Exploiting Domain-Specific Features to Enhance Domain Generalization

Manh-Ha Bui[1], Toan Tran[1], Anh Tuan Tran[1], Dinh Phung[1,2]

[1]VinAI Research, [2]Monash University

Advances in Neural Information Processing Systems, Dec 2021

Table of Contents

1. Problem setting and Definitions
2. A theoretical analysis under the Information bottleneck method
3. Algorithm: meta-Domain Specific-Domain Invariant (mDSDI)
4. Experiments
5. Summary

Table of Contents

1. Problem setting and Definitions
2. A theoretical analysis under the Information bottleneck method
3. Algorithm: meta-Domain Specific-Domain Invariant (mDSDI)
4. Experiments
5. Summary

Domain Generalization (DG)

Set up

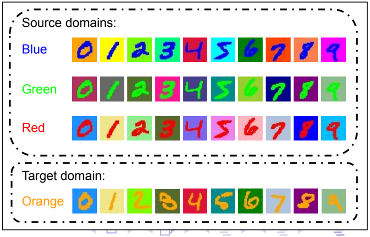
Given N source domains $S^{(i)} = \{(x_j^{(i)}, y_j^{(i)})\}_{j=1}^{n_i}$, $i = 1, \ldots, N$, where n_i is the number of data points in $S^{(i)}$, i.e., $(x_j^{(i)}, y_j^{(i)}) \overset{iid}{\sim} P^{(i)}(x, y)$ where $P^{(i)}(x, y) \sim \mathcal{P}$; and $x_j^{(i)} \sim P_{\mathcal{X}}^{(i)}$, in which $P_{\mathcal{X}}^{(i)} \sim P_{\mathcal{X}}$.

Goal

Training a model only on the set of source domains $\{S^{(i)}\}_{i=1}^{N}$ without any access to the data points but still perform well on the test dataset $S^T = \{(x_j^T, y_j^T)\}_{j=1}^{n_T}$, where $(x_j^T, y_j^T) \overset{iid}{\sim} P^T(x, y)$ and $P^T(x, y) \sim \mathcal{P}$.

Examples

Background-Colored-MNIST: Sketch color different across domains.



Domain-invariant

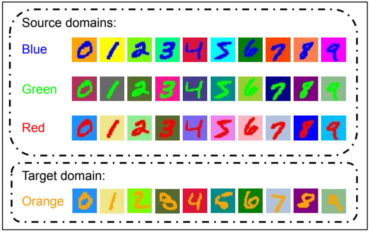
Definition 1: Domain-invariant

A feature extraction mapping $Q : \mathcal{X} \to \mathcal{Z}$ is said to be **domain-invariant** if the distribution $P_Q(Q(X))$ is unchanged across the source domains, i.e., $\forall i, j = 1, \ldots, N,\ i \neq j$ we have $P_Q^{(i)}(Q(X)) \equiv P_Q^{(j)}(Q(X))$, where $P_Q^{(i)}(Q(X)) = P_Q(Q(X)|X \sim P_{\mathcal{X}}^{(i)}),\ i = 1, \ldots, N$. In this case, the corresponding latent representation $Z_I = Q(X)$ is then called the domain-invariant representation.

Examples

The sketch of digits.

- Current methods often try to learn domain-invariant
- Ignore all color information, only concentrate on the sketch
- But, what if the sketch is blurred?



Domain-specific

Definition 2: Domain-specific

A feature extraction mapping $R : \mathcal{X} \to \mathcal{Z}$ is said to be **domain-specific** if $\forall i, j = 1, \ldots, N,\ i \neq j$ we have $P_R^{(i)}(R(X)) \neq P_R^{(j)}(R(X))$, where $P_R^{(i)}(R(X)) = P_R(R(X)|X \sim P_{\mathcal{X}}^{(i)}),\ i = 1, \ldots, N$. In this case, given $X \sim P_{\mathcal{X}}^{(i)}$ the corresponding latent representation $Z_S^{(i)} = R(X)$ is then called the domain-specific representation w.r.t. the domain $S^{(i)}$.

Examples

The background color.

- The background-color different across class and domain
- In the target domain, one of these is similar to the source
- **If the sketch is blurry, we can rely on the background to classify digits!**

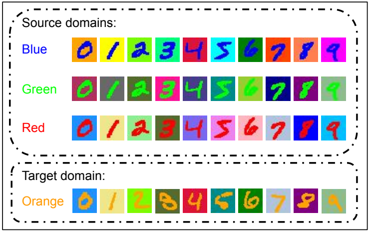
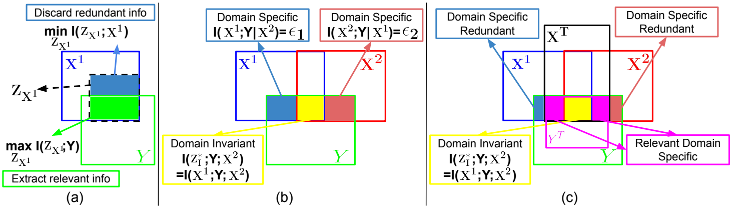


Table of Contents

1. Problem setting and Definitions
2. A theoretical analysis under the Information bottleneck method
3. Algorithm: meta-Domain Specific-Domain Invariant (mDSDI)
4. Experiments
5. Summary

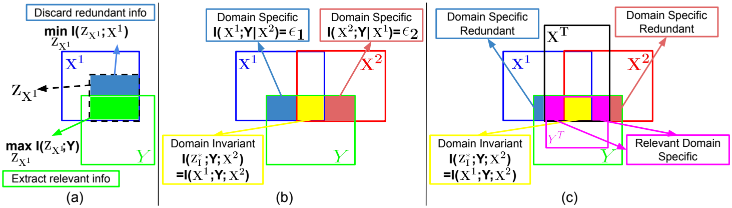
Label-correlated domain-specificity



Assumption 1: Label-correlated domain-specificity

Assuming that there exists a domain-specific representation $Z_S^{(1)}$ extracted by the deterministic mapping $Z_S^{(1)} = R(X^1)$ in definition 2, which correlates with label in domain S^1 such that $I(Z_S^{(1)}; Y|X^2) = I(X^1; Y|X^2) = \varepsilon_1$, where $\varepsilon_1 > 0$ is a constant.

Minimal and sufficient representations with label

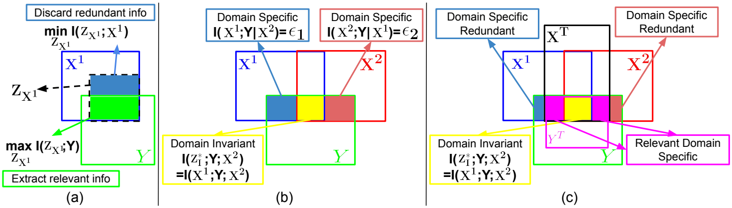


Definition 3: Minimal and sufficient representations with label

Let $Z_{X^1} = G(X^1)$ is the output of a deterministic latent mapping G. A representation Z_{sup} is said to be the sufficient label-related representation and Z_{sup^*} is said to be the minimal and sufficient representation if:

$$Z_{sup} = \underset{G}{\mathrm{argmax}} I(Z_{X^1}; Y) \text{ and } Z_{\text{sup}^*} = \underset{Z_{\text{sup}}}{\mathrm{argmin}} I(Z_{\text{sup}}; X^1) \text{ s.t. } I(Z_{\text{sup}}; Y) \text{ is maximized.}$$

Minimal and sufficient representations with domain-invariance

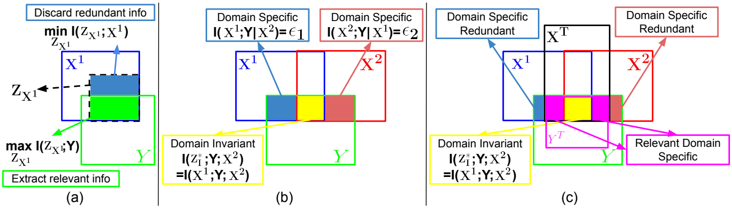


Definition 4: Minimal and sufficient representations with domain-invariance

Let $Z_{X^1} = Q(X^1)$ is the output of a deterministic domain invariant mapping Q in the definition 1. Then Z_I is said to be the sufficient domain-invariant representation and Z_{I^*} is said to be the minimal and sufficient representation if:

$$Z_I = \underset{Q}{\operatorname{argmax}} I(Z_{X^1}; X^2) \text{ and } Z_{I^*} = \underset{Z_I}{\operatorname{argmin}} I(Z_I; X^1) \text{ s.t. } I(Z_I; X^2) \text{ is maximized.}$$

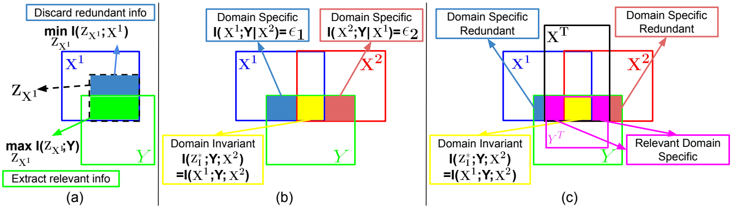
Label-related information with domain-specificity



Lemma 1: Determinism

If $P(Z_{X^1}|X^1)$ is a Dirac delta function, then the following conditional independence holds: $Y \perp\!\!\!\perp Z_{X^1}|X^1$ and $X^2 \perp\!\!\!\perp Z_{X^1}|X^1$, inducing a Markov chain $X^2 \leftrightarrow Y \leftrightarrow X^1 \rightarrow Z_{X^1}$.

Label-related information with domain-specificity



Lemma 1: Determinism

If $P(Z_{X^1}|X^1)$ is a Dirac delta function, then the following conditional independence holds: $Y \perp\!\!\!\perp Z_{X^1}|X^1$ and $X^2 \perp\!\!\!\perp Z_{X^1}|X^1$, inducing a Markov chain $X^2 \leftrightarrow Y \leftrightarrow X^1 \rightarrow Z_{X^1}$.

Theorem 1: Label-related information with domain-specificity

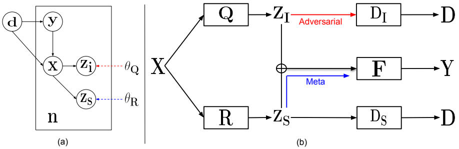
Assuming that there exists a domain-specific value $\varepsilon_1 > 0$ in domain S^1 (see Assumption 1), the label-related representation - based learning approach (i.e., using Z_{sup} and Z_{sup^*}) provides better prediction performance than the domain-invariant representation - based method (i.e., using Z_I and Z_{I^*}). Formally,

$$I(X^1; Y) = I(Z_{sup}; Y) = I(Z_{sup^*}; Y) = I(Z_{I^*}; Y) + \varepsilon_1 > I(Z_{I^*}; Y).$$

Table of Contents

1. Problem setting and Definitions
2. A theoretical analysis under the Information bottleneck method
3. Algorithm: meta-Domain Specific-Domain Invariant (mDSDI)
4. Experiments
5. Summary

Domain-Invariant and Domain-Specific Extraction



Objectives funcitons:

- Domain-Invariant Extraction

$$\min_{\theta_Q} \max_{\theta_{D_I}} \left\{ L_{Z_I} := -\mathbb{E}_{x,d \sim X,D} \left[d \log D_I(Q(x)) \right] \right\}. \tag{1}$$

- Domain-Specific Extraction

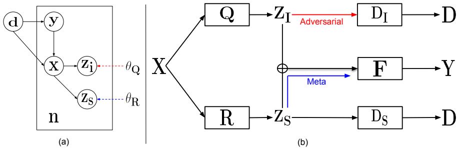
$$\min_{\theta_{D_S}, \theta_R} \left\{ L_{Z_S} := -\mathbb{E}_{x,d \sim X,D} \left[d \log D_S(R(x)) \right] \right\}. \tag{2}$$

- Domain-Invariant and Domain-Specific Disentanglement

$$\min_{\theta_Q, \theta_R} \left\{ L_D := \mathbb{E}_{x \sim X} \left[\| \mathrm{Cov}(Q(x), R(x)) \|_2 \right] \right\}, \tag{3}$$

where $\| \cdot \|_2$ is the L_2 norm.

Domain-Invariant and Domain-Specific Extraction



- Sufficiency of domain-specific and domain-invariant w.r.t. the classification task:

$$\hat{Y} = F_{\theta_F}(Z_I \oplus Z_S), \tag{4}$$

where $\oplus$ denotes the concatenation operation. Then, the training process of F is then performed by solving

$$\min_{\theta_Q, \theta_R, \theta_F} \left\{ L_T := -\mathbb{E}_{x,y \sim X,Y} \left[y \log F(Q(x), R(x)) \right] \right\}. \tag{5}$$

- Meta-training Domain-specific

$$\min_{w} \left\{ L_{T_m} := f\left(w - \nabla f\left(w, S_{mr} \right), S_{me} \right) \right\}, \tag{6}$$

where $w = (\theta_R, \theta_F)$ and

$$f\left(w, S_m \right) = -\mathbb{E}_{x,y \sim X,Y} \left[y \log F(Z_I, R(x)) \right]. \tag{7}$$

Training and Inference

Each iteration of the training process consists of two steps:

i) Integrating the objective functions (1), (2), (3) and (5) to construct L_A:

$$\min_{\theta_Q,\theta_{D_S},\theta_R,\theta_F} \max_{\theta_{D_I}} \left\{ L_A := \lambda_{Z_I} L_{Z_I} + \lambda_{Z_S} L_{Z_S} + \lambda_D L_D + L_T \right\}. \tag{8}$$

where λ_{Z_I}, λ_{Z_S} and λ_D are selected as the balanced parameters.

ii) In each mini-batch, the meta-train and meta-test are split, then the gradient transformation step from meta-train domains to the meta-test domain is performed by solving the optimization problem (6).

Algorithm 1: Training and Inference processes of mDSDI

Training Input: Source domain $S^{(i)}$, encoder Q_{θ_Q}, R_{θ_R}, domain classifier $D_{\theta_{D_I}}$, $D_{\theta_{D_S}}$ for Z_I, Z_S, task classifier F_{θ_F}, batch size B, learning rate η. **Output:** The optimal: $Q^*_{\theta_Q}$, $R^*_{\theta_R}$, $F^*_{\theta_F}$;

for *ite* $= 1 \rightarrow$ *iterations* **do**

  Sample S_B with a mini-batch B for each domain $S^{(i)}$;

  Compute L_A using Eq. (8) and perform gradient update $\nabla_{\theta_Q,\theta_R,\theta_{D_I},\theta_{D_S},\theta_F} L_A$ with η.;

  for $j = 1 \rightarrow N$ *(number of source domains)* **do**

    Split Meta-train $S_{B/j}$, Meta-test S_j;

    Meta-train: Perform gradient update $\nabla_{\theta_R,\theta_F}$ by minimizing Eq. (7) with $S_{B/j}$ and η;

    Meta-test: Compute L_{T_m} using Eq. (6) with S_j and updated gradient from **Meta-train**;

    Meta-optimization: Perform gradient update $\nabla_{\theta_R,\theta_F} L_{T_m}$ with η;

  end

end

Inference Input: Target domain S^T, optimal: $Q^*_{\theta_Q}$, $R^*_{\theta_R}$, $F^*_{\theta_F}$. **Output:** Y^T using Eq. (4);

Table of Contents

1 Problem setting and Definitions

2 A theoretical analysis under the Information bottleneck method

3 Algorithm: meta-Domain Specific-Domain Invariant (mDSDI)

4 Experiments

5 Summary

Results on benchmark dataset

Dataset	**Domains**					
Colored MNIST	10% flip  Degree of correlation between color and label	20% flip 	90% flip 			
Rotated MNIST	0' 	15' 	30' 	45' 	60' 	75' 
Background-Colored MNIST	Blue 	Green 	Red 	Orange 		
VLCS	Caltech101 	LabelMe 	SUN09 	VOC2007 		
PACS	Art 	Cartoon 	Photo 	Sketch 		
Office Home	Art 	Clipart 	Product 	Photo 		
Terra Incognita	L100 	L38 	L43 	L46 		
DomainNet	Clipart 	Infographic 	Painting 	QuickDraw 	Photo 	Sketch 

Results on benchmark dataset

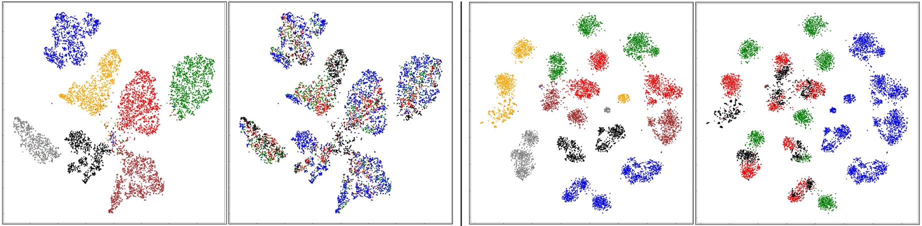
Table 1: Classification accuracy (%) for all algorithms and datasets summarization. Our mDSDI method achieves highest accuracy on average when comparing 14 popular DG algorithms across 7 benchmark datasets.

Method	CMNIST	RMNIST	VLCS	PACS	OfficeHome	TerraInc	DomainNet	Average
ERM [25]	51.5±0.1	98.0±0.0	77.5±0.4	85.5±0.2	66.5±0.3	46.1±1.8	40.9±0.1	66.6
IRM [18]	52.0±0.1	97.7±0.1	78.5±0.5	83.5±0.8	64.3±2.2	47.6±0.8	33.9±2.8	65.4
GroupDRO [26]	52.1±0.0	98.0±0.0	76.7±0.6	84.4±0.8	66.0±0.7	43.2±1.1	33.3±0.2	64.8
Mixup [34, 35, 36]	52.1±0.2	98.0±0.1	77.4±0.6	84.6±0.6	68.1±0.3	47.9±0.8	39.2±0.1	66.7
MLDG [11]	51.5±0.1	97.9±0.0	77.2±0.4	84.9±1.0	66.8±0.6	47.7±0.9	41.2±0.1	66.7
CORAL [29]	51.5±0.1	98.0±0.1	78.8±0.6	86.2±0.3	68.7±0.3	47.6±1.0	41.5±0.1	67.5
MMD [30]	51.5±0.2	97.9±0.0	77.5±0.9	84.6±0.5	66.3±0.1	42.2±1.6	23.4±9.5	63.3
DANN [31]	51.5±0.3	97.8±0.1	78.6±0.4	83.6±0.4	65.9±0.6	46.7±0.5	38.3±0.1	66.1
CDANN [32]	51.7±0.1	97.9±0.1	77.5±0.1	82.6±0.9	65.8±1.3	45.8±1.6	38.3±0.3	65.6
MTL [1, 27]	51.4±0.1	97.9±0.0	77.2±0.4	84.6±0.5	66.4±0.5	45.6±1.2	40.6±0.1	66.2
SagNets [37]	51.7±0.0	98.0±0.0	77.8±0.5	**86.3±0.2**	68.1±0.1	**48.6±1.0**	40.3±0.1	67.2
ARM [28]	**56.2±0.2**	**98.2±0.1**	77.6±0.3	85.1±0.4	64.8±0.3	45.5±0.3	35.5±0.2	66.1
VREx [33]	51.8±0.1	97.9±0.1	78.3±0.2	84.9±0.6	66.4±0.6	46.4±0.6	33.6±2.9	65.6
RSC [38]	51.7±0.2	97.6±0.1	77.1±0.5	85.2±0.9	65.5±0.9	46.6±1.0	38.9±0.5	66.1
mDSDI (Ours)	52.2±0.2	98.0±0.1	**79.0±0.3**	86.2±0.2	**69.2±0.4**	48.1±1.4	**42.8±0.1**	**67.9**

Observations:

- Our mDSDI still preserves domain-invariant information
- Our mDSDI could capture the usefulness of domain-specific information
- Extending beyond the invariance view to usefulness domain-specific information is important

How does mDSDI work?



(a) Domain-invariant (Classes) (b) Domain-invariant (Domains) (c) Domain-specific (Classes) (d) Domain-specific (Domains)

Figure: Feature visualization for domain-invariant: (a): different colors represent different classes; (b): different colors indicate different domains. Feature visualization for domain-specific: (c): different colors represent different classes; (d): different colors indicate different domains. Source domain includes: art (red), cartoon (green), sketch (blue) while target domain is photo (black) in the domain plots.

Observations:

- Our domain-invariant extractor can minimize the distance between the distribution of the domains (b)
- These domain-invariant features still make mistakes on the classification task (a)
- The domain-specific representation better distinguishes points by class label (c)
- The photo domain's specific features (black) are close to the art domain (red) (d)

Ablation study: Important of mDSDI on the Background-Colored-MNIST dataset

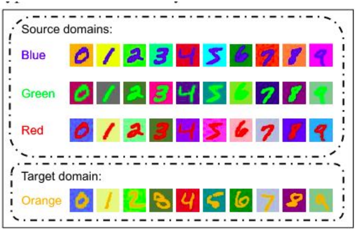


Figure 4: Background-Colored-MNIST Dataset, where source domains include {red, green, blue} digit colors and target domain has {orange} color.

Table 2: Classification accuracy (%) on Background-Colored-MNIST. Ablation study shows impact of domain-invariant when combined with meta-training on domain-specific in our method.

Method	**Accuracy**
DI	65.7±4.6
DI-Meta	63.6±5.1
DS	70.7±4.8
DS-Meta	75.3±3.4
DSDI-Without L_D	81.4±2.6
DSDI-Without Meta	80.4±1.7
DSDI-Meta	82.1±1.4
DSDI-Meta DI	79.0±2.3
mDSDI-Meta DS (Ours)	**89.7**±**0.8**

Observations:

- Combining domain-invariant and domain-specific is crucial
- Adding disentanglement loss L_D is essential to boost our model performance
- Meta-training is necessary but only for the domain-specific
- Our combination with domain-invariant and domain-specific is not easy: a deep ensemble between two neural networks

Table of Contents

1 Problem setting and Definitions

2 A theoretical analysis under the Information bottleneck method

3 Algorithm: meta-Domain Specific-Domain Invariant (mDSDI)

4 Experiments

5 Summary

Summary

Contributions:

- Provide theoretical analysis to point out the limitation of only learning domain-invariant and prove the essential of domain-specific
- Propose the meta-Domain Specific-Domain Invariant (mDSDI) - a novel theoretically sound framework that extends beyond the invariance view to further capture the usefulness of domain-specific information
- Empirically show that mDSDI provides competitive results with state-of-the-art techniques in DG and confirms the hypothesis that domain-specific is essential in ablation studies

For more information:

- PDF, code available at *https://github.com/VinAIResearch/mDSDI*
- **Come see our poster!**

See you at the NeurIPS 2021 conference!