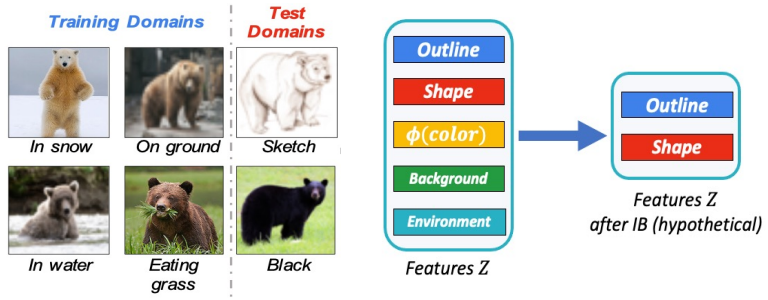
Invariant Information Bottleneck for Domain Generalization

Bo Li[1] Yifei Shen[2] Yezhen Wang[1] Wenzhen Zhu[3] Colorado Reed[4] Kurt Keutzer[4] Dongsheng Li[1] Han Zhao[5]
[1]Microsoft Research Asia, China [2]HKUST, China [3]WUSTL, USA [4]UC Berkeley, USA [5]UIUC, USA

AAAI-22

Motivation

Domain generalization(DG): learn a model from several training domains so that it generalizes to unseen test domains.



(a) Bears in training/test domains

(b) Constrain pseudo-invariant feature

Invariant risk minimization (IRM): a promising DG method but is susceptible to **pseudo-invariant feature**. In figure (a), training domains contain most *brown color bears* while test domains contain bears in *other fur color*. IRM would probably rely on *fur color* as **pseudo-invariant feature** in training domains.

Question:
How could we eliminate the usage of those features while they are pseudo-invariant in training domains?

Answer:
By constraining the information of feature Z and input X. We denote the concept with figure (b).

Method

Previous Invariant Risk Minimization objective:

$$\min_{\Phi} \sum_{e \in \mathcal{E}_{\text{train}}} R^e(\Phi) + \lambda \cdot \|\nabla_{w|w=1.0} R^e(w \circ \Phi)\|$$

Practical Optimization Objective of IRM

Invariant Information Bottleneck

We first write IRM's objective into a mutual information (MI) one as denoted in red underline part. Then we combine IB's objective in green underline part. IIB's integrated objective is listed below.

$$\max_{\Phi} \underline{I(\Phi(X), Y) - \lambda I(Y, D \mid \Phi(X))} \underline{- \beta I(X, \Phi(X))}.$$

Invariant Risk Minimization *Information* Bottleneck

The mutual information form indicates the following properties:

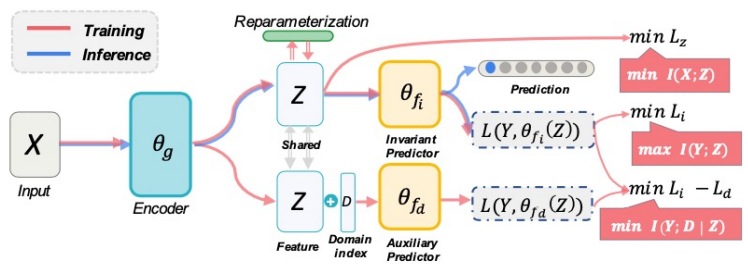
- $\max_{\Phi} I(\Phi(X), Y)$ denotes feature $\Phi(X)$ should be informative to predict the class label.
- $\min_{\Phi} I(Y, D \mid \Phi(X))$ denotes feature $\Phi(X)$ should not change across domains for the same class label *(e. g. outline and shape should always be the conditional invariant feature to predict a bear across domains).*
- $\min_{\Phi} I(X, \Phi(X))$ denotes feature $\Phi(X)$ should contain least information about input.

Variational Approximation

We denote $Z = \Phi(X)$ and formulate the MI into practical losses. More details are in **Loss Function Design** section.

- $I(Z,Y) - \beta I(Z,X) \geq \mathbb{E}_{p(x),y,z}[\log q(y \mid z)] - \beta \mathbb{E}_{p(x,z)}\left[\log \frac{p(z \mid x)}{r(z)}\right]$
- $I(Y, D \mid Z) = H(Y \mid Z) - H(Y \mid D, Z)$
  $= \sup_q \mathbb{E}_{p_{y,z}}[\log q(y \mid z)] - \sup_h \mathbb{E}_{p_{y,z,d}}[\log h(y \mid z, d)]$

Training/Inference procedure of IIB, IIB optimizes a model consisting of three parts (1) an invariant predictor. (2) a domain-dependent predictor. (3) an encoder.



Experiments

Cross Line: we create ten-valued spurious feature by adding cross lines to images. We set 10 line patterns for 10 classes. For certain class i, majority ($p_{ii} = 0.5$) images are added with line pattern i, minority images ($p_{ij} = 0.05$) are added with other line pattern j.

Type 1 (**Cross Lines**)

Methods	Validation Acc. (%) ↑	Test Acc.(%) ↑
ERM (Vapnik 1999)	90.12 ± 0.12	65.60 ± 0.27
IRM (Arjovsky et al. 2019)	63.82 ± 0.25	42.68 ± 0.32
IB-ERM (Ahuja et al. 2021)	83.93 ± 0.10	69.70 ± 0.42
IB-IRM (Ahuja et al. 2021)	81.61 ± 0.69	65.82 ± 0.77
IIB ($\lambda = 0$)	79.97 ± 0.50	69.52 ± 0.80
IIB ($\beta = 0$)	78.47 ± 0.50	66.93 ± 0.33
IIB	**92.86** ± 0.29	**71.04** ± 0.37

DomainBed: leave one domain out model selection (train-validation selection is provided in supplementary file)

Methods	Colored-MNIST	Rotated-MNIST	VLCS	PACS	OfficeHome	TerraIncognita	DomainNet	Average
ERM (Vapnik 1999)	36.7 ± 0.1	97.7 ± 0.0	77.2 ± 0.4	83.0 ± 0.7	65.7 ± 0.5	41.4 ± 1.4	40.6 ± 0.2	63.2
DANN (Ganin et al. 2017)	**40.7** ± 2.3	97.6 ± 0.2	76.9 ± 0.4	81.0 ± 1.1	64.9 ± 1.2	44.4 ± 1.1	38.2 ± 0.2	63.4
CDANN (Li et al. 2018b)	39.1 ± 4.4	97.5 ± 0.2	77.5 ± 0.2	78.8 ± 2.2	64.3 ± 1.7	39.9 ± 3.2	38.0 ± 0.1	62.2
MLDG (Li et al. 2018a)	36.7 ± 0.2	97.6 ± 0.0	77.2 ± 0.9	82.9 ± 1.7	66.1 ± 0.5	46.2 ± 0.9	41.0 ± 0.2	64.0
IRM (Arjovsky et al. 2019)	40.3 ± 4.2	97.0 ± 0.2	76.3 ± 0.6	81.5 ± 0.8	64.3 ± 1.5	41.2 ± 3.6	33.5 ± 3.0	62.0
GroupDRO (Sagawa et al. 2019)	36.8 ± 0.1	97.6 ± 0.1	**77.9** ± 0.5	83.5 ± 0.2	65.2 ± 0.2	44.9 ± 1.4	33.0 ± 0.3	62.7
MMD (Akuzawa, Iwasawa, and Matsuo 2019)	36.8 ± 0.1	97.8 ± 0.1	77.3 ± 0.5	83.2 ± 0.2	60.2 ± 5.2	46.5 ± 1.5	23.4 ± 9.5	60.7
VREx (Krueger et al. 2020a)	36.9 ± 0.3	93.6 ± 3.4	76.7 ± 1.0	81.3 ± 0.9	64.9 ± 1.3	37.3 ± 3.0	33.4 ± 3.1	60.6
ARM (Zhang et al. 2020)	36.8 ± 0.0	**98.1** ± 0.1	76.6 ± 0.5	81.7 ± 0.2	64.4 ± 0.2	42.6 ± 2.7	35.2 ± 0.1	62.2
Mixup (Yan et al. 2020)	33.4 ± 4.7	97.8 ± 0.0	77.7 ± 0.6	83.2 ± 0.4	67.0 ± 0.2	**48.7** ± 0.4	38.5 ± 0.3	63.8
RSC (Huang et al. 2020)	36.5 ± 0.2	97.6 ± 0.1	77.5 ± 0.5	82.6 ± 0.7	65.8 ± 0.7	40.0 ± 0.8	38.9 ± 0.5	62.7
MTL (Blanchard et al. 2021)	35.0 ± 1.7	97.8 ± 0.1	76.6 ± 0.5	83.7 ± 0.4	65.7 ± 0.5	44.9 ± 1.2	40.6 ± 0.1	63.5
SagNet (Nam et al. 2021)	36.5 ± 0.1	94.0 ± 3.0	77.5 ± 0.3	82.3 ± 0.1	67.6 ± 0.3	47.2 ± 0.9	40.2 ± 0.2	63.6
IIB(Ours)	39.9 ± 1.2	97.2 ± 0.2	77.2 ± 1.6	**83.9** ± 0.2	**68.6** ± 0.1	45.8 ± 1.4	**41.5** ± 2.3	**64.9**

Conclusion: Cross Line experiment performance suggest IIB can better overcome pseudo-invariant features than IRM and other methods. DomainBed experiment performance suggest IIB is applicable to realistic DG task.

Conclusion

In this paper, we have following conclusions:

- To mitigate pseudo-invariant features, inspired by the IB's principle, we propose to constrain the mutual information between inputs and features.
- We developed a novel information-theoretic approach IIB to overcome above issues.
- We further adopt variational approximation to develop tractable loss functions.
- We analyze IIB's performance with extensive experiments on both synthetic and large-scale benchmarks.