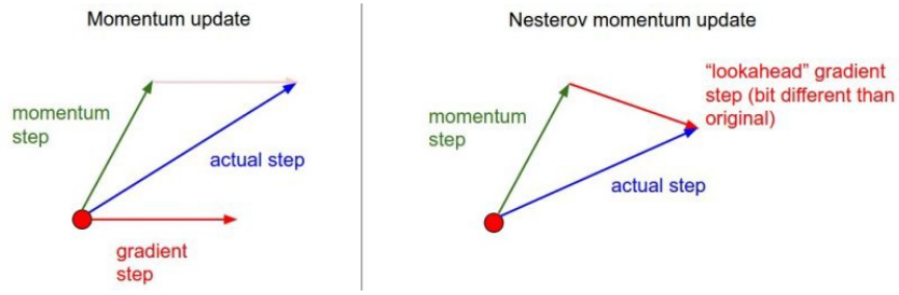
This method utilized the IRM and adversarial training with a partitioned dataset to discover confounding factor characteristics.

Long-tailed classification[Tang et al., 2020a]



(a) The Proposed Causal Graph

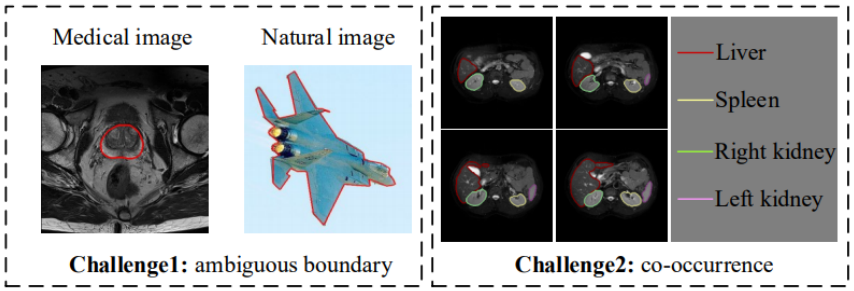


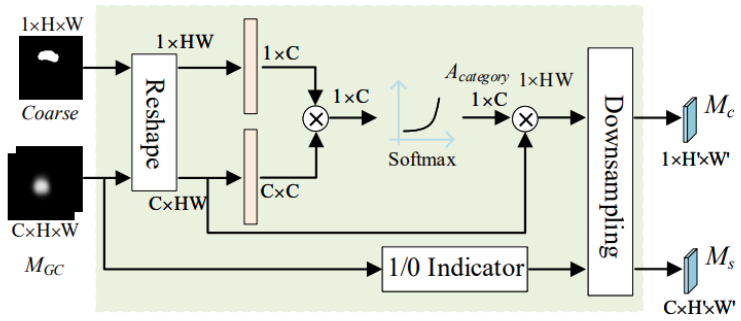


The update direction of SGD[a] contains the information on data distribution.
Solution: multi-head normalized classifier in training + counterfactual TDE inference in testing

[a]https://cs231n.github.io/neural-networks-3/

Weakly-supervised learning on medical image[Chen et al., 2022b]





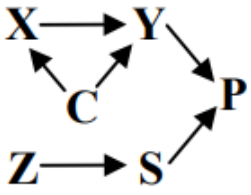
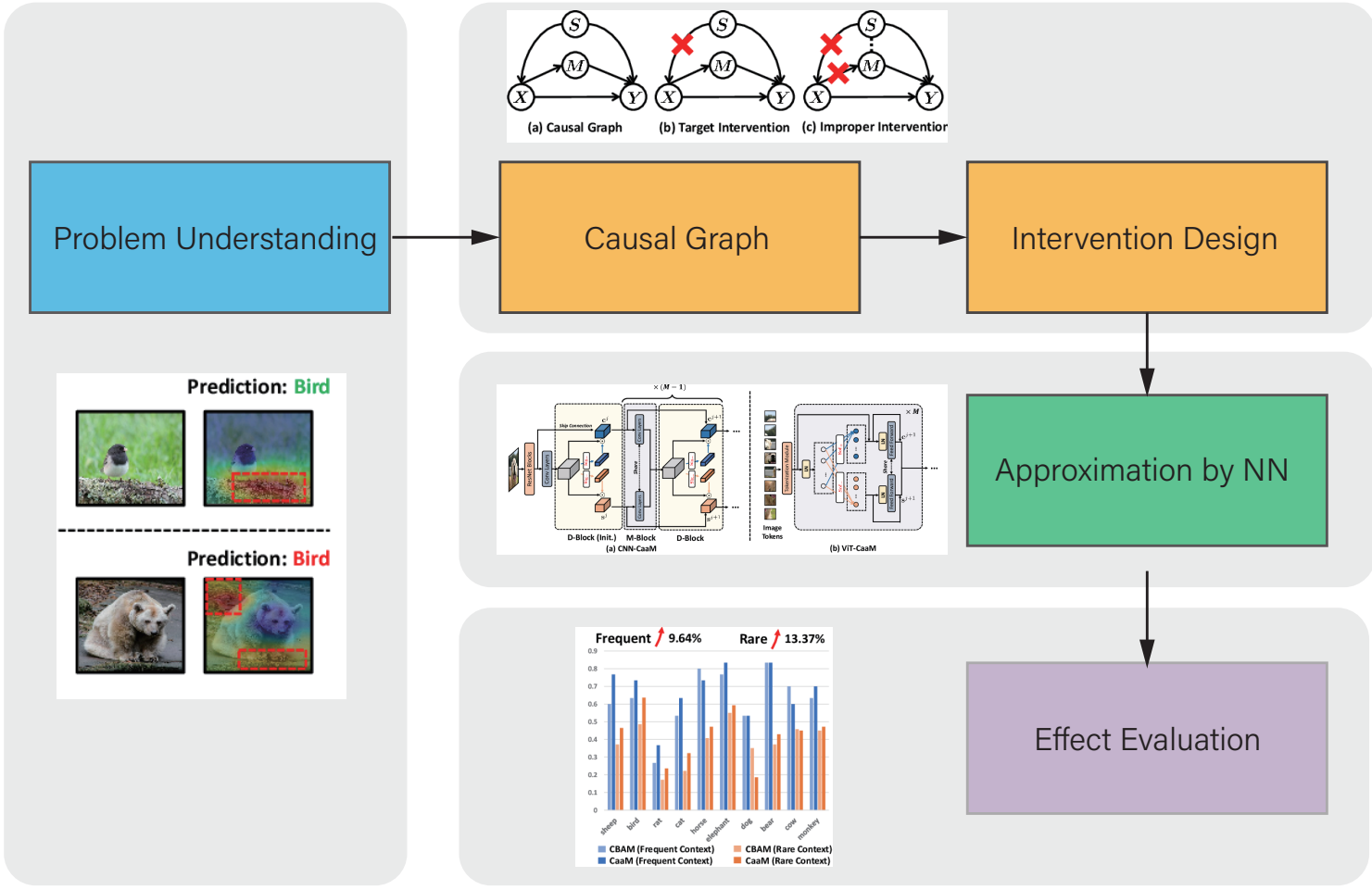


Figure 5. The causal graph of medical image WSSS. X denotes medical image, Y denotes classified category, C denotes context confounder. Z denotes anatomical structure, S denotes shape of segmentation and P denotes pseudo mask.

Solution:
A category causality chain for ambiguous boundary
A anatomy-causality chain for co-occurrence problem

Task	Performance Improvement over 2nd Model		
Scene Graph Generation [39]	Predicate Classification 51.1% Zero-Shot Relationship Retrieval 25.0%	Scene Graph Classification 56.4% Sentence-to-Graph Retrieval 33.7%	Scene Graph Detection 31.3%
Image Captioning [43]	Karpathy Split [64] 5 Captions 0.250%	Karpathy Split Whole Set 1.55%	MS-COCO [65] 1.02%
Attention Mechanism [45]	CNN-Based NICO [66] 8.72% ViT-Based NICO 8.08%	CNN-Based ImageNet-9 [67] 1.15% ViT-Based ImageNet-9 2.78%	CNN-Based ImageNet-A 8.33% ViT-Based ImageNet-A 12.7%
Few-shot Learning [49]	miniImageNet 2.40%	tieredImageNet 0.94%	
Long-tailed Classification [48]	LVIS V1.0 [68] val set 29.2% LVIS V0.5 eval test server 18.8%	ImageNet-LT 17.3%	LVIS V0.5 val set 19.1%
Incremental Learning [47]	CIFAR-100 [69] 6.17%	ImageNet-Sub 4.76%	ImageNet-Full 3.49%
Image Recognition [38]	MS-COCO 1.41%	Open Images [70] 1.09%	VQA2.0 test [71] 0.560%
Visual Grounding [40]	RefCOCO [72] 2.09% ReferIt Game 1.14%	RefCOCO+ 2.36% Flickr30K Entities [73] 0.36%	RefCOCOg 1.77%
Weakly-Supervised Learning [50]	PASCAL VOC 2012 [74] 1.84%	MSCOCO 2.45%	

Workflow


(a) Causal Graph
(b) Target Intervention
(c) Improper Intervention
Problem Understanding
Causal Graph
Intervention Design
Prediction: Bird
Prediction: Bird
D-Block (Init.)
M-Block
D-Block
(a) CNN-CaaM
(b) ViT-CaaM
Approximation by NN
Frequent 9.64%
Rare 13.37%
CBAM (Frequent Context)
CaaM (Frequent Context)
CBAM (Rare Context)
CaaM (Rare Context)
Effect Evaluation

Transfer learning

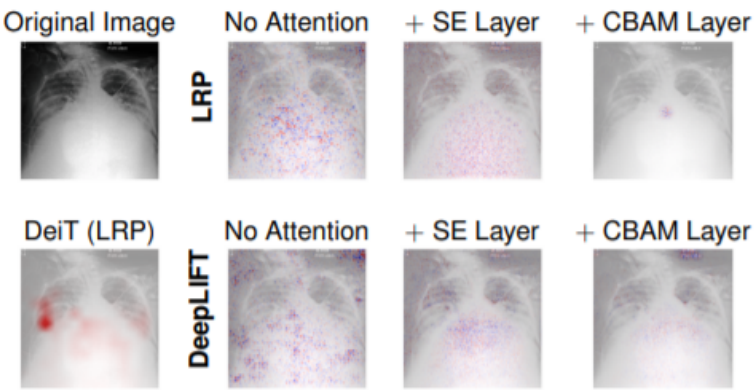
- Domain adaptation: [Subbaswamy and Saria, 2020] defined a invariance specification.
  [Zhang et al., 2020a] considered domain adaptation as an inference problem, constructing a DAG and solving it by Bayesian inference.
- Image-to-image translation: [Xie et al., 2021] concluded a DAG and developed an important reweighting-based learning method.
- General field:
  - Estimating heterogeneous treatment effects [Künzel et al., 2018]
  - Invariant models [Rojas-Carulla et al., 2018]
  - Multi-Armed Bandits in reinforcement learning [Zhang and Bareinboim, 2017]

Future Work on Medical Image Analysis

- Attention mechanism
- Deconfounded optimizer
- Domain adaptation/generalization
- Feature disentanglement
- Class imbalance learning

Attention mechanism

Attention is widely used in medical image analysis.
However, the performance is decreased when we apply it on o.o.d data.
”The results are very noisy, even for the cases of attention mechanisms, which is counter-intuitive.” —— A survey on **attention** in MIA[Gonçalves et al., 2022]



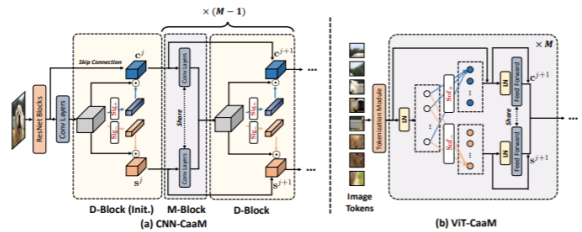


Figure 8: Causal attention module (Caam)[Wang et al., 2021b]

Deconfounded optimizer

MIA models are sensitive to the optimizer, especially in domain adaptation/generalization.

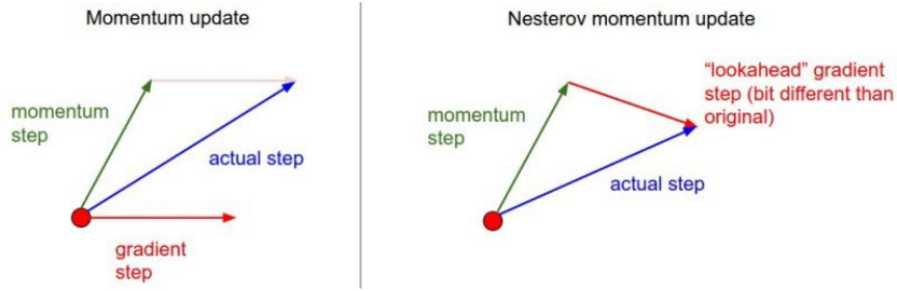


Figure 9: Deconfound Optimizer

Domain adaptation/generalization

Domain adaptation[Wu and Zhuang, 2020] [Chen et al., 2020]
Domain generalization[Liu et al., 2021c][Zhang et al., 2020b]
[Li et al., 2020][Li et al., 2021] techniques greatly improved the accuracy (from $\approx 10\%$ to $\approx 70\%$).
we could try to find an invariant representation among different sources.

$$\mathrm{L}_{\mathrm{IRM}}(\Phi, \mathrm{w}) = \sum_{\mathrm{e} \in \mathcal{E}_{\mathrm{tr}}} \mathrm{R}^{\mathrm{e}}(\mathrm{w} \circ \Phi) + \lambda \cdot \mathrm{D}(\mathrm{w}, \Phi, \mathrm{e}) \tag{18}$$

Feature disentanglement

The restriction for disentanglement is still loose and lacks interpretability in current work. [Wang et al., 2021a] introduced Iterative Partition-based Invariant Risk Minimization (IP-IRM), combining feature disentanglement (data partition) and IRM.

Figure 10: [Pei et al., 2021]

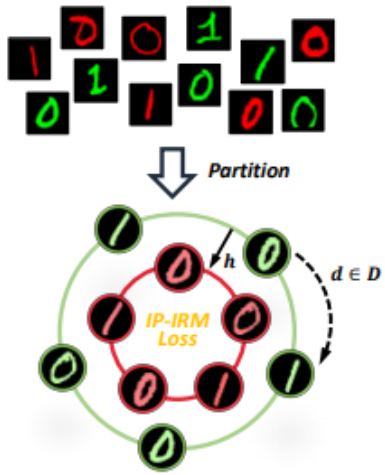


Figure 11: IP-IRM

Class imbalance learning

Traditional solution: data pre-processing (e.g. oversampling[Kotsiantis et al., 2006], re-weighting[Zhang and Pfister, 2021] etc. These methods need to know the data distribution before training. In reality, the data distribution is unknown.
We could refer to [Tang et al., 2020a] to invent a multi-head normalized classifier and make counterfactual TDE inference.

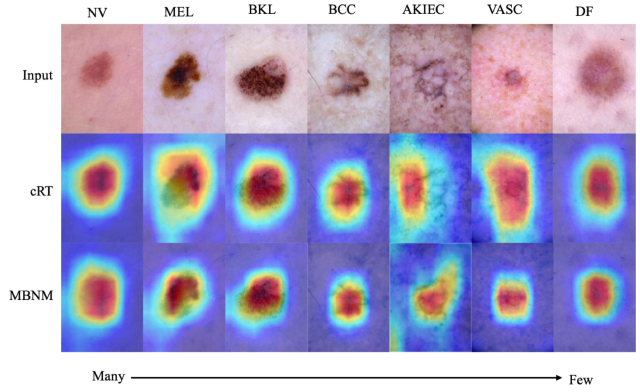


Figure 12: The visualized activation maps on skins [Zhang et al., 2021b]

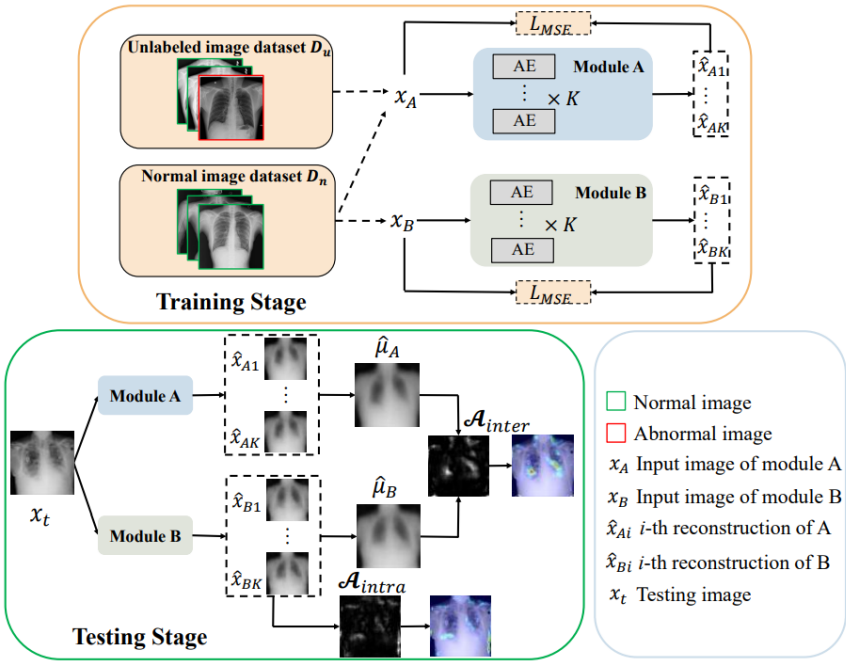


Figure 13: Dual-Distribution Discrepancy for Anomaly Detection in Chest X-Rays. (MICCAI22[Cai et al., 2022])

```
if opt.mode == "b":
    scale = torch.Tensor([1.0]).cuda().requires_grad_()
    loss_rec = rec_err.mean()*scale.mean()
    penalty_irm = torch.autograd.grad(loss_rec, scale, create_graph=True)[0]
    penalty = torch.sum(penalty_irm**2)*0.2
else:
    penalty = 0
```

	AE	AE+IRM
Rec	0.669	**0.673**
Intra	0.694	**0.715**
Inter	0.815	**0.816**
Test	0.672	**0.680**

$$\mathrm{L_{IRM}}(\Phi, \mathrm{w}) = \sum_{\mathrm{e}\in\mathcal{E}_{\mathrm{tr}}} \mathrm{R^e}(\mathrm{w}\circ\Phi)+\lambda\cdot\mathrm{D}(\mathrm{w}, \Phi, \mathrm{e}) \quad (19)$$

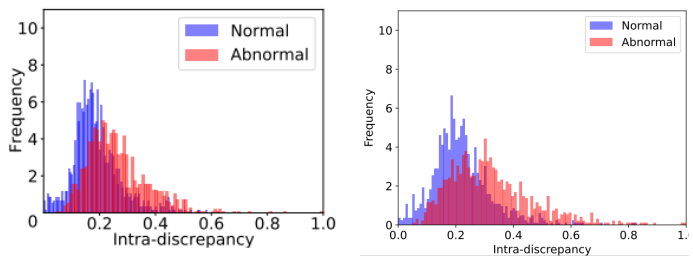


Figure 14: AE vs AE+IRM

Conclusion and Prospect

- Introduce the basic knowledge of causal inference.
- Analyze the IRM in theory and feature understanding and transfer learning in practice.
- Mention the future work on medical image analysis.
- Future research direction for CRL:
  - A well-defined causal graph is essential. However, some scenes, like anomaly detection, are very complicated. The performance of causal inference is sensitive to the causal graph.
  - The suitable approximation of operation (intervention, backdoor/frontdoor adjustment) on a node in the causal graph is hard to design.
  - Although causal inference is a promising way toward an explainable AI, a complete theory is required with mathematics definition.
  - We should explore the effect of causal representation learning in downstream applications or tasks (e.g. medical image analysis, low-level vision, etc.)

References I

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019).

Invariant risk minimization.
arXiv preprint arXiv:1907.02893.

Cai, Y., Chen, H., Yang, X., Zhou, Y., and Cheng, K.-T. (2022).

Dual-distribution discrepancy for anomaly detection in chest x-rays.
arXiv preprint arXiv:2206.03935.

Chen, C., Dou, Q., Chen, H., Qin, J., and Heng, P. A. (2020).

Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation.
IEEE transactions on medical imaging, 39(7):2494–2505.

Chen, H., Du, K., Yang, X., and Li, C. (2022a).

A review and roadmap of deep learning causal discovery in different variable paradigms.
arXiv preprint arXiv:2209.06367.

Chen, Z., Tian, Z., Zhu, J., Li, C., and Du, S. (2022b).

C-cam: Causal cam for weakly supervised semantic segmentation on medical image.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11676–11685.

References II

Gangl, M. (2010).

Causal inference in sociological research.
Annual review of sociology, 36:21–47.

Geer Jr, D. E. (2011).

Correlation is not causation.
IEEE Security & Privacy, 9(2):93–94.

Gonçalves, T., Rio-Torto, I., Teixeira, L. F., and Cardoso, J. S. (2022).

A survey on attention mechanisms for medical applications: are we moving towards better algorithms?

Harel, M. (2019).

Lmu, cmsi 498: your window into the cromulent world of cognitive systems.

Huang, J., Qin, Y., Qi, J., Sun, Q., and Zhang, H. (2022).

Deconfounded visual grounding.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 998–1006.

References III

Kamath, P., Tangella, A., Sutherland, D., and Srebro, N. (2021).

Does invariant risk minimization capture invariance?

In *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077. PMLR.

Kotsiantis, S., Kanellopoulos, D., Pintelas, P., et al. (2006).

Handling imbalanced datasets: A review.

GESTS international transactions on computer science and engineering, 30(1):25–36.

Künzel, S. R., Stadie, B. C., Vemuri, N., Ramakrishnan, V., Sekhon, J. S., and Abbeel, P. (2018).

Transfer learning for estimating causal effects using neural networks.

arXiv preprint arXiv:1808.07804.

Li, B., Shen, Y., Wang, Y., Zhu, W., Li, D., Keutzer, K., and Zhao, H. (2022).

Invariant information bottleneck for domain generalization.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7399–7407.

Li, L., Zimmer, V. A., Ding, W., Wu, F., Huang, L., Schnabel, J. A., and Zhuang, X. (2020).

Random style transfer based domain generalization networks integrating shape and spatial information.

In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 208–218. Springer.

References IV

Li, L., Zimmer, V. A., Schnabel, J. A., and Zhuang, X. (2021).

Atrialgeneral: Domain generalization for left atrial segmentation of multi-center lge mris.
In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 557–566. Springer.

Lin, Y., Dong, H., Wang, H., and Zhang, T. (2022).

Bayesian invariant risk minimization.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16021–16030.

Liu, J., Hu, Z., Cui, P., Li, B., and Shen, Z. (2021a).

Heterogeneous risk minimization.
In *International Conference on Machine Learning*, pages 6804–6814. PMLR.

Liu, J., Hu, Z., Cui, P., Li, B., and Shen, Z. (2021b).

Kernelized heterogeneous risk minimization.
arXiv preprint arXiv:2110.12425.

Liu, Q., Chen, C., Qin, J., Dou, Q., and Heng, P.-A. (2021c).

Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023.

References V

Pearl, J. (1994).

A probabilistic calculus of actions.
In *Uncertainty Proceedings 1994*, pages 454–462. Elsevier.

Pearl, J. (2009).

Causal inference in statistics: An overview.
Statistics surveys, 3:96–146.

Pei, C., Wu, F., Huang, L., and Zhuang, X. (2021).

Disentangle domain features for cross-modality cardiac image segmentation.
Medical Image Analysis, 71:102078.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. (2018).

Invariant models for causal transfer learning.
The Journal of Machine Learning Research, 19(1):1309–1342.

Rosenfeld, E., Ravikumar, P., and Risteski, A. (2020).

The risks of invariant risk minimization.
arXiv preprint arXiv:2010.05761.

References VI

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021).

Toward causal representation learning.
Proceedings of the IEEE, 109(5):612–634.

Subbaswamy, A. and Saria, S. (2020).

I-spec: An end-to-end framework for learning transportable, shift-stable models.
arXiv preprint arXiv:2002.08948.

Tang, K., Huang, J., and Zhang, H. (2020a).

Long-tailed classification by keeping the good and removing the bad momentum causal effect.
Advances in Neural Information Processing Systems, 33:1513–1524.

Tang, K., Niu, Y., Huang, J., Shi, J., and Zhang, H. (2020b).

Unbiased scene graph generation from biased training.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725.

Vlontzos, A., Rueckert, D., and Kainz, B. (2022).

A review of causality for learning algorithms in medical image analysis.
arXiv preprint arXiv:2206.05498.

References VII

Wagner, C. H. (1982).

Simpson's paradox in real life.
The American Statistician, 36(1):46–48.

Wang, T., Huang, J., Zhang, H., and Sun, Q. (2020).

Visual commonsense r-cnn.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770.

Wang, T., Yue, Z., Huang, J., Sun, Q., and Zhang, H. (2021a).

Self-supervised learning disentangled group representation as feature.
Advances in Neural Information Processing Systems, 34:18225–18240.

Wang, T., Zhou, C., Sun, Q., and Zhang, H. (2021b).

Causal attention for unbiased visual recognition.
In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100.

Wu, F. and Zhuang, X. (2020).

Cf distance: A new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation.
IEEE Transactions on Medical Imaging, 39(12):4274–4285.

References VIII

Xie, S., Gong, M., Xu, Y., and Zhang, K. (2021).

Unaligned image-to-image translation by learning to reweight.
In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14174–14184.

Zhang, J. and Bareinboim, E. (2017).

Transfer learning in multi-armed bandit: a causal approach.
In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1778–1780.

Zhang, K., Gong, M., Stojanov, P., Huang, B., Liu, Q., and Glymour, C. (2020a).

Domain adaptation as a problem of inference on graphical models.
Advances in Neural Information Processing Systems, 33:4965–4976.

Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B. J., Roth, H., Myronenko, A., Xu, D., et al. (2020b).

Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation.
IEEE transactions on medical imaging, 39(7):2531–2540.

Zhang, M., Marklund, H., Dhawan, N., Gupta, A., Levine, S., and Finn, C. (2021a).

Adaptive risk minimization: Learning to adapt to domain shift.
Advances in Neural Information Processing Systems, 34:23664–23678.

References IX

Zhang, R., Haihong, E., Yuan, L., He, J., Zhang, H., Zhang, S., Wang, Y., Song, M., and Wang, L. (2021b).

Mbnm: Multi-branch network based on memory features for long-tailed medical image recognition.
Computer Methods and Programs in Biomedicine, 212:106448.

Zhang, W., Ramezani, R., and Naeim, A. (2021c).

Causal inference in medicine and in health policy, a summary.
arXiv preprint arXiv:2105.04655.

Zhang, Z. and Pfister, T. (2021).

Learning fast sample re-weighting without reward data.
In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 725–734.

Zhou, X., Lin, Y., Zhang, W., and Zhang, T. (2022).

Sparse invariant risk minimization.
In *International Conference on Machine Learning*, pages 27222–27244. PMLR.