Towards Mitigating Spurious Correlations in Image Classifiers with Simple Yes-no Feedback

Seongmin Lee [1] **Ali Payani** [2] **Duen Horng (Polo) Chau** [1]

Abstract

Modern deep learning models have achieved remarkable performance. However, they often rely on spurious correlations between data and labels that exist only in the training data, resulting in poor generalization performance. We present CRAYON (**C**orrelation **R**ectification **A**lgorithms by **Y**es **O**r **N**o), effective, scalable, and practical solutions to refine models with spurious correlations using simple yes-no feedback on model interpretations. CRAYON addresses key limitations of existing approaches that heavily rely on costly human intervention and empowers popular model interpretation techniques to mitigate spurious correlations in two distinct ways: CRAYON-ATTENTION guides saliency maps to focus on relevant image regions, and CRAYON-PRUNING prunes irrelevant neurons to remove their influence. Extensive evaluation on three benchmark image datasets and three state-of-the-art methods demonstrates that our methods effectively mitigate spurious correlations, achieving comparable or even better performance than existing approaches that require more complex feedback.

1. Introduction

Modern deep learning models have achieved remarkable performance, surpassing humans in image classification tasks (He et al., 2015). However, these models often rely on spurious correlations between data and labels that exist only in the training data, resulting in poor generalization performance (Geirhos et al., 2018; Beery et al., 2018; Arjovsky et al., 2019; Sagawa et al., 2019; Geirhos et al., 2020; Singla et al., 2021; Singla & Feizi, 2021). For example, if a model is trained to classify *smiling* and *not smiling* faces using a dataset where the majority of smiling people coincidentally have black hair, the model often bases its predictions on hair color, which is irrelevant to smiling (Krishnakumar et al., 2021). It is important to refine these models so that their predictions leverage relevant data features (Oakden-Rayner et al., 2020; Rodolfa et al., 2021).

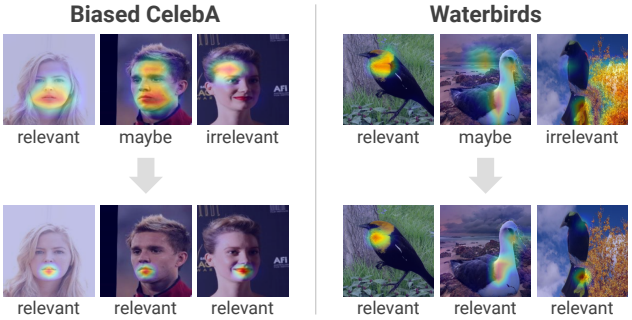

Figure 1. Our method corrects the model to focus on the relevant regions of an image. Here we shown an example of fixing a smile classifier that occasionally wrongly attends to a person's forehead, so that it now attends to the mouth regions. Similarly, our method adjusts the attention of a bird classifier, shifting it from irrelevant backgrounds to relevant bird bodies.

There have been many efforts to address spurious correlations in image classifiers by introducing new loss functions (Kim et al., 2019; Singla & Feizi, 2021; Zhang et al., 2021; Liu et al., 2022; Zhang et al., 2022; Asgari et al., 2022) and improving the balance of training data (Li & Vasconcelos, 2019; Nam et al., 2020; Liu et al., 2021; Kirichenko et al., 2022; Minderer et al., 2020; Lee et al., 2021; Kim et al., 2021; Chiu et al., 2022; Yao et al., 2022; Han et al., 2022). However, most of these approaches require explicitly identifying the attributes incorrectly attended to by the model. Furthermore, balancing the training dataset poses challenges, especially when there is limited or no data available without spurious correlations. To overcome these limitations, some researchers have involved humans in the training process (Ross et al., 2017; Schramowski et al., 2020; Plumb et al., 2021; Hagos et al., 2022; Gao et al., 2022b;a; Rao et al., 2023). They collect ground truth attention maps, which indicate where the model should or should not fo-

[1]Georgia Institute of Technology, GA, USA [2]Cisco Systems Inc., CA, USA. Correspondence to: Seongmin Lee <seongmin@gatech.edu>, Ali Payani <apayani@cisco.com>, Duen Horng (Polo) Chau <polo@gatech.edu>.

AI & HCI Workshop at the 40^{th} International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023.

cus, and guide the models' saliency maps to resemble the ground truth. However, most of these methods make the major assumption that humans can easily provide accurate ground truth maps, which can be costly and time-intensive to obtain (Gao et al., 2022a).

To address the above research gaps, our ongoing research presents CRAYON (**C**orrelation **R**ectification **A**lgorithms by **Y**es **O**r **N**o), which makes the following contributions:

- **Yes-No Feedback as a Simple, Practical Strategy to Mitigate Spurious Correlations.** We present the major idea that simple yes-no feedback on model interpretations can offer effective, scalable, and practical solutions for refining models with spurious correlations, addressing key limitations of existing approaches that heavily rely on costly human intervention. Our strategy empowers popular model interpretation techniques to mitigate spurious correlations in two distinct ways:
  1. **Guiding saliency maps to focus on relevant image regions**. (Sec. 3.2) We propose CRAYON-ATTENTION, a method that refines a model's ability to attend to the relevant regions of images by incorporating yes-no feedback on the relevance of saliency maps from the original unrefined models. CRAYON-ATTENTION guides the model to attend to the regions that are not highlighted in irrelevant maps, while preserving attention on the relevant saliency maps (Fig. 1).
  2. **Pruning irrelevant neurons to remove their influence.** (Sec. 3.3) CRAYON-PRUNING identifies irrelevant neurons in the penultimate layer of a model by presenting the visual concepts responsible for highly activating each neuron (Fig. 2). These irrelevant neurons are then pruned so that the model's predictions are not influenced by irrelevant regions.
- **Extensive evaluation on three benchmark image datasets against three state-of-the-art methods.** (Sec. 4) We demonstrate that our methods effectively mitigate spurious correlations, achieving comparable or even better performance when compared to competitors that rely on more complex feedback.

2. Related Work on Mitigating Spurious Correlations

Spurious correlations in deep learning models are often attributed to imbalances in training datasets (Sagawa et al., 2020). Efforts have been made to mitigate these spurious correlations by reweighting training data (Li & Vasconcelos, 2019; Nam et al., 2020; Liu et al., 2021; Kirichenko et al., 2022). However, these methods face challenges when there is limited or no data available without spurious correlations. To address this, some researchers have opted to collect or generate additional data to create balanced training datasets (Minderer et al., 2020; Lee et al., 2021; Kim et al., 2021; Chiu et al., 2022; Yao et al., 2022; Han et al., 2022). Nevertheless, obtaining such data is often costly or impractical in real-world scenarios. Moreover, most of these methods require the attributes responsible for the spurious correlations to be predefined.

A growing amount of research incorporated human involvement in the iteration of model training. The RRR loss (Ross et al., 2017) has been proposed to guide MLP models in avoiding irrelevant regions and has later been extended to deeper CNN models (Gao et al., 2022b;a; Hagos et al., 2022). Methods such as CDEP (Rieger et al., 2020) and SPIRE (Plumb et al., 2021) aim to minimize the importance of the irrelevant pixels by exploiting contextual decomposition and masking specific objects in images, respectively. Stammer et al. (Stammer et al., 2021) revise a model at both the pixel and concept levels by disentangling concepts in an image (Stammer et al., 2021). However, all of these methods require ground truth saliency maps for each data, which are often extremely costly to annotate. Collecting bounding boxes instead of the maps have partially addressed this challenge (Rao et al., 2023). In parallel, an explanatory interactive learning workflow has been introduced, where humans are asked to revise a DNN model based on the model's interpretations (Schramowski et al., 2020).

3. Methods

3.1. Overview

We aim to mitigate spurious correlations in a trained model by leveraging yes-no feedback on the relevance of the model's prediction reasoning. In this section, we introduce two methods: CRAYON-ATTENTION and CRAYON-PRUNING. CRAYON-ATTENTION uses yes-no feedback to guide the model's saliency maps to highlight the relevant regions of each image (Sec. 3.2). CRAYON-PRUNING identifies and prunes the neurons activated by irrelevant visual concepts (Sec. 3.3).

3.2. Refining with Saliency Maps

One of the most commonly used model interpretation methods is the generation of saliency maps (Simonyan et al., 2013; Selvaraju et al., 2017). For a model f and its training data $\mathbf{x}_1, \ldots, \mathbf{x}_N$, the saliency map $M_{\mathbf{x}_n}$ highlights the regions in the image $\mathbf{x}_n$ that the model f focuses on for its prediction. After generating the saliency maps for all N training data, we collect yes-no feedback on whether each map highlights relevant regions for the prediction task. We also provide a *maybe* option for maps that are difficult to evaluate. We denote the set of the indices of the training data

with relevant and irrelevant maps as R and I, respectively.

To refine the model f using the collected feedback, we modify the RRR loss (Ross et al., 2017), which guides the model to generate correct saliency maps with ground truth maps. For the data point $\mathbf{x}_n$ whose saliency map $M_{\mathbf{x}_n}$ highlights the relevant regions, the model f should generate similar saliency maps after the refinement. Therefore, we design the loss function $\mathcal{L}_{rel,n}$ as follows:

$$\mathcal{L}_{rel,n} = \sum_{h=1}^{H}\sum_{w=1}^{W} [M'_{\mathbf{x}_n}]_{hw}(1 - [M_{\mathbf{x}_n}]_{hw}) \quad (1)$$

where H and W are the height and width of the saliency maps, respectively, and $M'_{\mathbf{x}_n}$ is the saliency map for the model f' being trained and the data point $\mathbf{x}_n$. For better stability of the loss function, we normalize both $M_{\mathbf{x}_n}$ and $M'_{\mathbf{x}_n}$ to have values ranging from 0 to 1 by dividing each map with its maximum value.

On the other hand, for the data $\mathbf{x}_n$ with irrelevant saliency maps, the model should attend to the regions that are not highlighted in the map $M_{\mathbf{x}_n}$. We construct the loss function $\mathcal{L}_{irrel,n}$ as follows:

$$\mathcal{L}_{irrel,n} = \sum_{h=1}^{H}\sum_{w=1}^{W} [M'_{\mathbf{x}_n}]_{hw}[M_{\mathbf{x}_n}]_{hw} \quad (2)$$

We do not use the data with *maybe* feedback since the characteristics of their maps are unclear.

While guiding the model to attend to the right regions, we should keep the model prediction correct. Therefore, we add the prediction loss $\mathcal{L}_{pred,n}$ for the data point $\mathbf{x}_n$:

$$\mathcal{L}_{pred,n} = \sum_{k=1}^{K} -y_{nk} \log \hat{y}_{nk} \quad (3)$$

where y_{nk} is 1 if the label of the data $\mathbf{x}_n$ is k and 0 otherwise and $\hat{y}_{nk}$ is the probability of the data $\mathbf{x}_n$ being labeled as k computed by the model f' being trained.

Summing up the loss functions, we obtain the loss $\mathcal{L}_{map}$ that guides a model with yes-no feedback on saliency maps,

$$\mathcal{L}_{map} = \sum_{n=1}^{N} \mathcal{L}_{pred,n} + \alpha \sum_{n\in R} \mathcal{L}_{rel,n} + \beta \sum_{n\in I} \mathcal{L}_{irrel,n} \quad (4)$$

where α and β are the hyperparameters that control the weights of the loss terms.

3.3. Pruning Irrelevant Neurons

Neurons, also referred to as *channels*, in the penultimate layer of CNN models are known to be activated by specific high-level visual concepts in the input data (Bengio et al., 2013; Mahendran & Vedaldi, 2016). The visual concepts responsible for activating a neuron can be determined by investigating the regions of the neuron that are activated by different images (Hohman et al., 2020). Following the methods proposed in existing work (Hohman et al., 2020), we summarize the visual concepts responsible for a neuron's activation as a collection of image patches. These patches are generated by selecting the images that activate the neuron most strongly and cropping out the corresponding region (Fig. 2). For example, in a smile classifier, a neuron in the penultimate layer would have patches corresponding to the *mouth* concept, indicating that the neuron's activation is attributed to the presence of a mouth On the other hand, the patches of another neuron in the same model might indicate that its activation is attributed to *hair*.

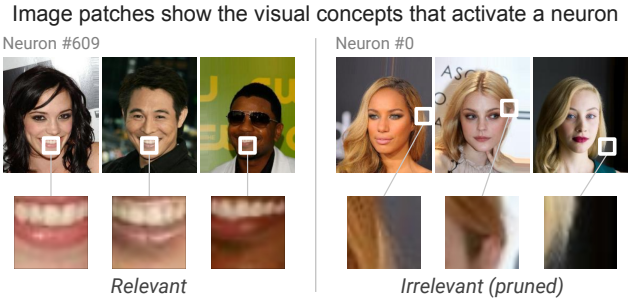

Figure 2. For each neuron in the penultimate layer of a smile classifier, we generate image patches that summarize the visual concepts responsible for the activation of the neuron. **Left:** Among these neurons, neuron #609 is activated by *mouth* patches relevant to smile classification. **Right:** Neuron #0, on the other hand, is activated by irrelevant *hair* patches. CRAYON-PRUNING prunes the neurons activated by irrelevant concepts in the penultimate layer and fine-tunes the last layer.

CRAYON-PRUNING identifies the neurons in the penultimate layer that are activated by irrelevant visual concepts by presenting the image patches of each neuron and collecting yes-no feedback on their relevance. For instance, for the smile classifier in Fig. 2, the neuron activated by the *mouth* concept is relevant while the neuron activated by the *hair* concept is irrelevant. We then prune the irrelevant neurons and fine-tune the last fully-connected layer of the model to remove the effect of the irrelevant concept on the model prediction. For this fine-tuning process, we use the prediction loss in the Equation 3.

4. Evaluation

As a litmus test of our hypothesis that using only yes-no feedback can provide comparable or even better performance than the existing methods that require more complex feedback, we conduct an evaluation of CRAYON-ATTENTION and CRAYON-PRUNING. We perform this

evaluation on three datasets with automated feedback provision before collecting feedback from a future large-scale human evaluation. In this section, we first describe our experiment setup and then present and discuss the results.

4.1. Datasets

Biased CelebA. Inspired by previous work (Krishnakumar et al., 2021), we intentionally introduce a spurious correlation between the attributes of "hair color" and "smiling" by subsampling the CelebA face image dataset (Liu et al., 2015a). The training set consists of 20,200 data instances:

- 10,000 with *black hair* and *smiling* attributes;
- 10,000 with *blond hair* and *not smiling* attributes;
- 100 with *black hair* and *not smiling* attributes; and
- 100 with *blond hair* and *smiling* attributes.

The test set contains a total of 8,000 data instances, with 2,000 instances for each group. We train classifiers to predict whether a face in an input image is smiling, which would incorrectly associate the prediction with hair color.

Waterbirds. The Waterbirds (Sagawa et al., 2019) dataset is a compilation of bird photographs (Wah et al., 2011) that are combined with backgrounds (Zhou et al., 2017) so that waterbirds and landbirds appear more frequently in water (e.g., ocean, lake) and land (e.g., forest) backgrounds, respectively. The training set consists of:

- 1,057 *waterbirds* on *water* backgrounds,
- 3,498 *landbirds* on *land* backgrounds,
- 56 *waterbirds* on *land* backgrounds, and
- 184 *landbirds* on *water* backgrounds.

Models trained on this dataset would classify waterbirds and landbirds based on the backgrounds rather than the bird bodies. The test set consists of 1,284 waterbirds and 4,510 landbirds; half of the waterbird images and half of the landbird images have water backgrounds, while the other half have land backgrounds.

DecoyMNIST. DecoyMNIST (Ross et al., 2017) is a synthetic variation of MNIST (LeCun et al., 2010). The training set of DecoyMNIST contains 60,000 images of digits (0 to 9). A 4px×4px square of patch with a shade of gray is overlaid at random one of the image's four corners (Fig. 3.1). The shade of gray is a function of the digit y, specifically $255 - 25y$ (Ross et al., 2017). In other words, every image of the digit 0 has the lightest square (placed at a random corner), and every image of the digit 9 has the darkest square. On the other hand, each image in test dataset, which contains 10,000 images, has a square with a random shade of gray placed in a random corner. Digit classifiers trained on DecoyMNIST are likely to focus on the shades of gray of the squares in the corners, which are spurious correlations present only in the training data, and poorly perform on the test set.

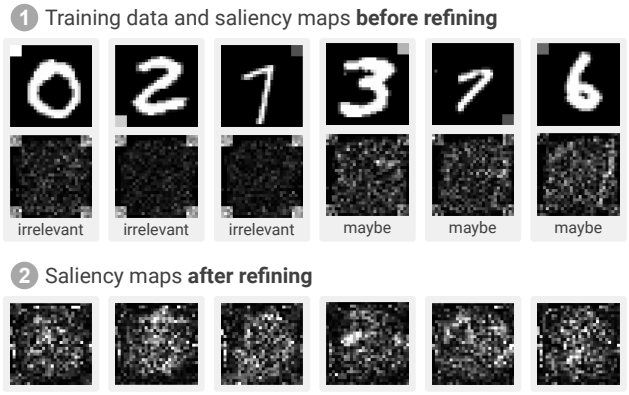

Figure 3. (1) Training data of the DecoyMNIST dataset contains a 4px×4px square patch in one of the four corners, with a shade of gray determined by the digit label. A digit classifier trained on the dataset often uses the square colors, which are dependent on the digit labels only in the training data. (2) After refining the model with CRAYON-ATTENTION, the model does not attend to the squares but to the central regions where digits are located.

4.2. Compared Methods and How They Are Trained

Original Models. We refer to the models that have been trained with data containing spurious correlations as the *original* models, meaning no mitigation has been applied.

- For the Biased CelebA dataset, the original model is ResNet50 (He et al., 2016) pretrained with ImageNet (Deng et al., 2009), using a learning rate of 0.0001 and a batch size of 64 for 5 epochs.
- For the Waterbirds dataset, the original model is ResNet50 pretrained with ImageNet, using a learning rate of 0.001 and a batch size of 128 for 50 epochs.
- For the DecoyMNIST dataset, the original model is a multilayer perceptron (MLP)consisting of two layers, each with an output dimension of 50 and 30, trained using a learning rate of 0.001 and a batch size of 256 for 30 epochs.

For consistent results, we train five variations of each original model, each using a different random seed. Our experiment results are averaged across these five variations. We use the Adam optimizer (Kingma & Ba, 2014) with a weight decay of 0.0001 for all training.

Compared Methods. Our goal is to compare how well our two methods, CRAYON-ATTENTION and CRAYON-PRUNING, would mitigate the spurious correlations in the original models, against five other methods. These include three state-of-the-art techniques that mitigate spurious correlations through the use of feedback on model attention[1]:

[1]We tried RES (Gao et al., 2022a) on our datasets and determined that it was computationally prohibitive. The algorithm did not finish one iteration even after 3 hours on an NVIDIA A6000 GPU; 2,205 iterations are needed for the Biased CelebA dataset.

- *RRR* (Ross et al., 2017) collects ground truth maps that annotate irrelevant pixels in the images and guides the model not to attend to the irrelevant regions.
- *GRADIA* (Gao et al., 2022b) identifies the images that the baseline model generates irrelevant saliency maps or incorrectly predicts, collects ground truth maps of relevant pixels in the images, and aligns the model's attention with the collected maps.
- *Energy loss* (Rao et al., 2023) collects bounding boxes that cover the relevant regions of each image and guides the model to keep its attention within the boxes.

And two basic baseline methods:

- *Train More* trains the model for more epochs without any feedback, to ensure that the result of mitigation is not simply due to further training.
- *Random Pruning* prunes random neurons in the penultimate layer and fine-tunes the last fully connected layer; the number of the pruned neurons is the same as in CRAYON-PRUNING.

For all methods, except CRAYON-PRUNING, we fine-tune the original models as follows[2]:

- Biased CelebA: we use a learning rate of 0.0001 and a batch size of 64 for 7 epochs; for CRAYON-ATTENTION, we set $\alpha = 10000$, $\beta = 100$.
- Waterbirds: we use a learning rate of 0.0001 and a batch size of 128 for 4 epochs; for CRAYON-ATTENTION, we set $\alpha = 2000$, $\beta = 20$.
- DecoyMNIST: we use a learning rate of 0.001, which exponentially decreases at each epoch by 0.975, and a batch size of 256 for 200 epochs; for CRAYON-ATTENTION, we set $\alpha = 0$, $\beta = 30$.

CRAYON-PRUNING prunes the irrelevant neurons in the penultimate layer of the original models and trains the last fully connected layer of the original models with the learning rate of 0.0001 for 10 epochs for both Biased CelebA and Waterbirds datasets. All methods use Adam optimizer (Kingma & Ba, 2014) with a weight decay of 0.0001.

4.3. Ground Truth & Feedback Generation

For a fair comparison, we generate the ground truth and feedback using all training data of each dataset. For the Biased CelebA dataset, in which the celebrities' eyes and mouths are at the same positions across all images (Liu et al., 2015b), the ground truth locations of the mouths can be automatically determined. As a result, we can label the relevance of Grad-CAM of an image for the CRAYON-ATTENTION as *yes* if the map puts significant attention on the mouth, *no* if it highlights other regions, and *maybe* if the attention is ambiguous (Fig. 1, left). To be specific, the relevance of Grad-CAM of an image is labeled as *yes* if the highest attention is on the mouth and all areas not covering the eyes and mouth receive less than half of the attention given to the mouth, *no* if the mouth gets less than 0.7 of the maximum attention and the areas not covering the eyes and mouth receive more than half of the maximum attention, and *maybe* otherwise. To identify relevant neurons for CRAYON-PRUNING, we forward all training data through the model and collect 20 image patches that summarize the visual concepts responsible for the activation of each neuron in the penultimate layer (Fig. 2). We determine the feedback on a neuron's relevance to be *yes* if all its patches are containing mouths and *no* otherwise. We generate the ground truth attention maps and bounding boxes required by the competitors to cover mouths.

For the Waterbirds dataset, which provides segmentation maps of bird bodies for each image, the relevance of Grad-CAM of an image for the CRAYON-ATTENTION is labeled as *yes* if more than 60% of the attention of the Grad-CAM of is on the bird body, *no* if less than 40% of the attention of the Grad-CAM of is on the bird body, and *maybe* otherwise. For CRAYON-PRUNING, we collect 10 image patches for each neuron in the penultimate layer by forwarding all training data through a model and label the neuron's relevance as *yes* if all or all but one of the 10 patches intersect with bird bodies and *no* if more than one of the patches do not cover bird bodies. We use the segmentation maps as the ground truth attention maps and generate the bounding boxes by drawing boxes around the segmentation maps.

For the DecoyMNIST dataset, it is challenging to automatically determine the precise pixels that are responsible for its digit prediction (see examples in Fig. 3). Therefore, we do not experiment with GRADIA because it requires ground truth pixels for model attention. Pruning methods are also inapplicable for this dataset since the models are MLPs and not CNNs. We generate yes-no feedback on the relevance of the *Input Gradient* saliency maps (Baehrens et al., 2010) by assigning *no* to the data that puts more than 20% of the total attention on the four corners and *maybe* otherwise (Fig. 3). For RRR, we use the map that annotates 4px×4px square of the four corners as the areas not to be attended for every image. For the bounding box, we draw a square on the center that does not cover any of the squares on the corners.

4.4. Results

4.4.1. Results for Biased CelebA & Waterbirds

Table 1 compares the performance of our methods and that of the competitors on the Biased CelebA and Waterbirds datasets. Following the convention of the literature on spurious correlation research (Sagawa et al., 2019), we use *mean accuracy* (MA) and *worst group accuracy* (WGA) as evalua-

[2]We report the number of epochs the fine-tuning took to reach the plateau of best performances

Table 1. Our methods based on simple yes-no feedback successfully mitigate spurious correlations in the models for the Biased CelebA and Waterbirds datasets, achieving comparable or even better performance than the existing SOTA approaches (RRR, GradIA, Energy loss) that require more complex feedback. Our methods achieve either the **best** or **second-best** accuracies. #FB stands for the number of feedback used; MA stands for *mean accuracy*; and WGA stands for *worst group accuracy*.

Method	Feedback	Biased CelebA			Waterbirds		
		#FB	MA	WGA	#FB	MA	WGA
Original	-	-	71.57	30.42	-	67.84	25.31
CRAYON-ATTENTION	Yes-No	20,200	**83.31**	**64.25**	4,795	71.50	37.76
CRAYON-PRUNING	Yes-No	2,048	**84.25**	62.81	2,048	**76.71**	**55.74**
RRR (Ross et al., 2017)	Map	20,200	81.08	55.32	4,795	75.35	39.10
GradIA (Gao et al., 2022b)	Yes-No, Map	20,200	75.51	35.16	4,795	74.31	36.01
Energy loss (Rao et al., 2023)	Bounding box	20,200	82.04	**63.52**	4,795	**79.86**	**54.98**
Train More	None	0	70.09	18.69	0	67.72	18.69
Random Pruning	None	0	71.42	29.86	0	66.49	15.89

tion metrics for these datasets. Specifically, we first evaluate the model accuracy for each *attribute group* introduced in Sec. 4.1. For example, the groups for the Biased CelebA dataset are:

- *black hair + smiling*
- *blond hair + not smiling*
- *black hair + not smiling*
- *blond hair + smiling*

We then calculate the mean and minimum of the accuracy values across the groups and denote them as *mean accuracy* (MA) and *worst group accuracy* (WGA), respectively. As we experiment with 5 original models with different random seeds for each dataset, we report the average of the MA and WGA values for the experiment with each random seed.

Overall, our methods based on yes-no feedback achieve comparable or even better performance than other competitors that require more complex feedback. For the Biased CelebA dataset, comparing with the unrefined original models (first row) demonstrates that both CRAYON-ATTENTION and CRAYON-PRUNING effectively mitigate spurious correlations, providing a significant boost to the *mean accuracy* (MA) by 12.68 percentage points (pp) (84.25 for CRAYON-PRUNING vs 71.57 for original) and the *worst group accuracy* (WGA) by 33.83pp (64.25 for CRAYON-ATTENTION vs 30.42 for original). The CRAYON-ATTENTION outperforms the competitors, which exploit ground truth maps and bounding boxes with richer information, in terms of both MA and WGA. We attribute the superiority of our method to the limitations of the ground truth maps and boxes. To be specific, the maps and boxes are represented as the binary values of 0 and 1, while the model-generated saliency maps have continuous real numbers. This inconsistency degrades the performance of the model attention guidance (Gao et al., 2022a). CRAYON-ATTENTION resolves the challenge by using the saliency maps of the unrefined model instead of binary ground truth. It is also notable that the CRAYON-PRUNING achieves the highest MA while using only one-tenth of the feedback compared to other methods (2,048 for CRAYON-PRUNING vs 20,200 for the others). This highlights the superiority of CRAYON-PRUNING in terms of both scalability and performance.

Our methods demonstrate the effectiveness in mitigating spurious correlations also for the Waterbirds dataset, enhancing the baseline models' MA from 67.84pp to 76.71pp and WGA from 25.31pp to 55.74pp. Especially, CRAYON-PRUNING shows significant superiority achieving the second best MA and the best WGA values among all the compared methods even though the number of feedback it takes is less than half of its competitors. The performance of the CRAYON-ATTENTION, which is greater than that of original models but lower than other competitors, indicates that there are rooms to be more improved.

We also conduct a qualitative evaluation to assess the effectiveness of CRAYON-ATTENTION, as shown in Fig. 1. For a model that irrelevantly attends to the *forehead* of an image from the Biased CelebA dataset, CRAYON-ATTENTION fixes its attention to the *mouth*. Similarly, CRAYON-ATTENTION rectifies the attention of a bird classifier that initially focuses on the background of a Waterbirds image to the *bird's body*.

4.4.2. RESULTS FOR DECOYMNIST

For the DecoyMNIST dataset with 100 attribute groups, we adopt the average accuracy as the evaluation metric, which is the proportion of correctly predicted data among all test data. This follows the convention of the literature (Ross et al., 2017; Friedrich et al., 2023). As previously mentioned in Sec. 4.3, we do not experiment with GRADIA and

Table 2. Accuracy for DecoyMNIST dataset.

Method	Feedback	#FB	Acc
Original	-	-	55.72
CRAYON-ATTENTION	Yes-No	60k	85.33
RRR	Map	60k	96.18
Energy loss	Bounding Box	60k	96.13
Train More	None	0	56.72

CRAYON-PRUNING because we do not have ground truth pixels for model attention and the models are MLPs not CNNs. Table 2 shows that CRAYON-ATTENTION mitigates the spurious correlations in the original models, significantly boosting the accuracy from 55.72pp to 85.33pp. However, it is not as high as the performance of RRR and Energy loss. Considering the simplicity of yes-no feedback and the complexity of attention maps needed for RRR and bounding boxes for Energy loss, we believe the performance differences between CRAYON-ATTENTION and those methods are a reasonable trade-off between the cost of feedback provision and performance.

These results demonstrate the strengths of our methods in terms of low cost of feedback collection, superior performance, scalability, and wide applicability. Specifically, CRAYON-PRUNING shows a remarkable ability to mitigate spurious correlations in CNN classifiers for the Biased CelebA and Waterbirds datasets with only a few numbers of yes-no feedback. Moreover, as it fine-tunes only the last fully connected layer, its training cost is much less than the competitors. CRAYON-ATTENTION proves its versatility by successfully mitigating spurious correlations in both MLP and CNN models.

5. Conclusion and Ongoing Work

We propose two methods, CRAYON-ATTENTION and CRAYON-PRUNING, which mitigate spurious correlations in image classifiers using simple yes-no feedback. CRAYON-ATTENTION collects yes-no feedback on the relevance of saliency maps and refines models to attend to the relevant regions of images, while CRAYON-PRUNING identifies and prunes irrelevant neurons in the penultimate layer of the models based on the yes-no feedback on the relevance of the neuron activation. Our experiments demonstrate that our methods effectively mitigate spurious correlations, achieving comparable or even better performance than other competitors using much more complex feedback.

Collecting human feedback. Based on the promising experiment results thus far, which highlight the potential of CRAYON-ATTENTION and CRAYON-PRUNING in mitigating spurious correlations in deep learning models, we plan to conduct a large-scale human evaluation to collect yes-no feedback through a crowd-sourcing platform such as Prolific (Lee et al., 2022). Through such a study, we will empirically verify our methods' ease of use and time saving, comparing with conventional fine-grained approaches like asking participants to draw pixel-wise maps or bounding boxes.

Integration of multiple mitigation methods. CRAYON-ATTENTION and CRAYON-PRUNING are currently designed to work as two independent methods, each with its own strengths. We are examining whether considering both methods in tandem may lead to superior outcomes compared to employing just a single method. Moreover, considering the trade-off between the performance and the cost of feedback collection as discussed in Sec. 4.4, we plan to experiment with potential the benefits of supplementing yes-no feedback with a very small amount of the more complex feedback, such as a few ground truth maps or bounding boxes so as to keep our overall approach simple and practical.

Acknowledgements

This research was supported in part by Cisco.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Asgari, S., Khani, A., Khani, F., Gholami, A., Tran, L., Mahdavi-Amiri, A., and Hamarneh, G. Masktune: Mitigating spurious correlations by forcing to explore. In *Advances in Neural Information Processing Systems*, 2022.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Chiu, M.-C., Chen, P.-Y., and Ma, X. Better may not be fairer: Can data augmentation mitigate subgroup degradation? *arXiv preprint arXiv:2212.08649*, 2022.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Friedrich, F., Steinmann, D., and Kersting, K. One explanation does not fit xil. *arXiv preprint arXiv:2304.07136*, 2023.

Gao, Y., Sun, T. S., Bai, G., Gu, S., Hong, S. R., and Liang, Z. Res: A robust framework for guiding visual explanation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 432–442, 2022a.

Gao, Y., Sun, T. S., Zhao, L., and Hong, S. R. Aligning eyes between humans and deep neural network through interactive attention alignment. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28, 2022b.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Hagos, M. T., Curran, K. M., and Mac Namee, B. Identifying spurious correlations and correcting them with an explanation-based learning. *arXiv preprint arXiv:2211.08285*, 2022.

Han, Z., Liang, Z., Yang, F., Liu, L., Li, L., Bian, Y., Zhao, P., Wu, B., Zhang, C., and Yao, J. Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. *arXiv preprint arXiv:2209.08928*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hohman, F., Park, H., Robinson, C., and Chau, D. H. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2020. URL https://fredhohman.com/summit/.

Kim, B., Kim, H., Kim, K., Kim, S., and Kim, J. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9012–9020, 2019.

Kim, E., Lee, J., and Choo, J. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14992–15001, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Krishnakumar, A., Prabhu, V., Sudhakar, S., and Hoffman, J. Udis: Unsupervised discovery of bias in deep visual recognition models. In *British Machine Vision Conference (BMVC)*, 2021.

LeCun, Y., Cortes, C., and Burges, C. J. The mnist database of handwritten digits. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2010. URL http://yann.lecun.com/exdb/mnist/.

Lee, J., Kim, E., Lee, J., Lee, J., and Choo, J. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34:25123–25133, 2021.

Lee, S., Afroz, S., Park, H., Wang, Z. J., Shaikh, O., Sehqal, V., Peshin, A., and Chau, D. H. Explaining website reliability by visualizing hyperlink connectivity. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pp. 26–30. IEEE, 2022.

Li, Y. and Vasconcelos, N. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9572–9581, 2019.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Liu, S., Zhang, X., Sekhar, N., Wu, Y., Singhal, P., and Fernandez-Granda, C. Avoiding spurious correlations via logit correction. *arXiv preprint arXiv:2212.01433*, 2022.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015a.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015b.

Mahendran, A. and Vedaldi, A. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120:233–255, 2016.

Minderer, M., Bachem, O., Houlsby, N., and Tschannen, M. Automatic shortcut removal for self-supervised representation learning. In *International Conference on Machine Learning*, pp. 6927–6937. PMLR, 2020.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33: 20673–20684, 2020.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.

Plumb, G., Ribeiro, M. T., and Talwalkar, A. Finding and fixing spurious patterns with explanations. *arXiv preprint arXiv:2106.02112*, 2021.

Rao, S., Böhle, M., Parchami-Araghi, A., and Schiele, B. Using explanations to guide models. *arXiv preprint arXiv:2303.11932*, 2023.

Rieger, L., Singh, C., Murdoch, W., and Yu, B. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pp. 8116–8126. PMLR, 2020.

Rodolfa, K. T., Lamba, H., and Ghani, R. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, 3(10):896–904, 2021.

Ross, A. S., Hughes, M. C., and Doshi-Velez, F. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.-G., Mahlein, A.-K., and Kersting, K. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Singla, S. and Feizi, S. Salient imagenet: How to discover spurious features in deep learning? *arXiv preprint arXiv:2110.04301*, 2021.

Singla, S., Nushi, B., Shah, S., Kamar, E., and Horvitz, E. Understanding failures of deep networks via robust feature extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12853–12862, 2021.

Stammer, W., Schramowski, P., and Kersting, K. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3619–3629, 2021.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. Caltech-UCSD Birds-200-2011 (CUB-200-2011). Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pp. 25407–25437. PMLR, 2022.

Zhang, M., Sohoni, N. S., Zhang, H. R., Finn, C., and Ré, C. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022.

Zhang, Y., Gong, M., Liu, T., Niu, G., Tian, X., Han, B., Schölkopf, B., and Zhang, K. Causaladv: Adversarial robustness through the lens of causality. *arXiv preprint arXiv:2106.06196*, 2021.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.