Learning from Failure:

Training Debiased Classficier from Biased Classifier

Jihyeon Lee, Oct *20th, 2020*

Motivation

Learning debias without supervision

- Neural networks often learn to make predictions using the unintended decision rule
- Recent approaches focus on how to utilize various types of human supervision effectively
- Another line of research focuses on developing algorithms tailored to a domain-specific type of bias in the target dataset
- An approach to train a debiased classifier without relying on such expensive supervision is warranted

$$\min_f \left\{ \mathcal{L}(f) + \lambda \max_g \Big(\mathrm{HSIC}_1(f,g) - \lambda_g \mathcal{L}(g) \Big) \right\}$$

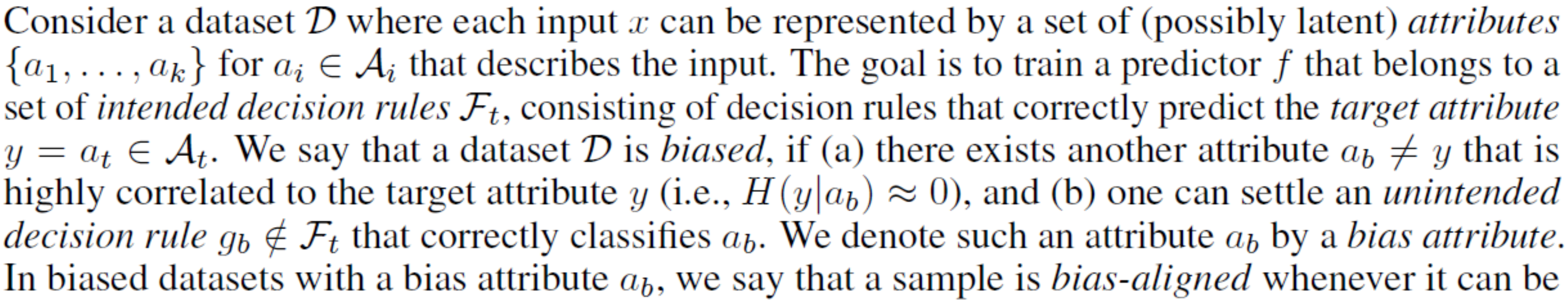
Contribution

Learning from Failure(LfF) – failure based debiasing scheme

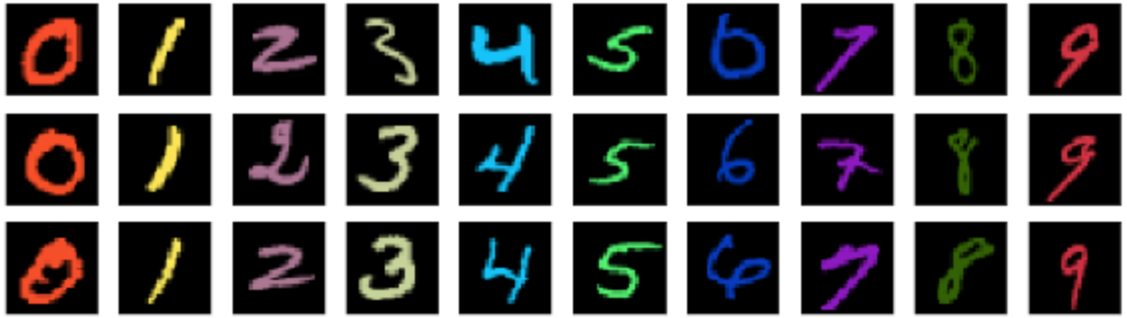
1. Propose debiasing network by simultaneously train two neural net, one to be biased and the other to be debiased
2. Classifier learns to fit samples aligned with the bias during the early stage of training and learns samples conflicting with the bias later
3. LfF Does not require expensive supervision on the bias or bias-tailored training technique
4. Show the effectiveness of LfF on various biased datasets
5. Constructed baised action recognition dataset (BAR)

Biased Dataset

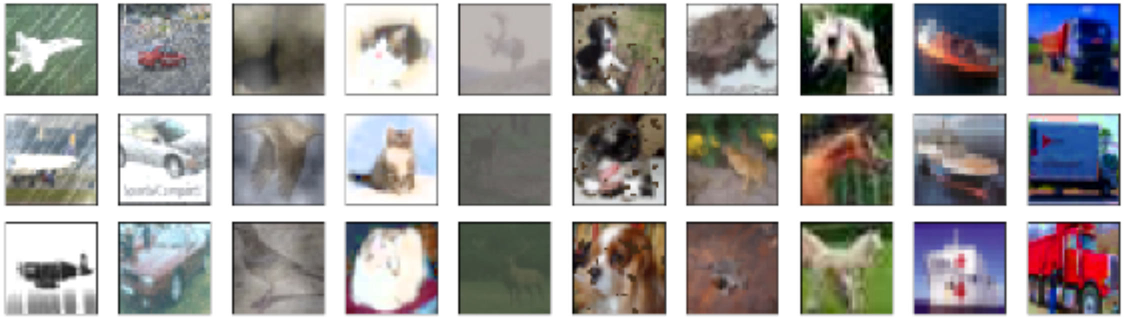
Definition of bias



Consider a dataset $\mathcal{D}$ where each input x can be represented by a set of (possibly latent) *attributes* $\{a_1, \dots, a_k\}$ for $a_i \in \mathcal{A}_i$ that describes the input. The goal is to train a predictor f that belongs to a set of *intended decision rules* $\mathcal{F}_t$, consisting of decision rules that correctly predict the *target attribute* $y = a_t \in \mathcal{A}_t$. We say that a dataset $\mathcal{D}$ is *biased*, if (a) there exists another attribute $a_b \neq y$ that is highly correlated to the target attribute y (i.e., $H(y|a_b) \approx 0$), and (b) one can settle an *unintended decision rule* $g_b \notin \mathcal{F}_t$ that correctly classifies a_b. We denote such an attribute a_b by a *bias attribute*. In biased datasets with a bias attribute a_b, we say that a sample is *bias-aligned* whenever it can be



(a) Colored MNIST

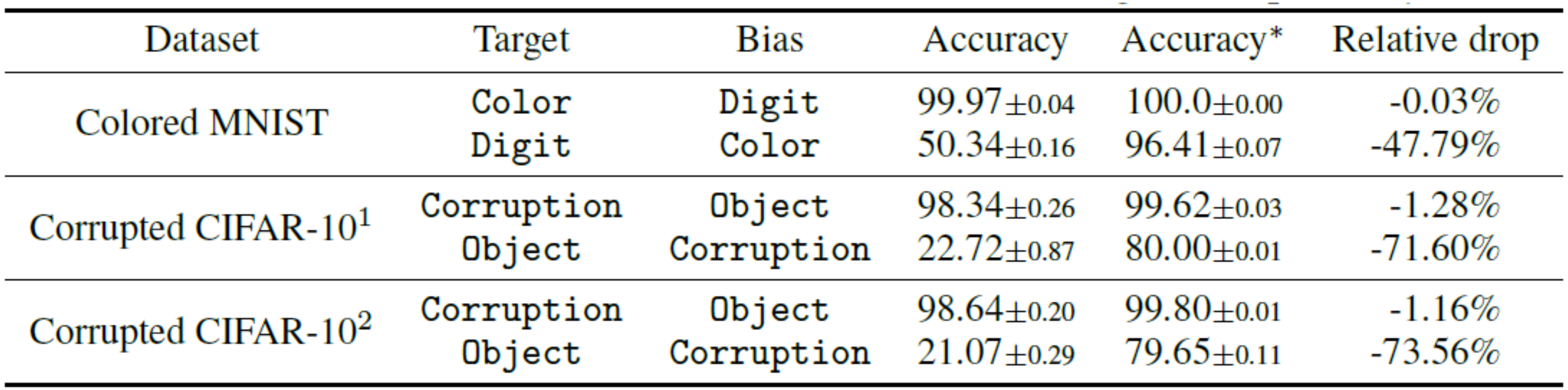


(b) Corrupted CIFAR-10[1]

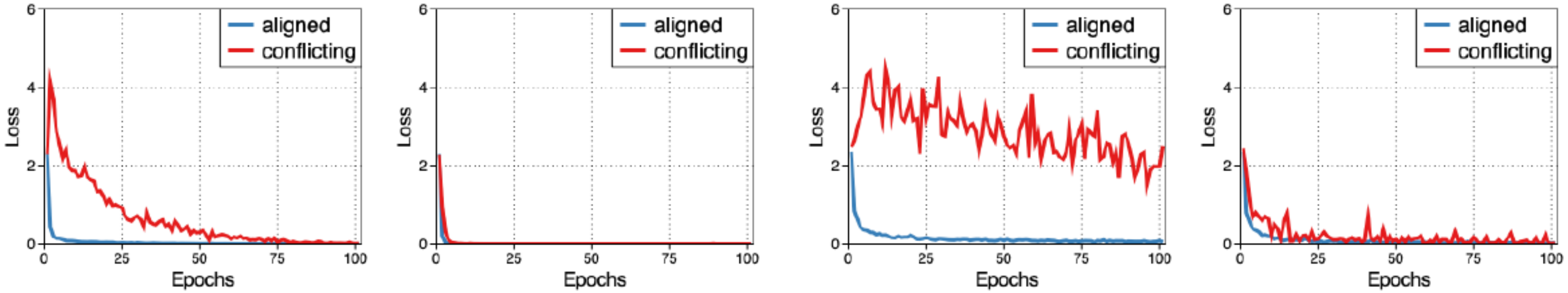
Biased Dataset

Two types of bias: malignant and benign

- Biased dataset does not necessarily lead the model to learn the unintended decision rule
- The bias negatively affects the model only when the bias attribute is “easier” to learn than the target attribute



Dataset	Target	Bias	Accuracy	Accuracy*	Relative drop
Colored MNIST	Color	Digit	99.97±0.04	100.0±0.00	-0.03%
	Digit	Color	50.34±0.16	96.41±0.07	-47.79%
Corrupted CIFAR-10[1]	Corruption	Object	98.34±0.26	99.62±0.03	-1.28%
	Object	Corruption	22.72±0.87	80.00±0.01	-71.60%
Corrupted CIFAR-10[2]	Corruption	Object	98.64±0.20	99.80±0.01	-1.16%
	Object	Corruption	21.07±0.29	79.65±0.11	-73.56%



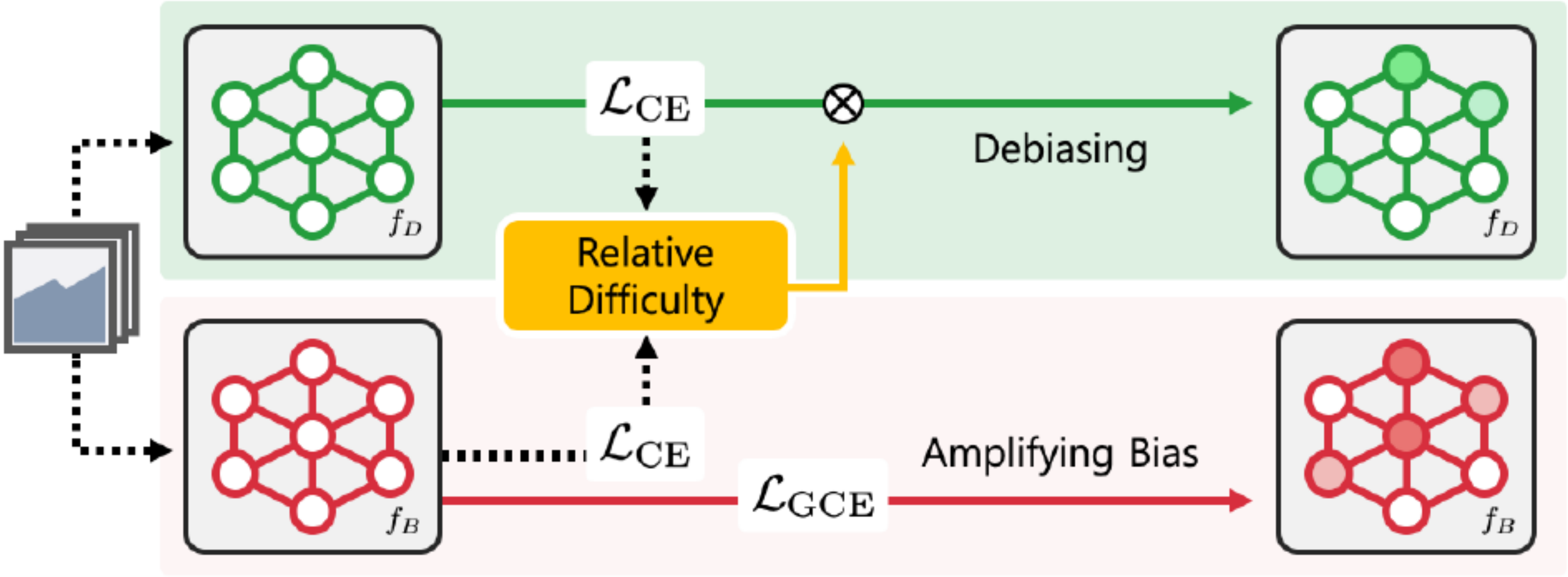
(a) Colored MNIST, (Digit, Color)

(b) Corrupted CIFAR-10[1], (Object, Corruption)

Model

Model Structure

Debiased Model


$\mathcal{L}_{\mathrm{CE}}$
Debiasing
f_D
f_D
Relative Difficulty
$\mathcal{L}_{\mathrm{CE}}$
Amplifying Bias
$\mathcal{L}_{\mathrm{GCE}}$
f_B
f_B

Biased Model

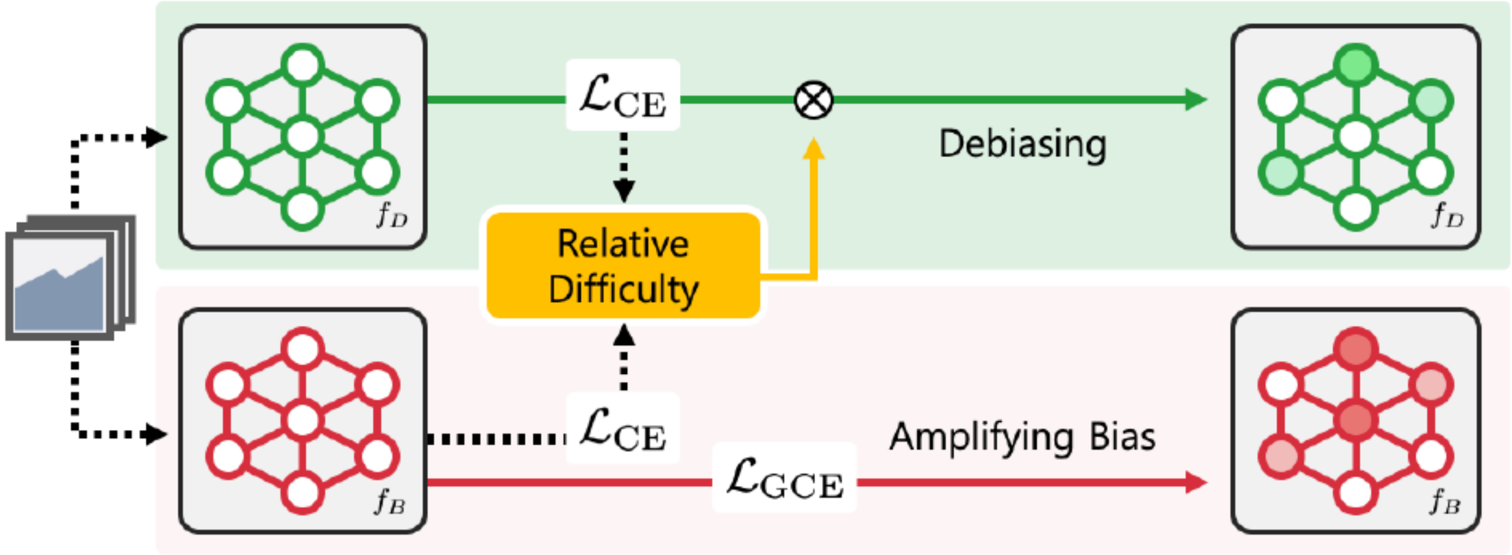
Model

Training biased model – amplify the bias

- GCE loss up-weights the gradient of the CE loss for the samples with a high probability p of predicting the correct target
- emphasizing the ‘easier’ samples with the strong agreement between softmax output and the target

$$\mathrm{GCE}(p(x;\theta), y) = \frac{1 - p_y(x;\theta)^q}{q} \qquad \lim_{q\to 0} \frac{1-p^q}{q} = -\log p$$

$$\frac{\partial \mathrm{GCE}(p, y)}{\partial \theta} = p_y^q \frac{\partial \mathrm{CE}(p, y)}{\partial \theta}$$

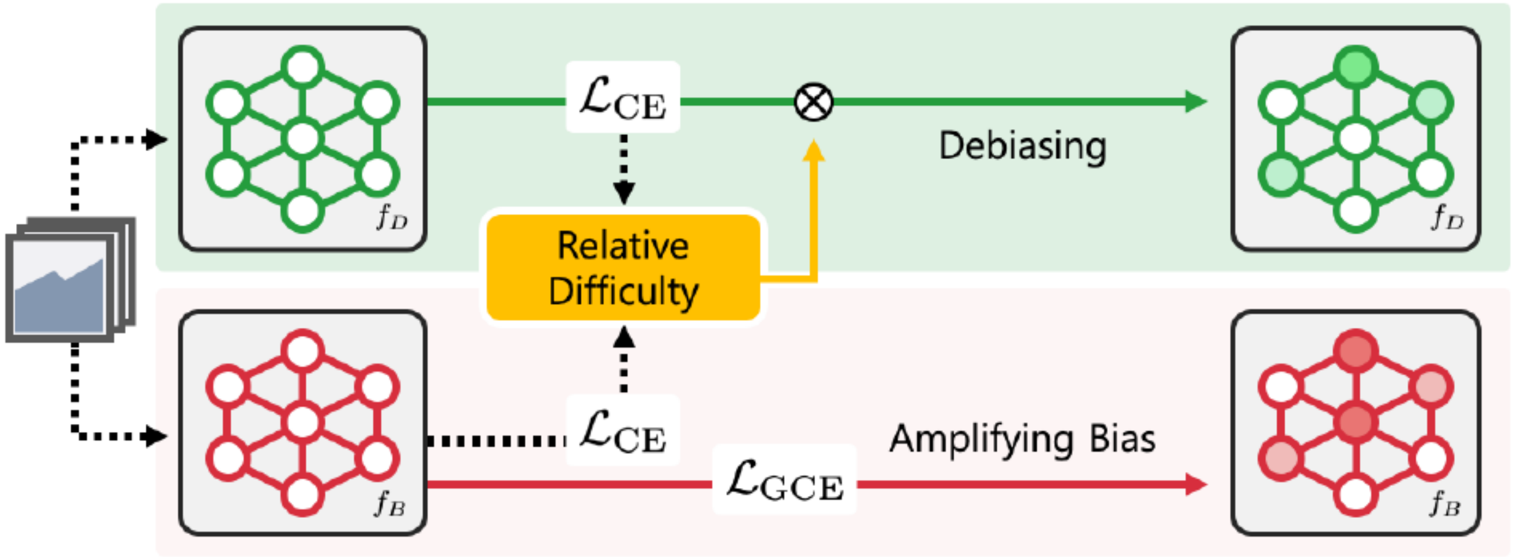

f_D
$\mathcal{L}_{\mathrm{CE}}$
$\otimes$
Debiasing
f_D
Relative Difficulty
f_B
$\mathcal{L}_{\mathrm{CE}}$
$\mathcal{L}_{\mathrm{GCE}}$
Amplifying Bias
f_B

Model

Training de-biased model – relative difficulty score

- Train simultaneously with the samples using the CE loss re-weighted by score
- Score indicates how much each sample is likely to be bias-conflicting
- For bias-alinged samples, biased model f_B tends to have smaller loss compare to debiased model f_D at the early stage, therefore having small weight for training debiased model

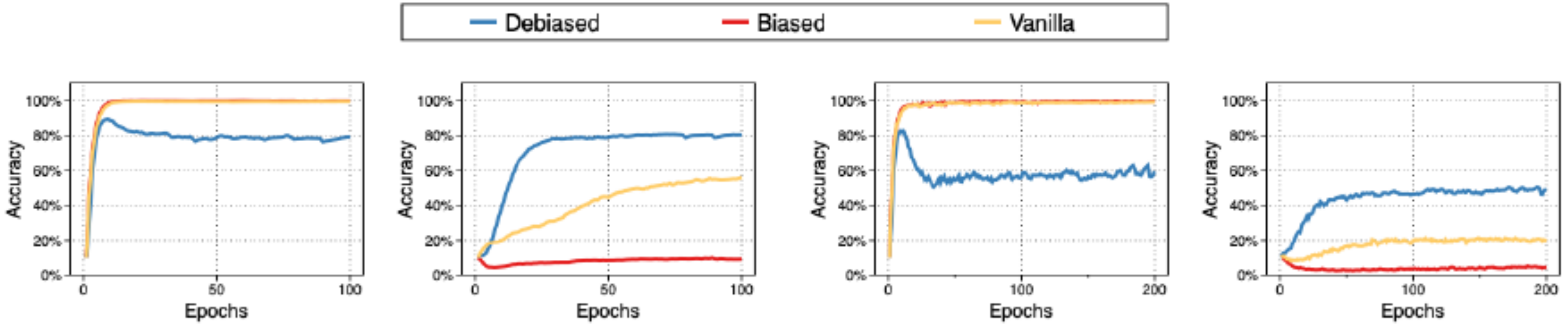
$$\mathcal{W}(x) = \frac{\mathrm{CE}(f_B(x), y)}{\mathrm{CE}(f_B(x), y) + \mathrm{CE}(f_D(x), y)}$$


f_D
$\mathcal{L}_{\mathrm{CE}}$
⊗
Debiasing
f_D
Relative Difficulty
f_B
$\mathcal{L}_{\mathrm{CE}}$
$\mathcal{L}_{\mathrm{GCE}}$
Amplifying Bias
f_B

Experiments

Controlled experiments

Dataset	Ratio (%)	Vanilla	Ours	HEX	REPAIR	Group DRO
		○	○	◑	●	●
Colored MNIST	95.0	77.63±0.44	**85.39**±0.94	70.44±1.41	82.51±0.59	84.50±0.46
	98.0	62.29±1.47	**80.48**±0.45	62.03±0.24	72.86±1.47	76.30±1.53
	99.0	50.34±0.16	**74.01**±2.21	51.99±1.09	67.28±1.69	71.33±1.76
	99.5	35.34±0.13	**63.39**±1.97	41.38±1.31	56.40±3.74	59.67±2.73
Corrupted CIFAR-10^1	95.0	45.24±0.22	**59.95**±0.16	21.74±0.27	48.74±0.71	53.15±0.53
	98.0	30.21±0.82	**49.43**±0.78	17.81±0.29	37.89±0.22	40.19±0.23
	99.0	22.72±0.87	**41.37**±2.34	16.62±0.80	32.42±0.35	32.11±0.83
	99.5	17.93±0.66	**31.66**±1.18	15.39±0.13	26.26±1.06	29.26±0.11
Corrupted CIFAR-10^2	95.0	41.27±0.98	**58.57**±1.18	19.25±0.81	54.05±1.01	57.92±0.31
	98.0	28.29±0.62	**48.75**±1.68	15.55±0.84	44.22±0.84	46.12±1.11
	99.0	20.71±0.29	**41.29**±2.08	14.42±0.51	38.40±0.26	39.57±1.04
	99.5	17.37±0.31	34.11±2.39	13.63±0.42	31.03±0.42	**34.25**±0.74



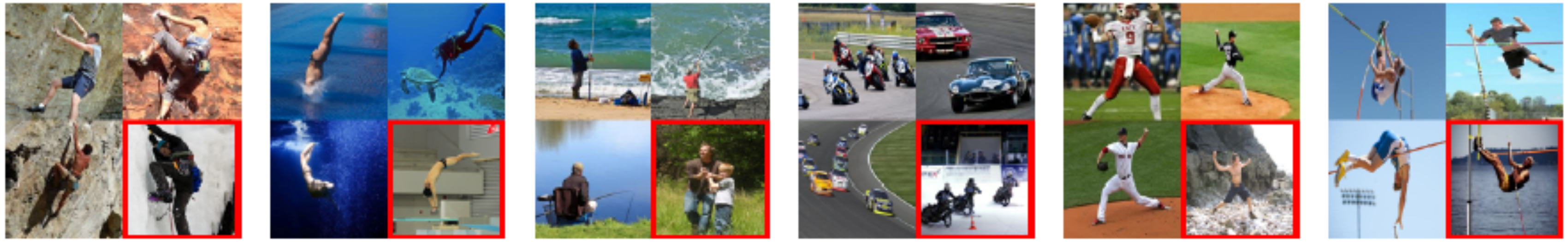
(a) Bias-{aligned, conflicting} Colored MNIST (b) Bias-{aligned, conflicting} Corrupted CIFAR-10^1

Experiments

Real-world experiments

Target attribute	Unbiased			Bias-conflicting		
	Vanilla	Ours	Group DRO	Vanilla	Ours	Group DRO
HairColor	70.25±0.35	84.24±0.37	85.43±0.53	52.52±0.19	81.24±1.38	83.40±0.67
HeavyMakeup	62.00±0.02	66.20±1.21	64.88±0.42	33.75±0.28	45.48±4.33	50.24±0.68

Action	Climbing	Diving	Fishing	Racing	Throwing	Vaulting	Average
Vanilla	59.05±17.48	16.56±1.58	62.69±3.64	77.27±2.62	28.62±2.95	66.92±7.25	51.85±5.92
Ours	**79.36**±4.79	**34.59**±2.26	**75.39**±3.63	**83.08**±1.90	**33.72**±0.68	**71.75**±3.32	**62.98**±2.76



(a) Climbing (b) Diving (c) Fishing (d) Racing (e) Throwing (f) Vaulting

Figure 5: Illustration of BAR images of six typical action-place pairs settled as (Climbing, RockWall), (Diving, Underwater), (Fishing, WaterSurface), (Racing, APavedTrack), (Throwing, PlayingField), and (Vaulting, Sky). The images with red border lines belong to BAR evaluation set, and others belong to BAR training set.

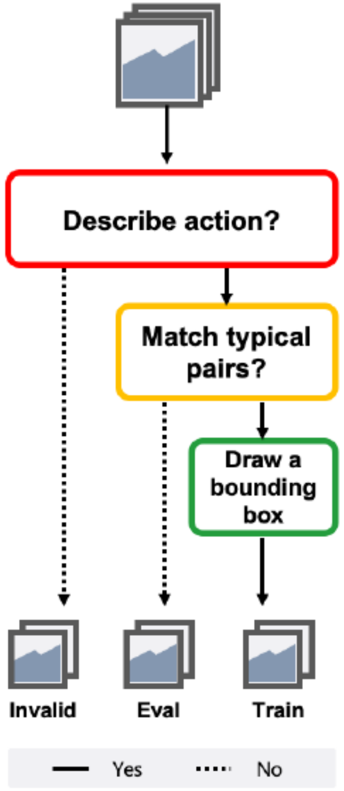


Figure 6: Illustration of BAR reasoning process.

Conclusion

- We consider general properties of bias from observations on training dynamics of bias-aligned and bias-conflicting samples, in a more straightforward approach
- We propose a simple yet widely applicable debiasing scheme free from the choice for form and amount of supervision on the bias
- Besides, our scheme introduces only a single additional hyperparameter(GCE), where one can enjoy simplicity in use
- Using this approach can increase awareness of underexplored biases