Examining and Combating Spurious Features under Distribution Shift

Chunting Zhou[1], Xuezhe Ma[2], Paul Michel[1], Graham Neubig[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Information Sciences Institute, University of Southern California

Carnegie Mellon University School of Computer Science
Language Technologies Institute

NEULAB

Background

Abstract

1. Group Distributionally Robust Optimization can fail when groups do not directly account for various spurious correlations that occur in the data (*imperfect partition*).
2. We propose an effective method — group-conditional DRO that minimizes the worst-case losses over a more flexible set of distributions that are defined on the *joint distribution* of groups and instances, instead of treating each group as a whole at optimization time.

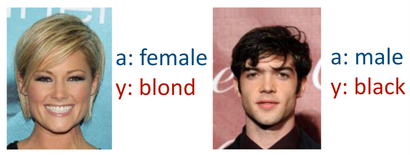
Code available at:
https://github.com/violet-zct/group-conditional-DRO

Problem: Poor worst-group performance
Models trained with empirical risk minimization (ERM) can latch on to spurious correlations in the training data and perform poorly on some groups.

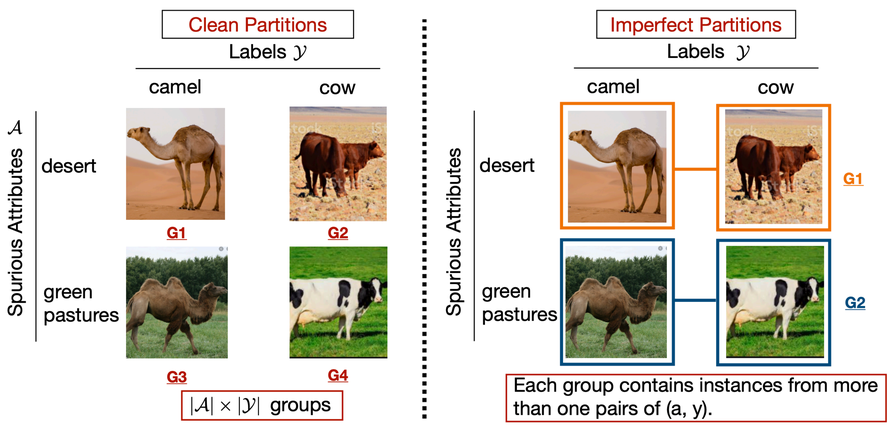
Toxicity Detection (Fortuna & Nunes., 18)
Twitter: "trump and his supporters can all burn in the pits of fucking hell."
Attribute (dialect): White-aligned / Hispanic / African American (AAE) / other
Label: abusive / normal / hateful / span **avg acc:** 79.7 **normal, AAE:** 34.3

MultiNLI (Williams et al., 18)
(P) Turned out, I wasn't completely wrong.
(H) I was 100 percent wrong.
Attribute: if negation word in Hypothesis
Label: entailment / negation / neutral
Minority group: no negation word, negation

CelebA (Liu et al., 15)



Perfect v.s. Imperfect Partitions



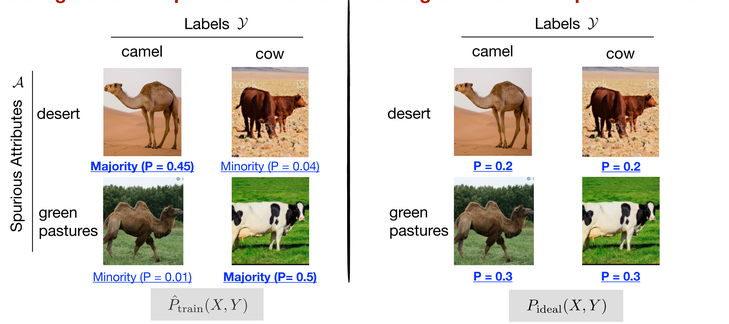
Imperfect partitions are common: (1) annotation is expensive (2) privacy concerns: sensitive attributes (3) spurious attributes are unknown

Examples of imperfect partitions: natural groups from topics/domains; groups from unsupervised clustering

Motivation and Approach: Group-Conditional Distributionally Robust Optimization

$\hat{P}_{\text{train}}(X,Y)$ versus $P_{\text{ideal}}(X,Y)$

$p(Y|a)$ is uniform distribution.



Group Distributionally Robust Optimization (group DRO)

$$\mathcal{L}_{GDRO}(\theta) = \sup_{q \in \mathcal{U}} \sum_{i \le N} q_i \mathcal{L}(\theta; g = i), \text{ where } \mathcal{U} \subset \Delta^{N-1}$$

Group DRO minimizes the worst expected loss over a set of potential test distributions $\mathcal{Q}$, which is an instance of DRO.

- **Defining $\mathcal{Q}$ that contains P_{ideal} is highly advantageous for learning robust features.**

Group DRO can fail under imperfect partitions:

- It's important to have a worst-case distribution q over the groups such that the spurious attribute no longer correlates with the labels.
- However, with imperfect partitions, the underlying conflicts prevent group DRO from formulating a worse-case distribution that can eliminate spurious correlations, i.e. $P_{\text{ideal}} \notin \mathcal{Q}$.

Examples: G1 — bg = desert; G2 — bg = green pastures
To prevent the model from learning spurious correlations between camel and desert, one would upweight G2; however, this exacerbates spurious correlations between green pastures and cows in G2.

Group-conditional DRO (GC-DRO)

- GC-DRO defines a more flexible uncertainty set over the *joint distribution of (x, y, g):*

$$\mathcal{Q}^{\alpha,\beta} = \left\{ q(g)q(x,y|g) : q(g) \le \frac{p_{\text{train}}(g)}{\alpha},\ \frac{1}{N} \le q(x,y|g) \le \frac{p_{\text{train}}(x,y|g)}{\beta}, \forall x,y,g \right\}$$

- *Efficient online greedy optimization*: interleave the updates of model parameters θ and q.

How and WHY Does GC-DRO Work?

Imperfect Partitions: (1) manually designed adversarial portions (2) supervised classifier (3) unsupervised clustering
Robust Acc: the worst accuracy across all groups (clean partitions of the test set)

Datasets	Methods	Clean Partition		Imperfect Partition	
		Robust Acc	Average Acc	Robust Acc	Average Acc
Celeb-A	ERM	40.14 ± 0.99	**95.92 ± 0.05**	40.14 ± 0.99	**95.92 ± 0.05**
	resampling	86.81 ± 1.26	92.72 ± 0.28	44.17 ± 1.15	95.58 ± 0.03
	group DRO	88.19 ± 2.31	92.65 ± 0.20	45.97 ± 1.73	95.81 ± 0.09
	GC-DRO	**88.75 ± 0.82**	92.92 ± 0.16	**82.85 ± 1.54**	89.32 ± 2.21
MNLI	ERM	70.84 ± 2.47	**86.18 ± 0.18**	70.84 ± 2.47	**86.18 ± 0.18**
	resampling	67.02 ± 2.43	85.72 ± 0.37	67.26 ± 1.63	85.22 ± 0.58
	group DRO	75.14 ± 3.96	85.82 ± 0.24	70.34 ± 2.19	86.02 ± 0.25
	GC-DRO	77.82 ± 1.45	85.04 ± 0.67	**75.32 ± 0.93**	84.82 ± 0.74
FDCL18	ERM	34.30 ± 1.83	**79.70 ± 1.05**	34.30 ± 1.83	79.70 ± 1.05
	resampling	55.44 ± 4.69	72.04 ± 1.99	26.10 ± 4.11	**80.66 ± 0.52**
	group DRO	56.83 ± 2.94	70.52 ± 1.99	36.24 ± 3.80	79.40 ± 1.12
	GC-DRO	**57.28 ± 2.71**	70.26 ± 0.94	**48.42 ± 6.72**	72.02 ± 2.96

How does GC-DRO perform?

- Under the clean partition, all the baseline methods outperform ERM greatly on the robust accuracy.
- Under the imperfect partition, baseline methods (resampling, group DRO) that leverage group information fail to perform well on the worst accuracy; GC-DRO still performs remarkably well due to the flexible weighting scheme.

Why does GC-DRO work? —— A study on MNLI

Groups in the imperfect partitions corresponds to cells in the same color in the left figure.

- Group DRO equally weighs examples in the same group of the imperfect partitions and pays less attention to minority groups.
- GC-DRO can handle sub-groups inside each group in a fine-grained way, which encourages the model to learn from minority groups that help combat spurious features.