Label-Efficient Group Robustness via Out-of-Distribution Concept Curation

Yiwei Yang[1] Anthony Z. Liu[2] Robert Wolfe[1] Aylin Caliskan[1] Bill Howe[1]
[1]University of Washington, [2]University of Michigan
{yanyiwei, rwolfe3, aylin, billhowe}@uw.edu anthliu@umich.edu

A. Experimental Details

A.1. Dataset Details

We further describe our evaluation benchmarks in detail.

CMNIST [1]: We used the same setup in [6]. The task is to classify MNIST digits into one of the five classes: $\mathcal{Y}$ = {(0,1), (2,3), (4,5), (6,7), (8,9)}. The spurious attribute is color. In the training data, p_{corr} fraction of each class's data points are colored with a corresponding color, and the rest is colored randomly from one of the five colors. The colors we used, in the order of the classes, are $\mathcal{A}$ = {#ff0000, #85ff00, #00fff3, #6e00ff, #ff0018}. The validation and test data are colored randomly with $a \in \mathcal{A}$. The training set of MNIST is split into training and validation according to a ratio of 80/20 split. The test set if the default test of MNIST. We set $p_{corr} = 0.995$ for the main result.

Waterbirds [3]: We used the same setup in [3]. The task is to classify birds into one of the 2 classes: $\mathcal{Y}$ = {waterbird, landbird}. The spurious attribute is background. The dataset is created by taking images of birds from the CUB dataset [5], and pasting on top of the images from the Places dataset [7]. Bird images are either waterbirds or landbirds; background images are either water or land. Waterbirds are either seabirds or waterfowl, and the rest are landbirds. Water backgrounds are either ocean or lake, and land backgrounds are either bamboo forest or broadleaf forest. We used the default training, validation, test splits from [3], where in training, 95% of waterbird images co-occur with 95% of water background images, and 95% of landbird images co-occur with 95% of land background images. In validation and test sets, waterbird and landbird images are evenly split between the water and land background images.

CelebA [2]: We used the same setup in [3]. The task is to classifiy celebrities' hair color $\mathcal{Y}$ = {blonde, not blonde}, correlated with the celebrities' identified gender $\mathcal{A}$ = {female, male}, respectively. We used the default training, validation, test splits from [3]. Blond men makes up for only 6% of the data.

A.2. Implementation Details

We now describe the models and hyperparameters used on each of the tasks. For all tasks: We used Logistic Regression from Scikit-learn (sklearn.linear_model.LogisticRegression) with max_iter=1000) for generating CAV ;we used Gaussian Mixture from Scikit-learn (sklearn.mixture.GaussianMixture) with default hyperparameters for inferring group labels.

- CMNIST: LeNet-5 CNN in the pytorch image classification tutorial (We used same hyperparameters as [6])
  - 1st stage: weight decay = 5e-4, learning rate = 1e-3, optimizer = SGD, momentum = 0.9, number of epochs = 5, batch size = 32, early stopping = False
  - 2nd stage: weight decay = 5e-4, learning rate = 1e-3, optimizer = SGD, momentum = 0.9, number of epochs = 20, batch size = 32, early stopping = True
- Waterbirds: ResNet-50 (torchvision.models.resnet50(pretrained=True)) (We used same hyperparameters as [4])
  - 1st stage: weight decay = 1e-4, learning rate = 1e-3, optimizer = SGD, momentum = 0.9, number of epochs = 300, batch size = 128, early stopping = True
  - 2nd stage: weight decay = 1.0, learning rate = 1e-5, optimizer = SGD, momentum = 0.9, number of epochs = 300, batch size = 128, early stopping = False
- CelebA: ResNet-50 (torchvision.models.resnet50(pretrained=True)) (We used same hyperparameters as [4])
  - 1st stage: weight decay = 1e-4, learning rate = 1e-4, optimizer = SGD, momentum = 0.9, number of epochs = 300, batch size = 128, early stopping = True
  - 2nd stage: weight decay = 0.1, learning rate = 1e-5, optimizer = SGD, momentum = 0.9, number of epochs = 300, batch size = 128, early stopping = False

B. Concept Prompt Quality

We now describe our reasoning for the choices of prompts used to generate the concept images. We design the prompts based on how much effort a hypothetical practitioner would

	Precision(%)	Recall(%)
CMNIST	98.2	98.2
Waterbirds	88.6	86.4
CelebA	85.2	82.3

Table 1. We show the average precision and recall of the pseudo-group labels inferred by CDRO compared to the ground truth. We show CDRO effectively infers group labels.

spend on curating the concepts. For example, to curate the concept *Land Background* and the contrastive set *Water Background*, a practitioner might simply use the prompt "A photo of land background" and "A photo of water background." If the practitioner browses through some examples of the dataset and learns that the dataset consists of waterbird and landbird images, they might use prompts "A photo of landbird habitat" and "A photo of waterbird habitat." But the practitioner may know that the land backgrounds are primarily either bamboo or broadleaf trees and that the water backgrounds are primarily either ocean or lake settings, they might use the more specific prompts "A photo of bamboo trees" and "A photo of broadleaf trees" for land backgrounds and "A photo of ocean" and "A photo of lake" for water backgrounds. Lastly, we also drew images with land background and water background from the training data itself as a best-case baseline. We refer to these strategies as *distant from training distribution*, *somewhat near training distribution*, *near training distribution*, *in distribution*, respectively.

For CelebA, we designed the prompts in a similar manner, such that prompts with more information about the training data correspond to concepts that are nearer or further from the training data distribution. Specifically, ordering from *distant* to *near*, we used prompts "A photo of a female/male person", "A photo of a female/male celebrity", and "A photo of a female/male celebrity face" for concept sets *female* and *male*, respectively. We also drew labeled male and female images from the training data as a baseline.

C. Concept DRO Algorithm

We present our algorithm 1 for Concept DRO.

D. Additional Results on Effectiveness of CDRO in Inferring Group Labels

In figure 1, we present the distributions of the train data by their cosine similarity scores to the corresponding CAV for all three datasets.

In table 1, we show that CDRO estimates group labels with high precision and recall on the validation set in all three datasets.

Algorithm 1 Concept DRO

Input: Training dataset (X_{train}, Y_{train}), validation dataset (X_{val}, Y_{val}), spurious attribute $\mathcal{A} = \{a_1, ..., a_m\}$, for each a_i, a concept set $C_i = \{c_{i1}, ..., c_{ik}\}$, and a contrastive set $N_i = \{n_{i1}, ..., n_{ik}\}$
Stage 1: Infer group labels
Train an ERM model on (X_{train}, Y_{train}), take the first to penultimate layer as the feature extractor f
for $a_i \in \mathcal{A}$ **do**
    Train a linear classifier on features of the corresponding concept and contrastive set, $f(C_i)$ and $f(N_i)$, the coefficient of the classifier gives us the CAV_i in the direction of the concept a_i
    for $X \in \{X_{train}, X_{val}\}$ **do**
        Compute cosine similarity between CAV_i and $f(X)$, which gives us similarity scores between the data representations and CAV_i, denoted as $s_{CAV}(X)$
        Train a GMM on $s_{CAV}(X)$ with m mixtures
        Label the spurious attribute of data points in the mixture of the highest mean as a_i
    end for
end for
For the unlabeled data points, randomly assign a_i from $\mathcal{A}$
Stage 2: Optimize for worst-group loss
Use Group DRO with the inferred labels to optimize for worst-group loss

E. Images Used to Learn Concepts

We present image samples used to learn concepts for both the Waterbirds and CelebA datasets.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1

[2] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 1

[3] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 1

[4] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020. 1

[5] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona,

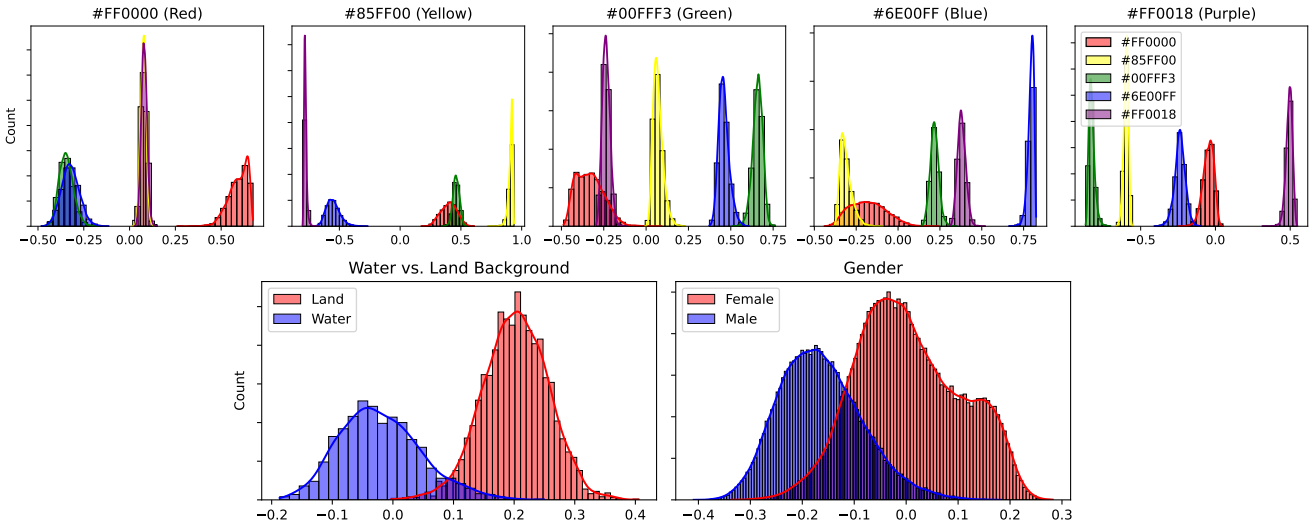

Figure 1. Distributions of similarity between the CAV trained with out-of-distribution concepts and training samples. Each peak is colored by the ground truth group label. The peaks are separable, and the peak furthest to the right (highest similarity to the concept) corresponds to the ground truth label in all cases, indicating how we can infer spurious attribute labels. Top row: CMNIST train samples for each of five color concepts. Bottom left: Waterbird train samples with land as the concept set and water as the contrastive set. Bottom right: CelebA train samples with women as the concept set and men as the contrastive set.

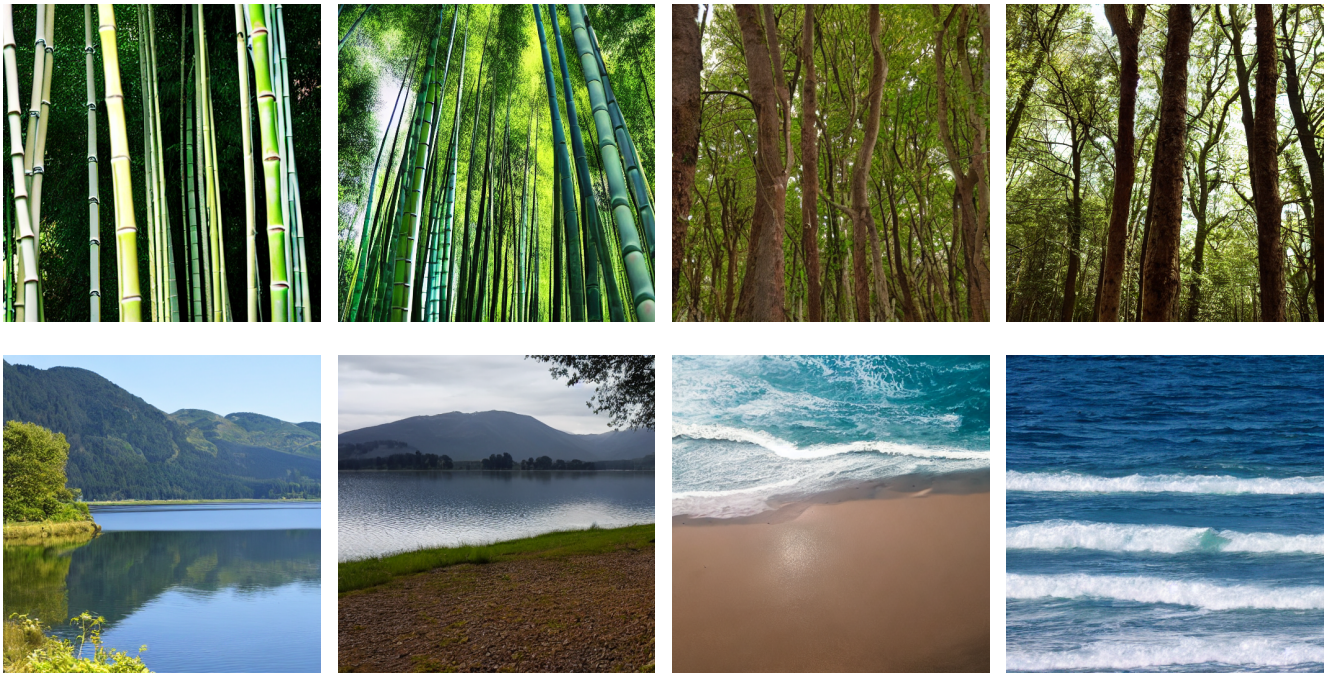

Figure 2. Sampled Stable Diffusion generated images used for the Waterbirds dataset. The first four images are generated with prompts "A photo of bamboo trees" (images 1 and 2) and "A photo of broadleaf trees" (images 3 and 4) to represent the concept of "land background". The latter four images are generated with prompts "A photo of lake" (images 5 and 6) and "A photo of ocean" (images 7 and 8) to represent the concept of "water background".

and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

[6] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022. 1

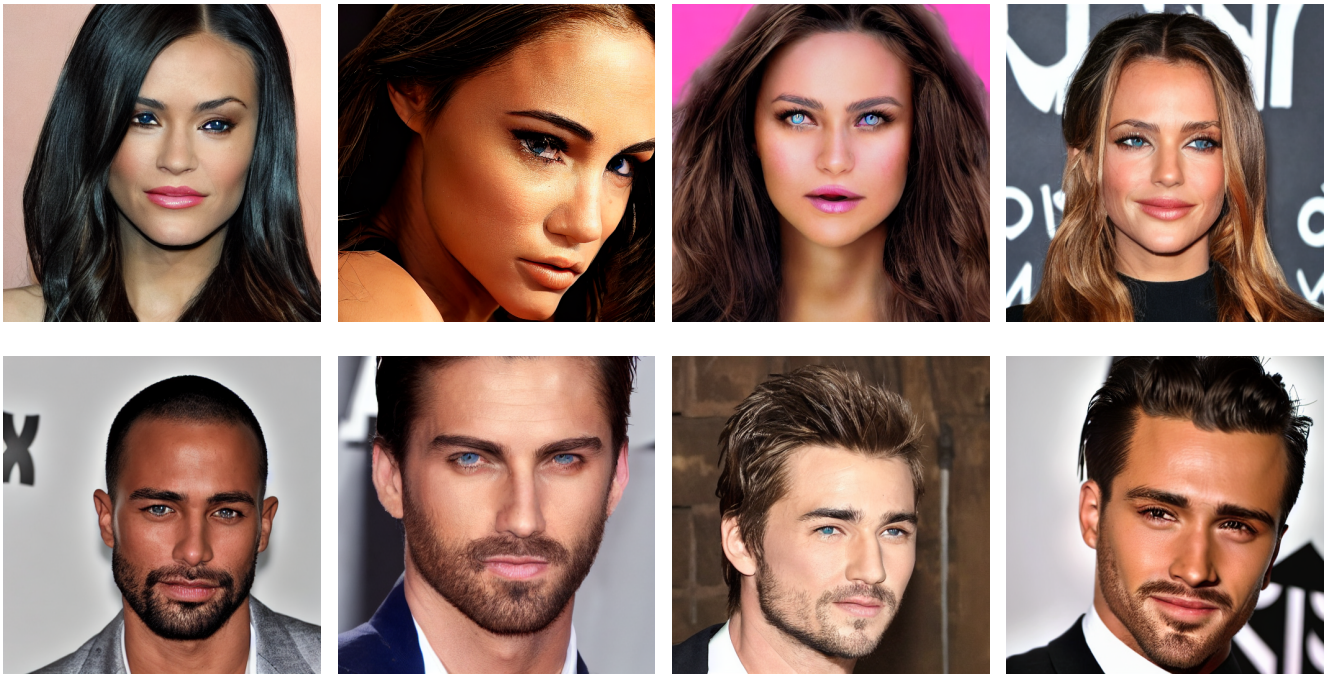

Figure 3. Sampled Stable Diffusion generated images used for the CelebA dataset. The first four images are generated with prompts “A photo of a female celebrity face” to represent the concept of “femaleness”. The latter four images are generated with prompts “A photo of a male celebrity face” to represent the concept of “maleness”.

[7] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014. 1