NEURAL INFORMATION PROCESSING SYSTEMS

When Does Group Invariant Learning Survive Spurious Correlations?

Yimeng Chen, Ruibin Xiong, Zhi-ming Ma, Yanyan Lan

Outline

- The question (motivation)
- 3 highlights of this paper
  - Two group criteria
  - Failures of existing methods
  - New method: SCILL
- Main experimental results
- Conclusion

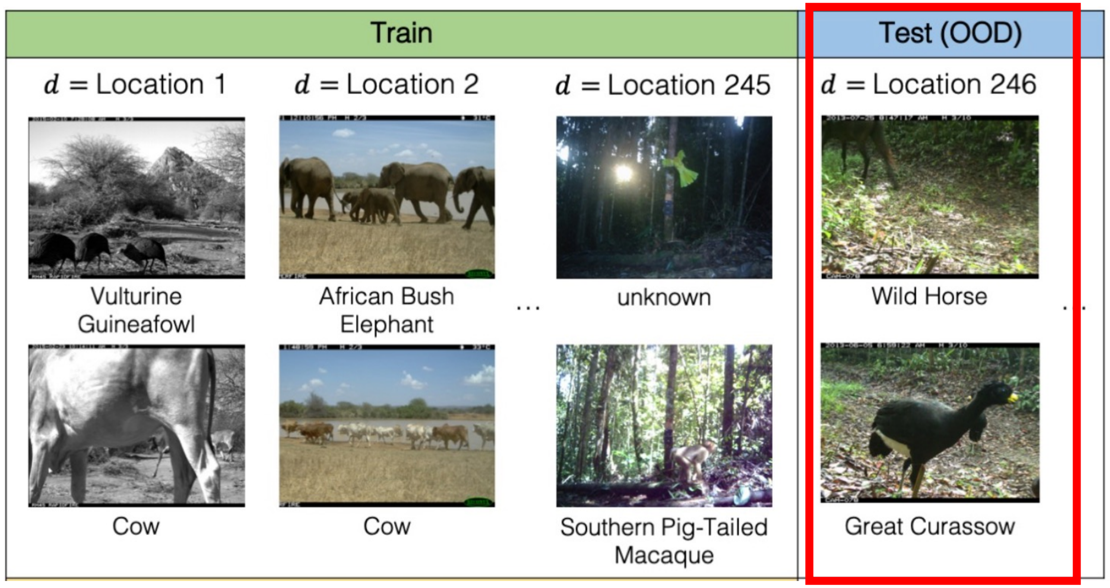
Code is available at:
https://github.com/Beastlyprime/group-invariant-learning

Motivation

When Does **Group Invariant Learning** Survive Spurious Correlations?

Invariant learning

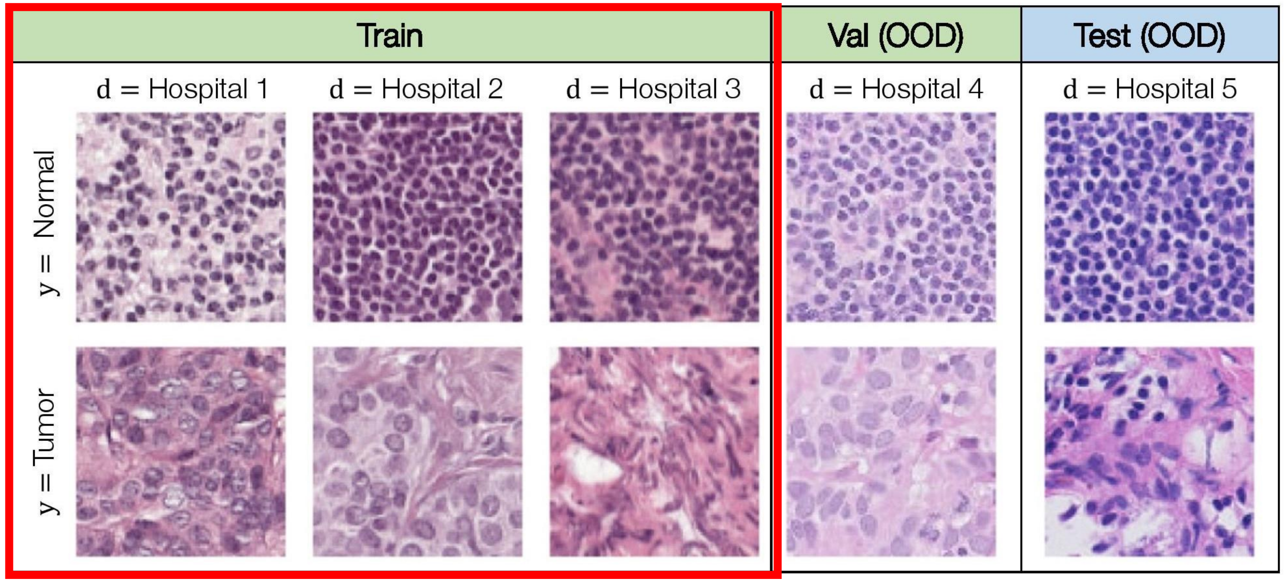
In real world applications, machine learning model encounters out-of-distribution (OOD) data



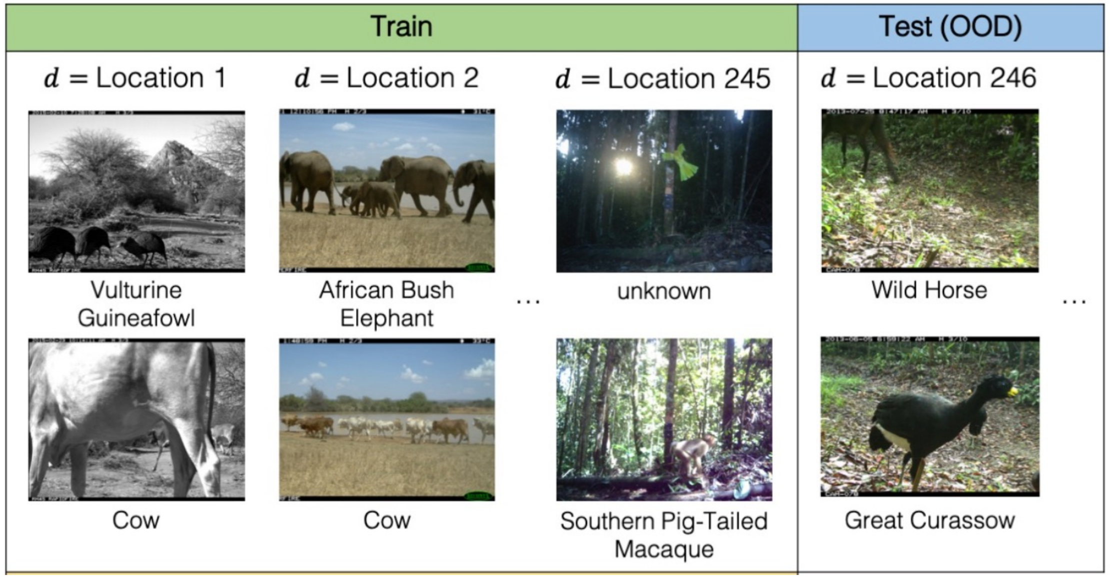
photos from new locations

Invariant learning and environments

Invariant learning: a notable kind of method for OOD generalization



samples from different hospitals



photos from different locations

Invariant learning is designed for the case when environment labels are available

Invariant learning and environments

Intuitively, the target is to learn the common rule on different environments

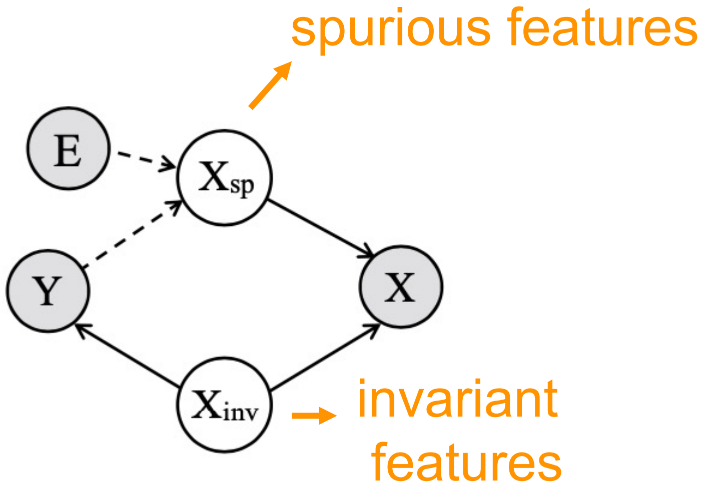
label sample

$$\mathbb{P}(Y|\Phi(X), E = e) = \mathbb{P}(Y|\Phi(X), E = e'), \forall e, e' \in \mathcal{E}.$$

feature encoder

Invariant learning and environments

Formally, invariance is deduced by assumptions on the data generating process



X_{sp}: the color green

X_{inv}: the shape "0"

$Y = 0$

X

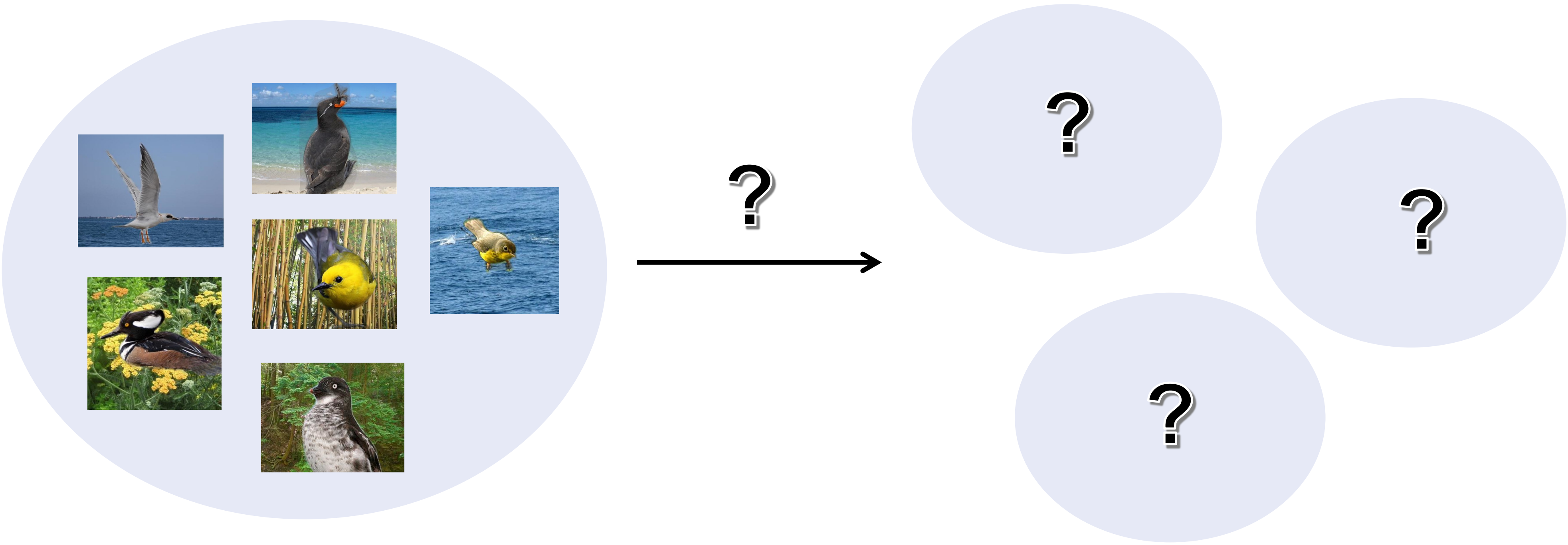
$\mathbb{P}^e(Y|X_{inv}) := \mathbb{P}(Y|X_{inv}, E = e)$ keeps invariant across different environments

Correlation between Y and X_{sp} is **spurious**, which changes across environments

Group invariant learning

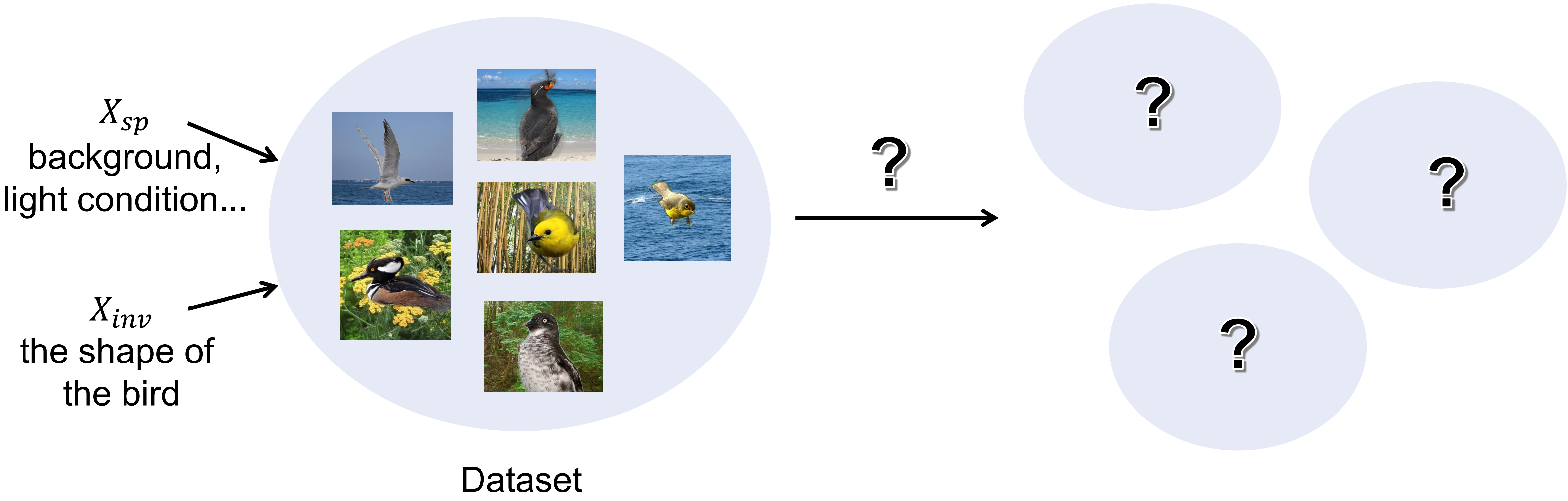
- Limitation: we need the environment labels are known
- "Group invariant learning" extend IL to the case when environments are unknown

Infer Environments for Invariant Learning



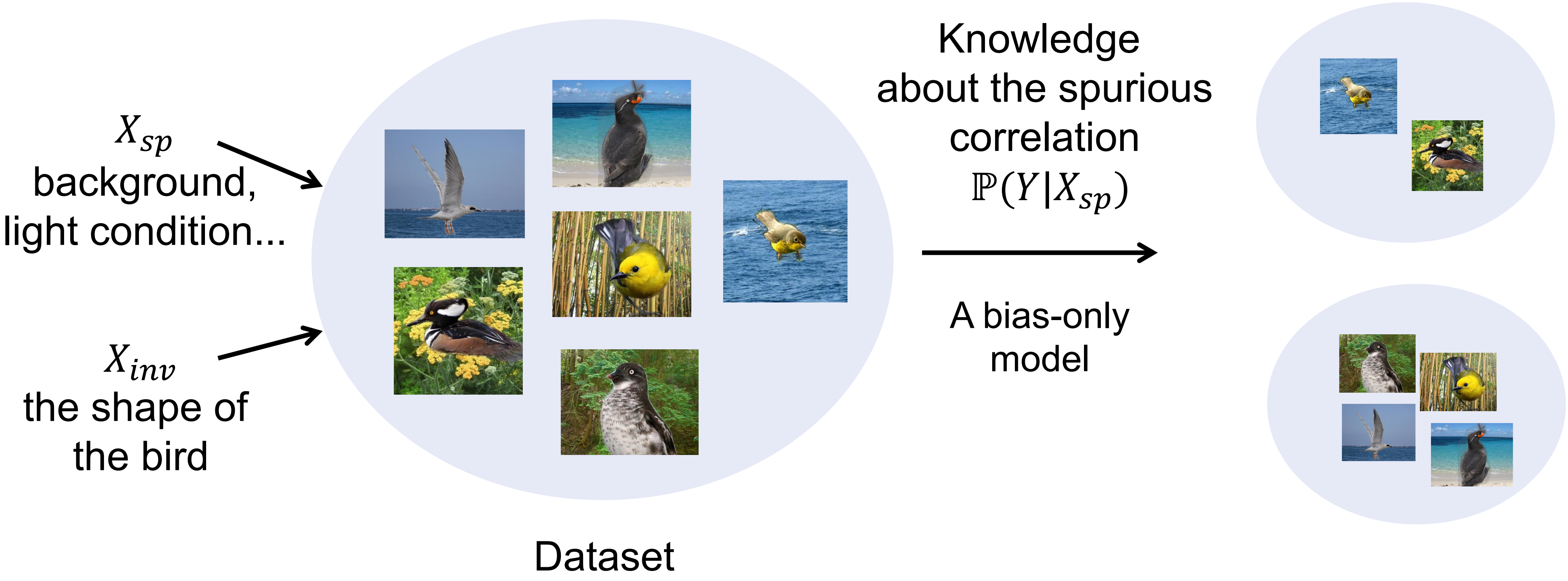
Dataset

Infer Environments for Invariant Learning



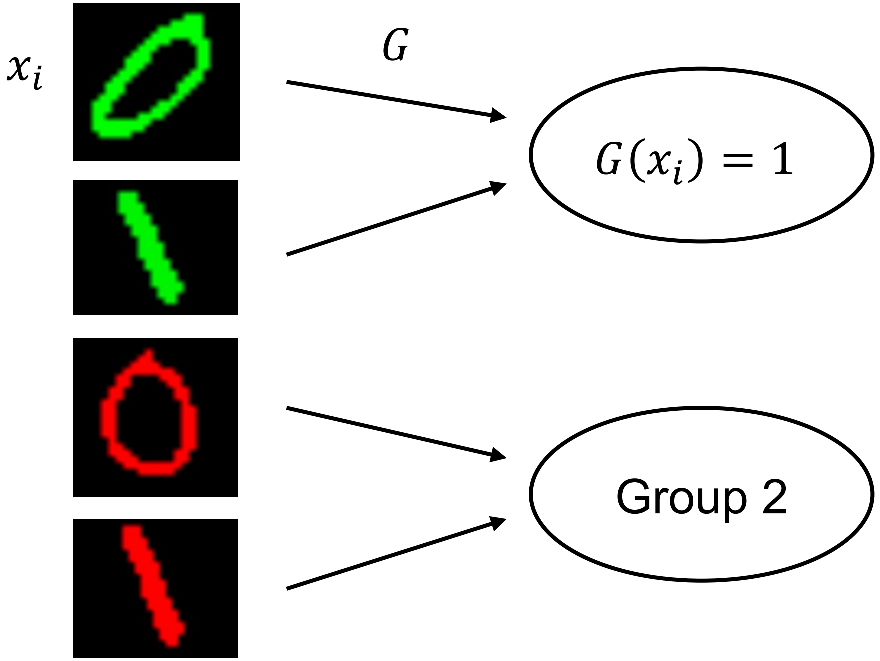
Infer Environments for Invariant Learning

Infer Environments for Invariant Learning



“Groups”: posterior environments

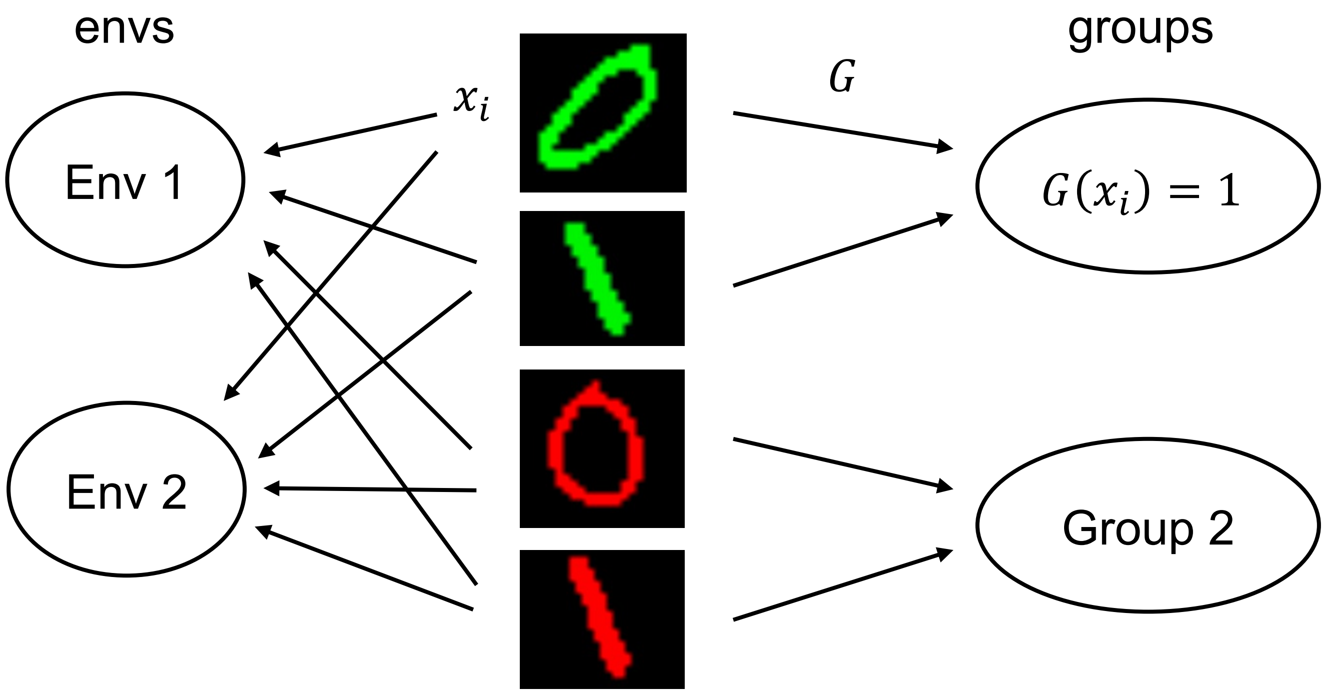
Group invariant learning extend IL to the case when environments are unknown

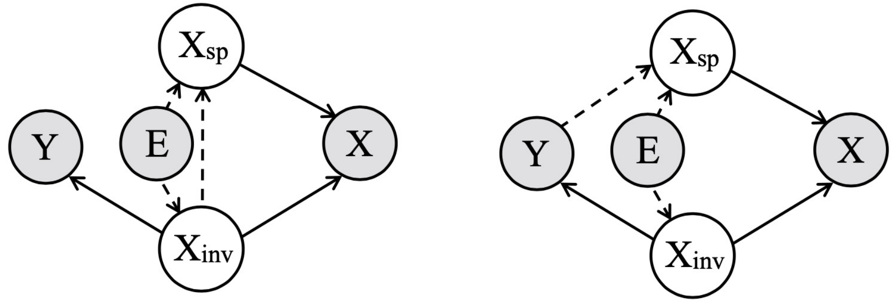


Infer posterior environments (groups) with knowledge about X_{sp} and Y

“Groups”: posterior environments

Groups are different from priori environments





(c) confounded [2; 25] (d) hybrid [39; 23]

Cannot be inferred without X_{inv}

We need new theory for group invariant learning

Highlight 1: Two group criteria

- Falsity exposure criterion:
  Groups should fully expose the falsity of spurious correlations (informally)

Criterion 4.1 (Falsity Exposure). For any $\sigma(X_{sp})$-measurable function h that satisfies $\forall g, g' \in \mathcal{G}$, $\mathbb{P}(Y|h(X_{sp}), g) = \mathbb{P}(Y|h(X_{sp}), g')$, it must satisfies $\mathbb{P}(Y|h(X_{sp})) = \mathbb{P}(Y)$.

- Label balance criterion:
  The label proportion between groups should be the same (informally)

Criterion 4.3 (Label Balance). For any $g, g' \in \mathcal{G}$ and $y, y' \in \mathcal{Y}$ with non-zero $\mathbb{P}(Y = y|g), \mathbb{P}(Y = y'|g), \mathbb{P}(Y = y|g')$ and $\mathbb{P}(Y = y'|g')$, the following equation holds.

$$\mathbb{P}(Y = y|g)/\mathbb{P}(Y = y'|g) = \mathbb{P}(Y = y|g')/\mathbb{P}(Y = y'|g') \quad (2)$$

Both criterion are necessary !

Highlight 2: Failures of existing methods

more ⟶ additional information ⟶ less

Existing methods	**clustering X_{sp}**	**clustering $P(Y\|X_{sp})$**	**majority/minority split**
Falsity exposure	No guarantee	✅	?
Label balance	No guarantee	❌	?

- We focus on the majority/minority split (EIIL, ICML2021)
- On **some dataset** (e.g. colored-MNIST), the majority/minority split satisfy both criteria
- In the presence of **multivariate** spurious features, it fails both criteria

Highlight 3: a new method SCILL

- Same as EIIL, it relies on a reference model f_r which approximates $\mathbb{P}^e(Y|X_{sp})$
- For falsity exposure:
  Construct groups such that $Y \perp f_r(X) | \, g$
- For label balance:
  Attach a weight $\omega^g(y) := \mathbb{P}(Y = y)/\mathbb{P}(Y = y|g)$ to samples in group g.
- Learning objective: $\mathcal{L}(f) := \sum_{g \in \mathcal{G}} \tilde{\mathcal{R}}^g(f) + \lambda \cdot \mathit{penalty}(\{S_g(f)\}_{g \in \mathcal{G}})$

Invariance penalty

Highlight 3: a new method SCILL

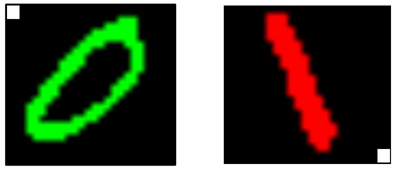
- Sufficiency of SCILL:

Theorem 5.1. *If $\mathcal{G}$ satisfies $f_r^*(X) \perp\!\!\!\perp Y|g, \forall g \in \mathcal{G}$, where $f_r^* : \mathcal{X} \to \mathcal{Y}$ is spurious-only, i.e. $\sigma(X_{sp})$-measurable, and minimizes the prediction loss $\mathcal{L}_{ce}^r = \mathbb{E}[\sum_y \mathbb{P}(Y = y|X) \log f_r(X)_y]$, the optimal model minimizing the objective (3) satisfies SFC.*

SCILL can survive spurious correlations with an ideal reference model

Experimental Results

Patched-Colored-MNIST
(PC-MNIST)

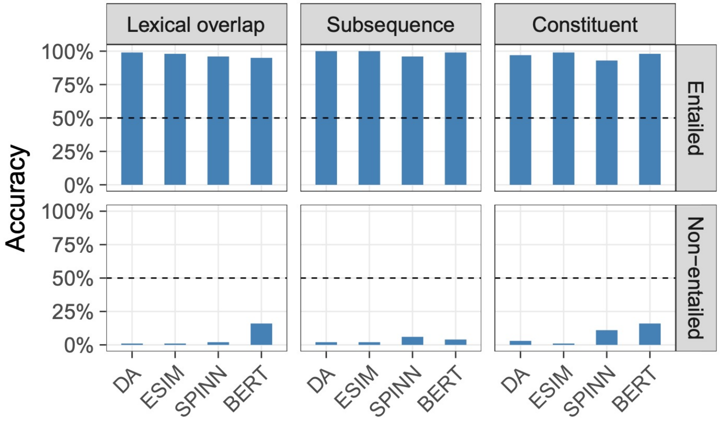


Two spurious features:
color and patch

MNLI-HANS

Heuristic	Supporting Cases	Contradicting Cases
Lexical overlap	2,158	261
Subsequence	1,274	72
Constituent	1,004	58

on MNLI, multiple syntactic
features and
the labels have spurious
correlations.



ERM models fail on HANS

Experimental Results

Patched-Colored-MNIST(PC-MNIST)

Method	Penalty	ID Val	ID Test	Oracle Val	Oracle Test	TEV Val	TEV Test
ERM	-	90.22 $\pm$ 0.56	50.64 $\pm$ 0.56	89.95 $\pm$ 0.45	54.53 $\pm$ 0.60	-	-
EIIL	IRM	90.21 $\pm$ 0.48	50.63 $\pm$ 0.45	78.01 $\pm$ 0.45	63.63 $\pm$ 0.71	69.81 $\pm$ 0.27	50.99 $\pm$ 0.58
	REx	90.24 $\pm$ 0.45	51.21 $\pm$ 0.64	79.10 $\pm$ 0.43	64.04 $\pm$ 0.80	70.05 $\pm$ 0.23	51.01 $\pm$ 0.68
	cMMD	90.24 $\pm$ 0.43	51.36 $\pm$ 0.61	77.27 $\pm$ 0.28	65.09 $\pm$ 0.63	70.15 $\pm$ 0.25	52.70 $\pm$ 1.40
	PGI	90.19 $\pm$ 0.46	51.07 $\pm$ 0.54	80.03 $\pm$ 1.41	64.27 $\pm$ 0.26	70.37 $\pm$ 0.14	50.64 $\pm$ 0.38
SCILL	IRM	79.65 $\pm$ 0.76	62.49 $\pm$ 0.55	71.54 $\pm$ 0.35	67.46 $\pm$ 0.19	71.54 $\pm$ 0.35	67.46 $\pm$ 0.19
	REx	80.23 $\pm$ 0.83	62.13 $\pm$ 0.99	72.59 $\pm$ 1.44	**67.60** $\pm$ 0.24	70.77 $\pm$ 0.50	67.33 $\pm$ 0.30
	cMMD	83.13 $\pm$ 0.93	59.76 $\pm$ 0.92	73.12 $\pm$ 0.47	67.49 $\pm$ 0.52	72.38 $\pm$ 0.51	**67.81** $\pm$ 0.34
	PGI	80.67 $\pm$ 1.75	**62.52** $\pm$ 0.32	71.73 $\pm$ 1.43	67.26 $\pm$ 0.14	71.35 $\pm$ 0.24	67.36 $\pm$ 0.33

Across 4 invariance penalties and 3 selection protocols, SCILL shows significant improvement

Experimental Results

MNLI-HANS

Method	Penalty	ID		Oracle		TEV	
		Val	Test	Val	Test	Val	Test
ERM	-	84.12 ± 0.15	64.88 ± 3.00	84.12 ± 0.15	64.88 ± 3.00	-	-
EIIL	IRM	84.01 ± 0.08	65.35 ± 0.93	83.82 ± 0.17	66.42 ± 0.98	84.01 ± 0.08	65.35 ± 0.93
	REx	84.10 ± 0.13	65.16 ± 0.19	83.91 ± 0.20	66.87 ± 2.92	84.00 ± 0.48	66.43 ± 1.00
	cMMD	83.56 ± 0.03	63.22 ± 1.76	83.22 ± 0.13	64.25 ± 1.63	83.38 ± 0.20	62.72 ± 2.03
	PGI	84.17 ± 0.08	65.57 ± 2.25	83.78 ± 0.03	66.02 ± 0.93	83.94 ± 0.64	65.57 ± 2.25
SCILL	IRM	82.75 ± 0.17	69.11 ± 1.76	82.56 ± 0.33	68.72 ± 1.24	82.67 ± 0.14	69.82 ± 1.29
	REx	82.68 ± 0.28	**69.73** ± 1.63	82.59 ± 0.22	**71.20** ± 1.81	82.56 ± 0.33	69.75 ± 1.53
	cMMD	82.74 ± 0.26	69.15 ± 1.39	82.39 ± 0.45	70.77 ± 1.40	82.61 ± 0.04	**70.92** ± 0.79
	PGI	82.79 ± 0.30	68.57 ± 0.54	81.69 ± 0.28	70.99 ± 0.48	82.79 ± 0.30	68.57 ± 0.54

Across 4 invariance penalties and 3 selection protocols, SCILL shows significant improvement

Conclusion

- The first theoretical study on group invariant learning
- Two criteria for group invariant learning to survive spurious correlations
- Failures of existing methods on multivariate spurious features
- New method guided by the two criteria: SCILL

Code is available at:
https://github.com/Beastlyprime/group-invariant-learning

Thanks for Your Attention !