Optimal Representations for Covariate Shift

Yangjun Ruan

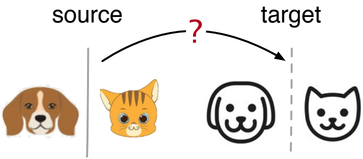
Joint with Yann Dubois, Chris J. Maddison

ICLR 2022

Overview

Overview

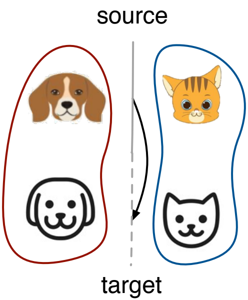
ML experiences distribution shifts from train (source) to test (target)



Overview

ML experiences distribution shifts from train (source) to test (target)

Goal: learn robust representations Z of data X from which source (d_s) predictors perform well on target (d_t)

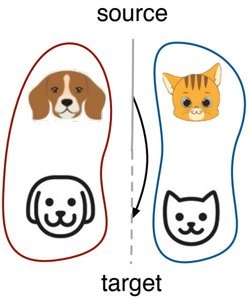


Overview

ML experiences distribution shifts from train (source) to test (target)

Goal: learn robust representations Z of data X from which source (d_s) predictors perform well on target (d_t)

Optimal Z^*: all source optimal predictors minimize target risk



Overview

We characterize the optimally robust Z^* to covariate shift

☺ prove sufficient and necessary condition for optimal Z^*

Overview

We characterize the optimally robust Z^* to covariate shift

- ☺ prove sufficient and necessary condition for optimal Z^*
- ☺ derive practical self-supervised objectives for learning Z^*

Overview

We characterize the optimally robust Z^* to covariate shift

- ☺ prove sufficient and necessary condition for optimal Z^*
- ☺ derive practical self-supervised objectives for learning Z^*
- ☺ show why CLIP [4] is more robust over other SSL methods

Overview

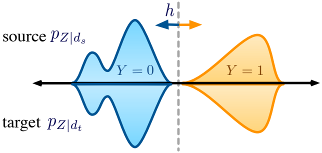
We characterize the optimally robust Z^* to covariate shift

- ☺ prove sufficient and necessary condition for optimal Z^*
- ☺ derive practical self-supervised objectives for learning Z^*
- ☺ show why CLIP [4] is more robust over other SSL methods
- ☺ improve CLIP's robustness with our objectives

Theory: Characterizing Z^*

Intuition

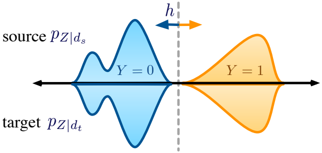
Desiderata: reduce to typical ML setup in Z space



match distribution

Intuition

Desiderata: reduce to typical ML setup in Z space

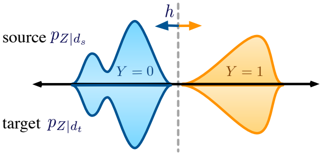


match distribution

✔ Sufficient condition (...most previous work hinted towards)

Intuition

Desiderata: reduce to typical ML setup in Z space



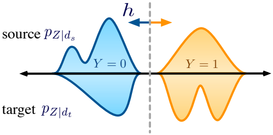
match distribution

- ✔ Sufficient condition (...most previous work hinted towards)
- ✘ Necessary? Achievable?

Intuition

Minimal sufficiency: Z^* should

- remain discriminative about Y
- have invariant support

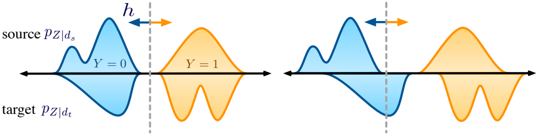


discriminative & support

Intuition

Minimal sufficiency: Z^* should

- remain discriminative about Y
- have invariant support

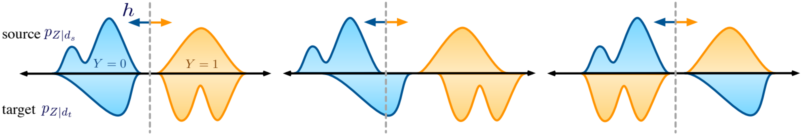


discriminative & support   only discriminative

Intuition

Minimal sufficiency: Z^* should

- remain discriminative about Y
- have invariant support



discriminative & support | only discriminative | only support match

Problem Setup

Formalization with domain generalization (DG) language:

1. Given
    - A set of domains $\mathcal{D}$ [Asm: discrete finite]
    - Domain-specific $\left\{p_{X,\,Y\,|\,d}\right\}_{d\in\mathcal{D}}$ [Asm: gen. covariate shift]
    - Loss $\ell : \mathcal{Y} \times \Gamma \to \mathbb{R}_{\geq 0}$

Problem Setup

Formalization with domain generalization (DG) language:

1. Given
   - A set of domains $\mathcal{D}$ [Asm: discrete finite]
   - Domain-specific $\{p_{X,Y\,|\,d}\}_{d\in\mathcal{D}}$ [Asm: gen. covariate shift]
   - Loss $\ell : \mathcal{Y} \times \Gamma \to \mathbb{R}_{\geq 0}$
2. Learn an encoder $p_{Z\,|\,X}$

Problem Setup

Formalization with domain generalization (DG) language:

1. Given
   - A set of domains $\mathcal{D}$ [Asm: discrete finite]
   - Domain-specific $\left\{p_{X,Y\,|\,d}\right\}_{d\in\mathcal{D}}$ [Asm: gen. covariate shift]
   - Loss $\ell : \mathcal{Y} \times \Gamma \to \mathbb{R}_{\geq 0}$
2. Learn an encoder $p_{Z\,|\,X}$
3. Measure DG risk:
   - Select a random source D_s and target D_t
   - Train a source predictor: $h \in \mathcal{H}^*_{D_s} := \arg\min_h \mathrm{R}_h^{D_s}\left[Y\,|\,Z\right]$
   - Measure target risk $\mathrm{R}_h^{D_t}\left[Y\,|\,Z\right]$

     where $\mathrm{R}_h^{d}\left[Y\,|\,Z\right] := \mathbb{E}_{p_{Z,Y\,|\,d}}[\ell(Y, h(Z))]$

Problem Setup

Goal: minimize the *idealized domain generalization* (IDG) risk w.r.t. Z

$$\mathrm{R}_{\mathrm{IDG}}\left[Y \mid Z\right] := \underbrace{\mathbb{E}_{p_{D_s,D_t}}}_{\text{random domains}} \underbrace{\sup_{h \in \mathcal{H}^*_{D_s}}}_{\text{worst source risk. min.}} \underbrace{\mathrm{R}_h^{D_t}\left[Y \mid Z\right]}_{\text{target risk}}$$

Uniform guarantees:

- random domains
- worst-case source predictor

Problem Setup

Goal: minimize the *idealized domain generalization* (IDG) risk w.r.t. Z

$$\mathrm{R}_{\mathrm{IDG}}\left[Y \mid Z\right] := \underbrace{\mathbb{E}_{p_{D_s,D_t}}}_{\text{random domains}} \underbrace{\sup_{h \in \mathcal{H}^*_{D_s}}}_{\text{worst source risk. min.}} \underbrace{\mathrm{R}_h^{D_t}\left[Y \mid Z\right]}_{\text{target risk}}$$

Uniform guarantees:

- random domains
- worst-case source predictor

Idealized setup for simplicity:

- population risk used for source predictor selection
- universal hypothesis class

Characterization of Z^*

Theorem (Optimality conditions, informal)

Under generalized covariate shift and some mild assumptions, Z^* is optimal for IDG if and only if it

- remains discriminative: $\mathrm{R}[Y \mid Z^*] = \mathrm{R}[Y \mid X]$
- has invariant support: $\mathrm{supp}(p_{Z^* \mid d_s}) = \mathrm{supp}(p_{Z^* \mid d_t}), \ \forall d_s, d_t \in \mathcal{D}$

Characterization of Z^*

Theorem (Optimality conditions, informal)

Under generalized covariate shift and some mild assumptions, Z^* is optimal for IDG if and only if it

- remains discriminative: $\mathrm{R}[Y | Z^*] = \mathrm{R}[Y | X]$
- has invariant support: $\mathrm{supp}(p_{Z^* \mid d_s}) = \mathrm{supp}(p_{Z^* \mid d_t}),\ \forall d_s, d_t \in \mathcal{D}$

☺ achievable sufficient and necessary condition

Characterization of Z^*

Theorem (Optimality conditions, informal)

Under generalized covariate shift and some mild assumptions, Z^* is optimal for IDG if and only if it

- remains discriminative: $\mathrm{R}\left[Y \mid Z^*\right] = \mathrm{R}\left[Y \mid X\right]$
- has invariant support: $\mathrm{supp}(p_{Z^* \mid d_s}) = \mathrm{supp}(p_{Z^* \mid d_t}),\ \forall d_s, d_t \in \mathcal{D}$

☺ achievable sufficient and necessary condition

☹ requires access to labeled target domain

Characterization of Z^*

Proposition (No free lunch for IDG, informal)

Let Z_{d_s} be any rep. chosen on some source d_s and C a constant rep.

Under mild assumptions, if Z_{d_s} outperforms C on some "good" targets outside the source's support, there are many "bad" targets on which Z_{d_s} is strictly worse than C.

Characterization of Z^*

Proposition (No free lunch for IDG, informal)

Let Z_{d_s} be any rep. chosen on some source d_s and C a constant rep.

Under mild assumptions, if Z_{d_s} outperforms C on some "good" targets outside the source's support, there are many "bad" targets on which Z_{d_s} is strictly worse than C.

- ✔ implies the failure of current DG methods
  - ☹ unable to outperform ERM on a unified benchmark [3]
  - ☹ insufficient access to or strong asmp. on targets

Characterization of Z^*

Proposition (No free lunch for IDG, informal)

Let Z_{d_s} be any rep. chosen on some source d_s and C a constant rep.

Under mild assumptions, if Z_{d_s} outperforms C on some "good" targets outside the source's support, there are many "bad" targets on which Z_{d_s} is strictly worse than C.

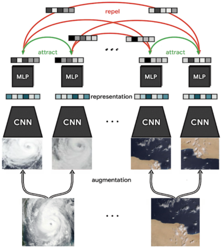
- ✔ implies the failure of current DG methods
  - ☹ unable to outperform ERM on a unified benchmark [3]
  - ☹ insufficient access to or strong asmp. on targets
- ✘ how to deal with necessary (but unrealistic) access to targets?

Method: Learning Z^* with SSL

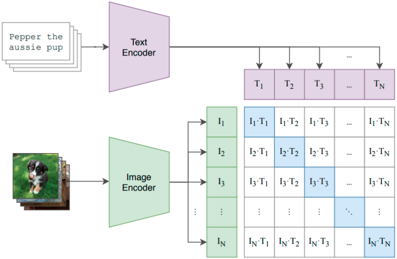
Deviation: Self-Supervised Learning (SSL)

Recent SSL methods learn transferable and robust reps.:

- train on large-scale unlabelled data ($\gg=$ ImageNet)
- use augmentations as surrogate information for Y



SimCLR [1]: image aug.

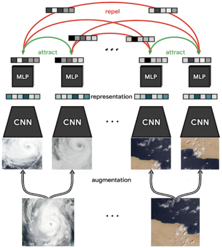


CLIP [4]: text caption as aug.

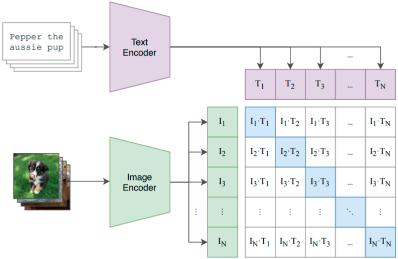
Deviation: Self-Supervised Learning (SSL)

Recent SSL methods learn transferable and robust reps.:

- train on large-scale unlabelled data ($\gg=$ ImageNet)
- use augmentations as surrogate information for Y



SimCLR [1]: image aug.



CLIP [4]: text caption as aug.

Robustness of different SSL methods varies:

☺ CLIP achieves incredible robustness to distribution shifts

Learning Z^* with SSL

Augmentation A for learning Z^*:

- Label-perserving: retain information about Y

Learning Z^* with SSL

Augmentation A for learning Z^*:

- Label-perserving: retain information about Y
- Domain-agnostic: *no correlation with domain*

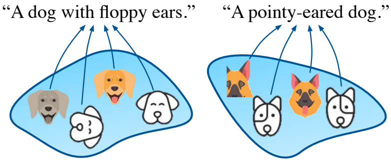
Learning Z^* with SSL

Augmentation A for learning Z^*:

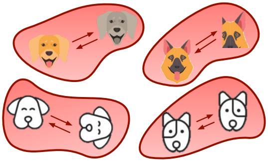
- Label-perserving: retain information about Y
- Domain-agnostic: *no correlation with domain*

Domain-agnostic A

- ✔ Example: image-text aug. (e.g., CLIP [4])
- ✘ Counterexample: standard image aug. (e.g., SimCLR [1])



CLIP aug. ⇒ domain-agnostic rep.



SimCLR aug. ⇒ domain-correlated rep.

Learning Z^* with SSL

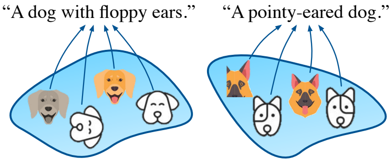
Augmentation A for learning Z^*:

- Label-perserving: retain information about Y
- Domain-agnostic: *no correlation with domain*

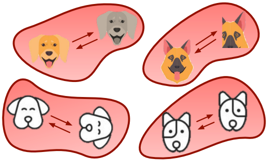
Domain-agnostic A

✔ Example: image-text aug. (e.g., CLIP [4])

✘ Counterexample: standard image aug. (e.g., SimCLR [1])



CLIP aug. ⇒ domain-agnostic rep.



SimCLR aug. ⇒ domain-correlated rep.

☺ implies the incredible robustness of CLIP over other SSL models

Learning Z^* with SSL

Proposition (Learning Z^* with domain-agnostic A)

Let $p_{A \mid X}$ be a domain-agnostic augmenter. Then any optimal solution $p_{Z^* \mid X}$ of the following objective is optimal for IDG:

$$\max_{p_{Z \mid X}} \mathrm{I}[A; Z]$$
$$\text{s.t.} \quad \operatorname{supp}(p_{Z \mid d}) = \operatorname{supp}(p_Z), \ \forall d \in \mathcal{D}$$

Learning Z^* with SSL

Proposition (Learning Z^* with domain-agnostic A)

Let $p_{A \mid X}$ be a domain-agnostic augmenter. Then any optimal solution $p_{Z^* \mid X}$ of the following objective is optimal for IDG:

$$\max_{p_{Z \mid X}} \mathrm{I}[A; Z]$$
$$\text{s.t.} \quad \mathrm{supp}(p_{Z \mid d}) = \mathrm{supp}(p_Z),\ \forall d \in \mathcal{D}$$

- ☺ No Y anymore!
- ☹ support invariance constraint

Learning Z^* with SSL

Practical objectives:

$$\arg\min_{p_{Z \mid X}} \underbrace{-\mathrm{I}[A; Z]}_{\text{max. MI}} + \lambda \underbrace{\mathrm{B}[Z, D]}_{\text{dom. bottleneck}}$$

Learning Z^* with SSL

Practical objectives:

$$\arg\min_{p_{Z \mid X}} \underbrace{-\mathrm{I}[A; Z]}_{\text{max. MI}} + \lambda \underbrace{\mathrm{B}[Z, D]}_{\text{dom. bottleneck}}$$

- Maximize $\mathrm{I}[A; Z]$: MI lower bound (e.g., InfoNCE)
- Domain bottleneck $\mathrm{B}[Z, D]$: enforce support invariance

Learning Z^* with SSL

Practical objectives:

$$\arg\min_{p_{Z\mid X}} \underbrace{-\mathrm{I}[A;Z]}_{\text{max. MI}} + \lambda \underbrace{\mathrm{B}[Z,D]}_{\text{dom. bottleneck}}$$

- Maximize $\mathrm{I}[A;Z]$: MI lower bound (e.g., InfoNCE)
- Domain bottleneck $\mathrm{B}[Z,D]$: enforce support invariance

Domain bottleneck: previous DG methods (e.g., DANN [2]) can apply

☐ Contrastive adversarial domain (CAD) bottleneck $\mathrm{I}[Z;D]$

☺ Requires no explicit trainable domain classifier

☺ Constructs an implicit domain classifier from contrastive var. dist.

☐ Entropy (Ent) bottleneck $\mathrm{H}[Z]$

☺ Requires no access to domain information

Learning Z^* with SSL

Summary: one can learn optimal Z^* with SSL using:

- large-scale unlabeled data
- contrastive learning with domain-agnostic augmentations
- domain bottlenecks

Experiments

Exploiting Pretrained CLIP for Z^*

Motivation: CLIP was trained

- ✔ with 400M image-text augmentations
- ✘ without explicit domain bottlenecks

Idea:

- Finetune CLIP with bottlenecks on available data
- Evaluate with linear probe on DomainBed [3]

Exploiting Pretrained CLIP for Z^*

Algorithm	VLCS	PACS	OfficeHome	DomainNet
ERM	77.6 ± 0.3	86.7 ± 0.3	66.4 ± 0.5	41.3 ± 0.1
DomainBed SOTA	79.9 ± 0.2	87.2 ± 0.1	68.4 ± 0.2	41.8 ± 0.1
DINO + CAD	69.6 ± 0.6	76.1 ± 0.1	56.9 ± 0.5	33.6 ± 0.1
CLIP	80.7 ± 0.4	93.7 ± 0.8	79.6 ± 0.1	52.8 ± 0.1
CLIP + CAD	81.6 ± 0.1	94.9 ± 0.3	80.0 ± 0.2	53.7 ± 0.1

☺ SOTA result with domain-agnostic aug. and bottlenecks!

Towards Generic Robust Representations with SSL

Idea: learn task- and domain-agnostic robust reps.

- Task: use LAION-400M [5] with text-image contrastive loss
- Domain: finetune CLIP with Ent bottleneck

Towards Generic Robust Representations with SSL

Idea: learn task- and domain-agnostic robust reps.

- Task: use LAION-400M [5] with text-image contrastive loss
- Domain: finetune CLIP with Ent bottleneck

Evaluate: natural distribution shift [6]

	IN	IN-V2	IN-S	YT-BB	IN-Vid	ObjNet	IN-A	IN-R	Avg.
Pretrained	75.2	64.2	41.0	58.4	71.6	42.8	27.5	62.9	52.6
Tuned w/o Ent	73.8	62.1	37.0	56.9	68.8	41.3	26.0	58.1	50.0
Tuned w/ Ent	74.2	62.7	38.9	58.1	70.1	42.1	26.2	60.8	51.3

☺ Consistently improved robustness with bottlenecks!

☺ Gains could be larger if end-to-end trained with bottlenecks!

Future Directions

- Non-idealized setups: finite sample case, constrained hypothesis?
- Approx. optimality: relaxed constraints?
- More practical methods for learning Z^*?
- Implicit regularization effect for learning Z^*?
- ...

Thank you!

Amazing co-authors:

Yann Dubois

Chris J. Maddison

References i

[1] T. Chen et al. **A simple framework for contrastive learning of visual representations.** In *ICML*, 2020.

[2] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. **Domain-adversarial training of neural networks.** *The journal of machine learning research*, 17(1):2096–2030, 2016.

[3] I. Gulrajani and D. Lopez-Paz. **In search of lost domain generalization.** In *ICLR*, 2021.

[4] A. Radford et al. **Learning transferable visual models from natural language supervision.** In *ICML*, 2021.

[5] C. Schuhmann et al. **Laion-400m: Open dataset of clip-filtered 400 million image-text pairs.** *arXiv preprint arXiv:2111.02114*, 2021.

[6] R. Taori et al. **Measuring robustness to natural distribution shifts in image classification.** *arXiv preprint arXiv:2007.00644*, 2020.