VECTOR INSTITUTE

Optimal Representations for Covariate Shift

Yangjun Ruan* [12] Yann Dubois* [2] Chris J. Maddison [12]

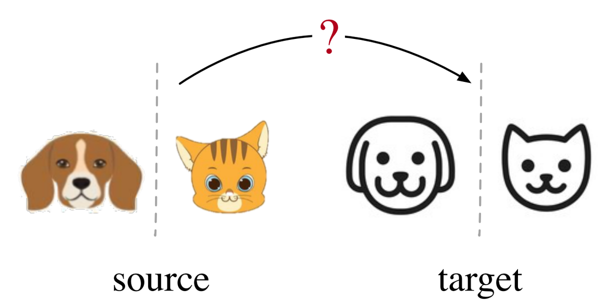
[1]University of Toronto [2]Vector Institute *Equal contribution

UNIVERSITY OF TORONTO

Overview

ML experiences **distribution shifts** from train (source) and test (target)

Goal: learn representations Z of data X from which source predictors perform well on target



Previous work:

☹ lack of theoretical characterization of optimal Z^*

☹ no practical methods uniformly outperform ERM [2]

Our work:

☺ prove **minimal sufficient** condition for optimal Z^*

☺ derive practical **SSL** objectives for learning Z^*

☺ show why CLIP [3] is so robust

☺ **SOTA** results on DomainBed!

Characterizing Optimally Robust Representations

Optimal Z^*: all source (d_s) optimal predictors achieve target (d_t) Bayes risk

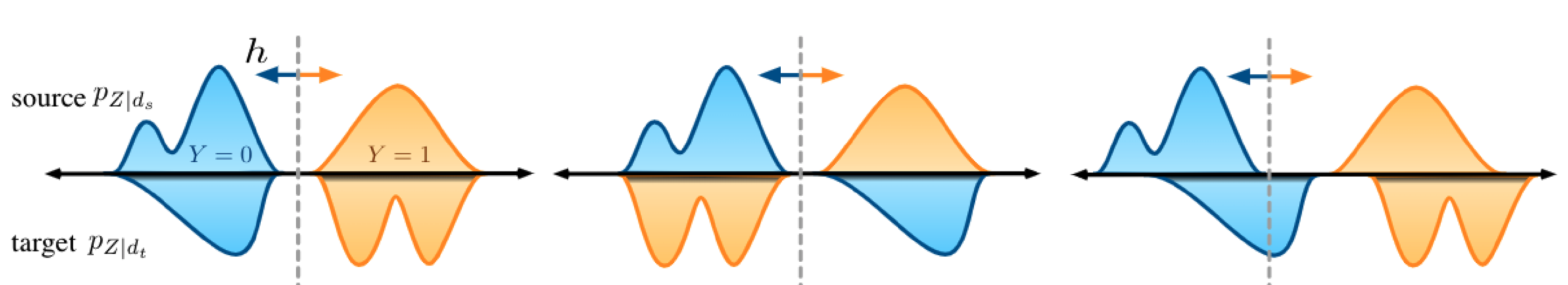
Goal: minimize the *idealized domain generalization* (IDG) risk w.r.t. Z

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

Theorem (Optimal coniditions)

Assume weak covariate shift, Z^* is **optimal if and only if** it

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



(a) discriminative & support match (b) only support match (c) only discriminative

☺ **achievable** sufficient and **necessary** condition

☺ provide **minimal sufficient** objectives for learning Z^*

☹ requires access to labeled target domain

No Free Lunch Without Target Information

Theorem (No free lunch)

Without accessing to target you **cannot learn useful** Z. You can construct many "bad" target domains where any Z will be **worse** than a constant C.

☺ explain the failure of current practical methods

☹ is getting access to targets realistic?

Learning Optimal Representations with SSL

Key idea: exploit large unlabeled data with self-supervised learning (SSL)

Proposition (Learning Z^* in practice)

One can learn optimal Z^* with **SSL** using:

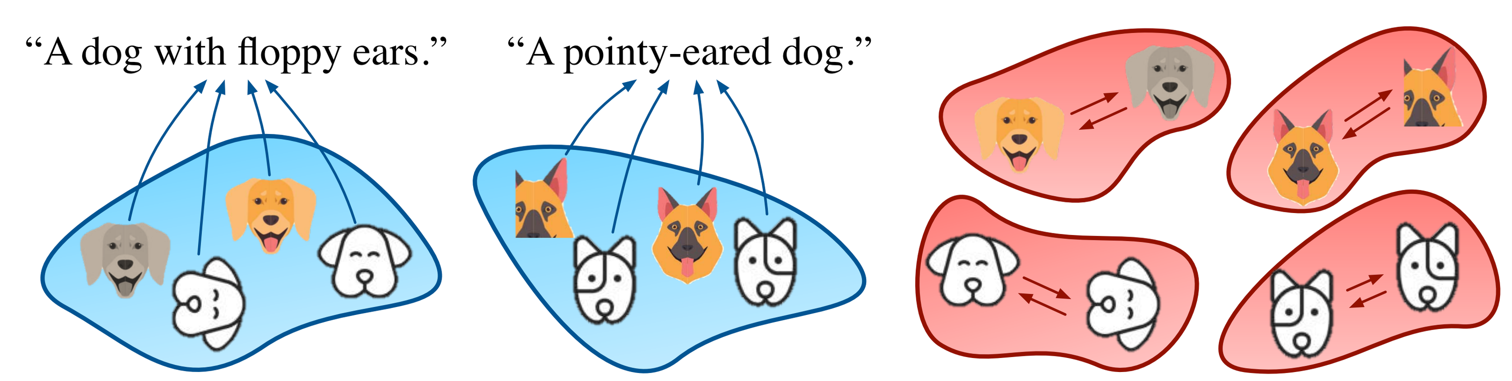
- **large-scale** unlabeled data
- contrastive learning with **domain-agnostic** augmentations
- **domain bottlenecks**

Domain-agnostic augmentations

- Require: **uncorrelated with domain**

✔ Example: **image-text** aug. (e.g., CLIP [3])

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



(a) image-text augmentations **(b)** standard augmentations

☺ explain the incredible robustness of CLIP over other SSL models

Domain bottleneck: enforce **support invariance**

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

☺ Requires **no trainable** domain classifier

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

☺ Requires **no access to domain** information

Exploiting Pretrained CLIP for Robust Representations

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 [2]

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	79.9 ± 0.1	52.8 ± 0.1
CLIP + CAD	**81.4 ± 0.8**	**94.7 ± 0.4**	**80.2 ± 0.2**	**54.1 ± 0.1**

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

Towards Generic Robust Representations with SSL

Idea: learn **task- and domain-agnostic** robust representations

- Task-agnostic: use large-scale data [4] with image-text contrastive loss
- Domain-agnostic: finetune CLIP with Ent bottleneck

Evaluate: natural distribution shift [5]

	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!

References

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

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

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

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

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