Let Samples Speak: Mitigating Spurious Correlation by Exploiting the Clusterness of Samples

WeiWei Li[1], Junzhuo Liu[1], Yuanyuan Ren[2], Yuchen Zheng[2], Yahao Liu[1], Wen Li[1]

[1]University of Electronic Science and Technology of China [2]Shihezi University

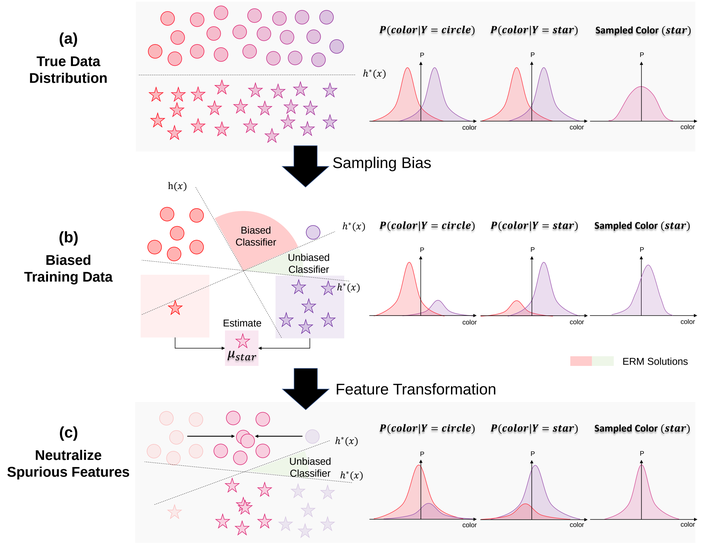
TL;DR

Project Page

- **Problem** DNNs learn false decision rules (bias) from the spurious features.
- **Observation** Samples affected by the spurious feature exhibit a dispersed distribution.
- **Solution** We develop an efficient (costing few minutes) debiasing pipeline of identifying, neutralizing, eliminating and updating, from this observation.
- **Results** Experiments show a +20% improvement over the ERM baseline.

Motivation

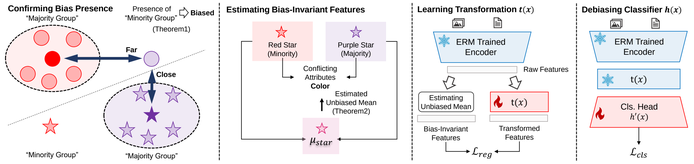
Deep learning models are known to often learn features that spuriously correlate with the class label during training but are irrelevant to the prediction task. 1) Ideally, bias attributes (e.g., color) should be evenly distributed and non-predictive of the class; 2) Sampling bias can introduce unintended patterns, like most circles being red and most stars being purple, causing some features to mistakenly correlate with class labels. Since ERM training minimizes the mean loss, an ERM-trained model is highly likely to fit these spurious correlations due to their large population in the data; 3) Intuitively, a transformation producing invariant representation for different values of bias attributes reduces the possible of learning bias.



Method

We introduce Neutralizing Spurious Features (NSF), a debiasing method that does not require prior knowledge of bias attributes. NSF consists of four key steps:

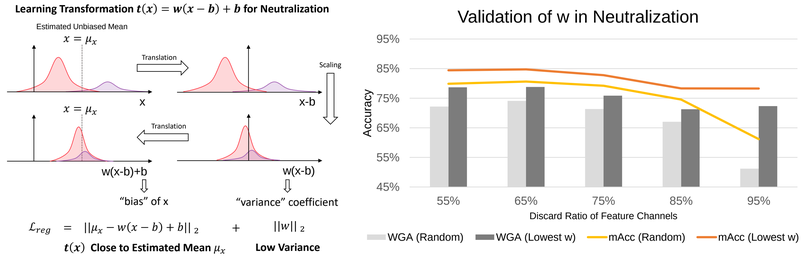
- *(1) Identifying Bias Presence*: Minority samples that deviate from the class centroid are identified, as such deviations indicate the presence of spurious features.
- *(2) Neutralizing Spurious Feature for Bias-Invariant Features*: Use identified groups to estimate a bias-invariant representation for each class.
- *(3) Eliminating Spurious Feature*: Learn a common transformation across all classes that aligns all training samples within a class to the estimated bias-invariant features. This transformation eliminates spurious features while preserving core features.
- *(4) Updating Classifier*: Finetune the classifier on these bias-invariant features, forcing reliance on core features alone.



Eliminating the Spurious Features with Channel-wise Transformation $t(x)$

Spurious features can be eliminated by learning a channel-wise transformation $t(\vec{x}) = \vec{w}(\vec{x} - \vec{b}) + \vec{b}$ where $\vec{w} \in R^{1\times D}$ and $\vec{b} \in R^{1\times D}$ to make all data points close to their corresponding conditioned mean value μ by minimizing

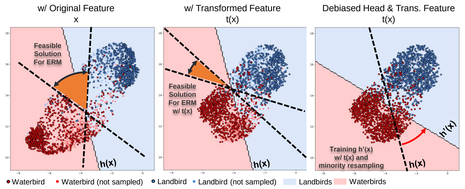
$$\mathcal{L}_{reg} = \lambda||\vec{w}||_2 + \frac{1}{N}\sum_{i}^{N} o_i||t(\mathbf{sg}[\vec{x}_i]) - \mu_{y_i}||_2. \quad (1)$$



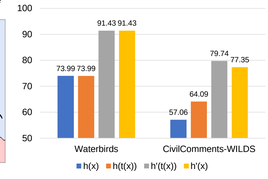
The WGA and mAcc of discarding channels by the lowest of the coefficient w in the transformation $t(x)$ are higher than choosing randomly, validating lower w highly correlate with spurious features so that they are eliminated.

Experiments

Features and the decision boundary using models trained on the waterbirds dataset, using group-balanced sampling for better visualization. 1) Original features $\vec{x}$ allows more biased solutions for the ERM training; 2) Elimination of spurious features using $t(\vec{x})$ leaves a smaller room for biased solutions; 3) Finetuning $h'(\vec{x})$ using $t(\vec{x})$ results in a unbiased classifier.



(a) Visualization of Features and Decision Boundary of Classifier

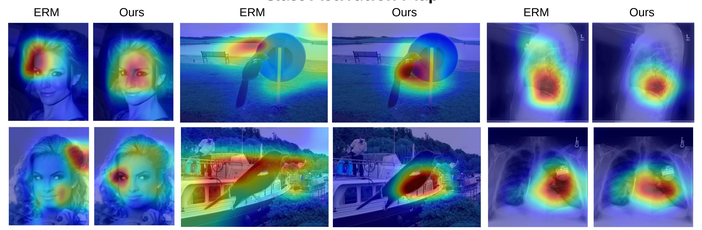


(b) Combination of Classifiers & Features

Ablation Studies

Combinations of classifiers of ERM $h(\vec{x})$ and debiased $h'(\vec{x})$, and the features of raw $\vec{x}$ and transformed $t(\vec{x})$. The debiased classifier $h'(\vec{x})$ performs well using the original features $\vec{x}$ indicating $h'(\vec{x})$ relies on core rather than spurious features.

Class Activation Map



(a) CelebA (b) Waterbirds (c) Chexpert

SOTA Comparison

	Labels		Waterbirds		CelebA		MultiNLI		CivilComments		Mean
	Tr.	Val	i.i.d.	WGA	i.i.d.	WGA	i.i.d.	WGA	i.i.d.	WGA	WGA
ERM	✗	✗	97.30	72.60	95.60	47.20	82.09	68.11	92.34	57.06	61.24
JTT	✗	✗	93.30	86.70	88.00	81.10	78.60	72.60	91.10	69.30	77.43
MT	✗	✗	93.00	86.40	91.30	78.00	——— Not Applicable ———				
CNC	✗	✗	90.90	88.50	89.90	**88.80**	-	-	81.70	68.90	-
AFR	✗	✗	94.20	90.40	91.30	82.00	81.40	**73.40**	89.80	68.70	78.63
Ours	✗	✗	95.65± 0.0011	**91.12**± 0.0063	88.70± 0.0036	84.27± 0.0047	80.43± 0.0003	73.12± 0.0008	87.19± 0.0010	**79.51**± 0.0022	**82.01**
GDRO	✓	✓	93.50	91.40	92.90	88.90	81.40	77.70	88.90	69.90	81.98