Out-of-Distribution Failure through the Lens of Labeling Mechanisms:

An Information Theoretic Approach

VECTOR INSTITUTE

Soroosh Shahtalebi[1], Zining Zhu[1,2], Frank Rudzicz[1,2,3]
[1]Vector Institute for Artificial Intelligence, [2]University of Toronto, [3]Unity Health Toronto

UNIVERSITY OF TORONTO

ICML Workshop on Spurious Correlations, Invariance, and Stability
July 2022

Table of Contents

Introduction

- Conventional machine learning models are developed based on the assumption that the test set is i.i.d with respect to the training set.
- This assumption is often violated in real-world problems
- The mismatch between the training and test distributions is of three types:

Covariate shift:	$P_{train}(X) \neq P_{test}(X)$
Correlation shift:	$P_{train}(Y\|X) \neq P_{test}(Y\|X)$
Label shift:	$P_{train}(Y) \neq P_{test}(Y)$

Introduction

- Neural networks typically fail to yield their optimum performance in domains with shifted distribution
- The reason for this failure is believed to be neural network’s inability in capturing generalizable and invariant features
- Our hypothesis is that the labeling mechanism employed by humans is also a contributing factor to this matter.
- Providing one label for a datapoint maximizes the risk that a model pick up a spurious correlation as the main differentiating feature for a classification task.

Intuition

Cow-Camel classification problem

vs.

Theory

Definitions

$$\text{empirical risk:} \quad \mathcal{L}_{emp}(A, S, R) = \frac{1}{n}\sum_{i=1}^{n} \ell(W, Z_i),$$
$$\text{population risk:} \quad \mathcal{L}(A, S, R) = \mathbb{E}_{Z' \sim D}\ell(W, Z'),$$

- S: a dataset of n *i.i.d* samples
- R: a random variable representing the stochasticity of data
- A: a learning algorithm that provides W, the parameters of the learning model, as a function of S and R
- Generalization gap:

$$|\mathbb{E}_{S,R}[\mathcal{L}(A, S, R) - \mathcal{L}_{emp}(A, S, R)]|$$

Theory

Theorem 2.1 (Xu & Raginsky (2017)). *If $\ell(w, Z')$, where $Z' \sim \mathcal{D}$, is $\sigma-$ subgaussian for all $w \in \mathcal{W}$, then*

$$|\mathbb{E}_{S,R}[\mathcal{L}(A,S,R) - \mathcal{L}_{emp}(A,S,R)]| \leq \sqrt{\frac{2\sigma^2 I(W;S)}{n}} \tag{3}$$

Theorem 2.2 (Harutyunyan et al. (2021)). *Let U be a random subset of $[n]$ with size m, independent of S and R. If $\ell(w, Z')$, where $Z' \sim \mathcal{D}$, is $\sigma-$ subgaussian for all $w \in \mathcal{W}$, then*

$$\begin{aligned} &|\mathbb{E}_{S,R}[\mathcal{L}(A,S,R) - \mathcal{L}_{emp}(A,S,R)]| \\ &\leq \mathbb{E}_{u\sim U}\sqrt{\frac{2\sigma^2 I(W;S_u)}{m}} \end{aligned} \tag{4}$$

Theory

Theorem 2.4. *Let $\ell(w, Z')$ is $\sigma - subgaussian$ for all $w \in \mathcal{W}$, and $Z' \sim \mathcal{D}$. Given a dataset S of n samples where each sample has K labels, for all $w \in \mathcal{W}$, the expected generalization bound is tighter by a factor of $\frac{1}{\sqrt{K}}$ than the case where each sample of a dataset with the same size has only 1 label. In other words,*

$$|\mathbb{E}_{S,R}[\mathcal{L}(A,S,R) - \mathcal{L}_{emp}(A,S,R)]| \\ \leq \sqrt{\frac{2\sigma^2 I(W;S)}{n}} = \sqrt{\frac{2\sigma^2 I(W;S)}{Km}}. \tag{5}$$

- Please note that across the single-label and multi-label scenarios, we assume that the number of parameters and the stochasticity of dataset does not change. What can be inferred from this theorem is that m=n/K number of multi-label training samples provide the same upper bound on the expected generalization gap that n number of single-label datapoints from the same distribution would do. In other words, given equal number of training examples from both scenarios, the upper bound of expected generalization gap for the multi-label scheme is $1/\sqrt{K}$ times tighter than the one of single-label case.

Results

Themes

- When the final label is among concepts (CelebA and Waterbirds)
- When final label is inferred from underlying concepts (Colored-MNIST)

- **Independent Bottleneck,** where the modules are trained independent from each other, i.e., $\hat{g} = \arg\min_g \sum_{i,j} L_y(g^j(x_i); y_i^j)$ and $\hat{f} = \arg\min_f \sum_i L_l(f(y_i); l_i)$.
- **Sequential Bottleneck,** where the concept bottleneck is trained first based on $\hat{g} = \arg\min_g \sum_{i,j} L_y(g^j(x_i); y_i^j)$, and then the inference module is trained on the outputs of the concept bottleneck, i.e., $\hat{f} = \arg\min_f \sum_i L_l(f(\hat{g}(x_i)); l_i)$.
- **Joint Bottleneck,** where the two modules are trained simultaneously based on a weighted sum of the loss for the two modules, i.e., $\hat{g}, \hat{f} = \arg\min_{g,f} \sum_i \left[L_l(f(g(x_i)); l_i) + \sum_j \lambda L_y(g^j(x_i); y_i^j)\right]$.

Results

Table 1: Accuracy of concept-based learning in OoD generalization over the Colored-MNIST dataset.

Method	Concept Accuracy			Label Accuracy		
	+90%	+80%	$\{+90\%\}\bigcup\{+80\%\}$	+90%	+80%	$\{+90\%\}\bigcup\{+80\%\}$
Independent	98.98	98.87	99.24	10.95	26.90	11.82
Sequential	98.82	98.89	99.35	**57.09**	**54.09**	**57.59**
Joint	98.93	99.07	99.16	12.93	27.01	13.00
ERM	50.55	26.18	74.32	17.08	29.82	28.51

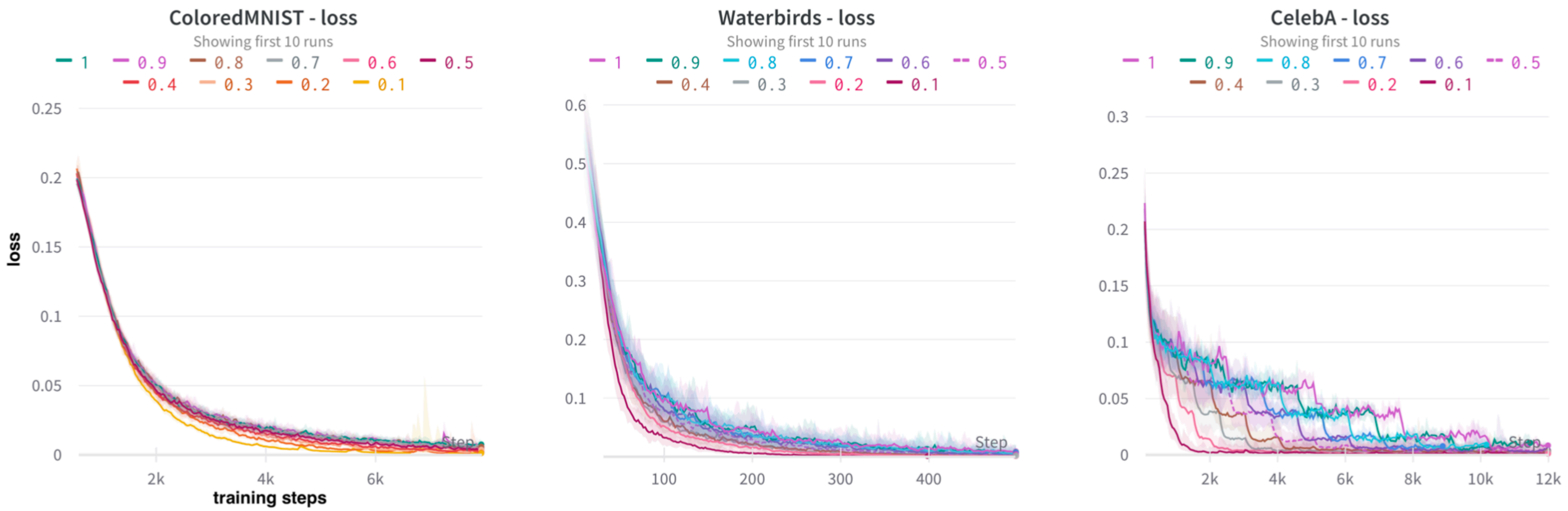
Results

Table 2: Accuracy of concept learning in OoD generalization over Waterbirds and CelebA datasets.

Model	Waterbirds		CelebA	
	Worst group	**Average**	**Worst group**	**Average**
GDRO (Sagawa et al., 2019)	83.80	89.40	88.30	91.80
ERM	60.00	97.30	41.10	94.80
VIB (Alemi et al., 2016)	75.31	95.39	78.13	91.94
CIM (Taghanaki et al., 2021)	73.35	89.78	81.25	89.24
CIM+VIB (Taghanaki et al., 2021)	77.23	95.60	83.59	90.61
Ours	**88.99**	91.85	**97.65**	**98.13**

Results

Sample efficiency



Thanks for your attention

Questions?

soroosh.shahtalebi@vectorinstitute.ai