A journey through Bias mitigation approaches

A journey through model debiasing: from methods to applications
Tutorial for ICIAP 2025
https://a-journey-through-model-debiasing.github.io/
15/09/2025

Vito Paolo Pastore
Assistant Professor,
MaLGa-DIBRIS, Università degli studi di Genova

vito.paolo.pastore@unige.it

Università di Genova | MaLGa

ISTITUTO ITALIANO DI TECNOLOGIA

Outline

Introduction

Datasets

Supervised approaches

Unsupervised approaches

Recent trends

Conclusions

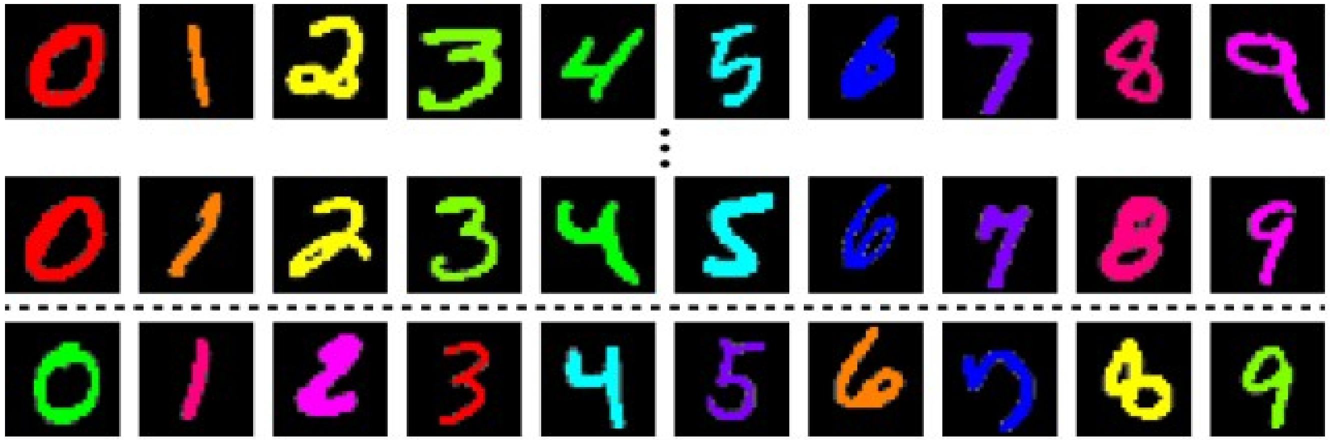
Outline

- *Introduction*
- *Datasets*
- *Supervised approaches*
- *Unsupervised approaches*
- *Recent trends*
- *Conclusions*

Bias in image classification

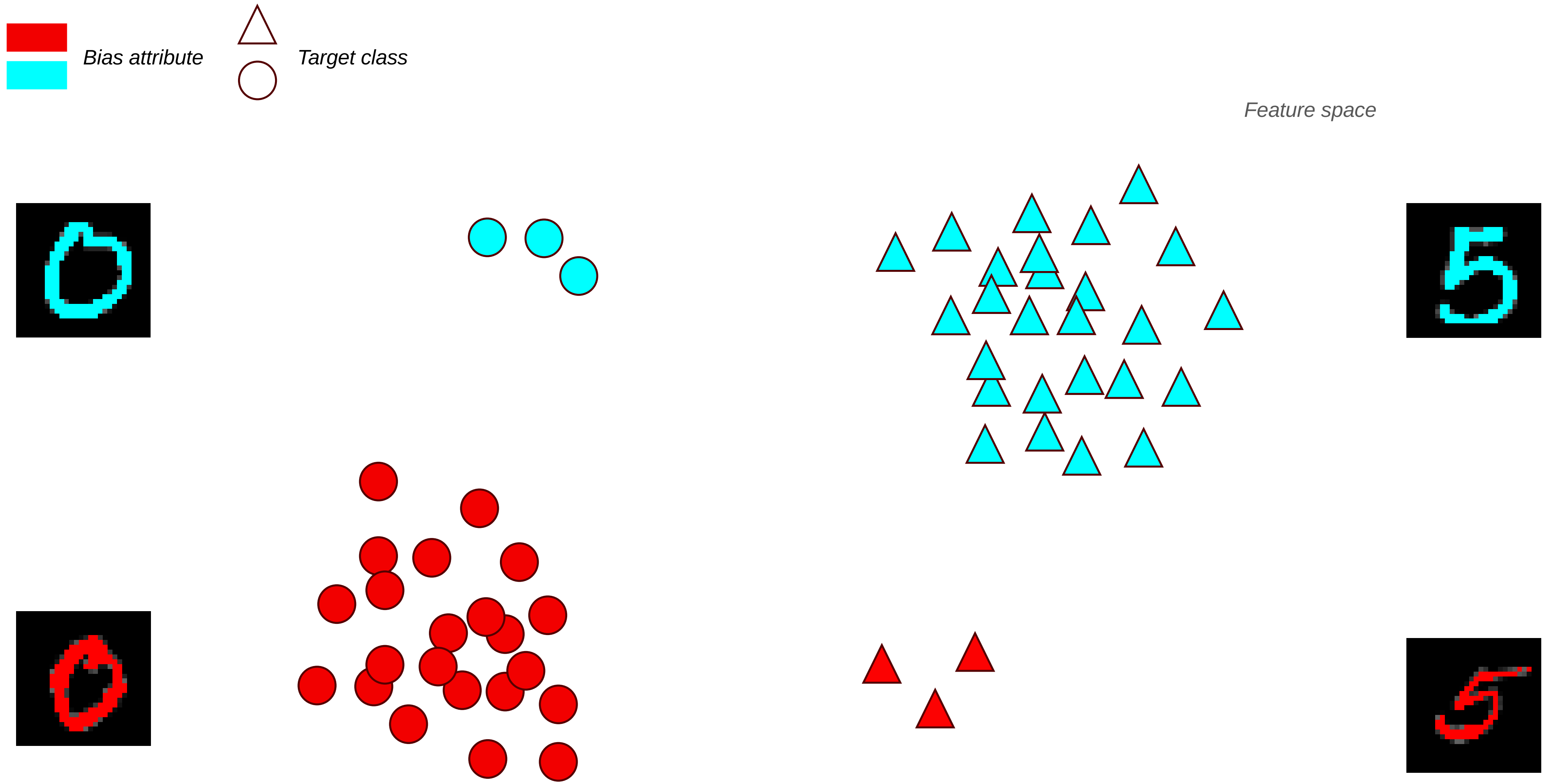
- **Spurious correlations** between class labels and samples;
- **Shortcuts** learned by models to minimize empirical risk;
- Present in most training samples (**bias-aligned**);
- Absent in a small percentage (**bias-conflicting**);
- A model learns these spurious correlations (instead of semantic attributes).

Bias-Aligned

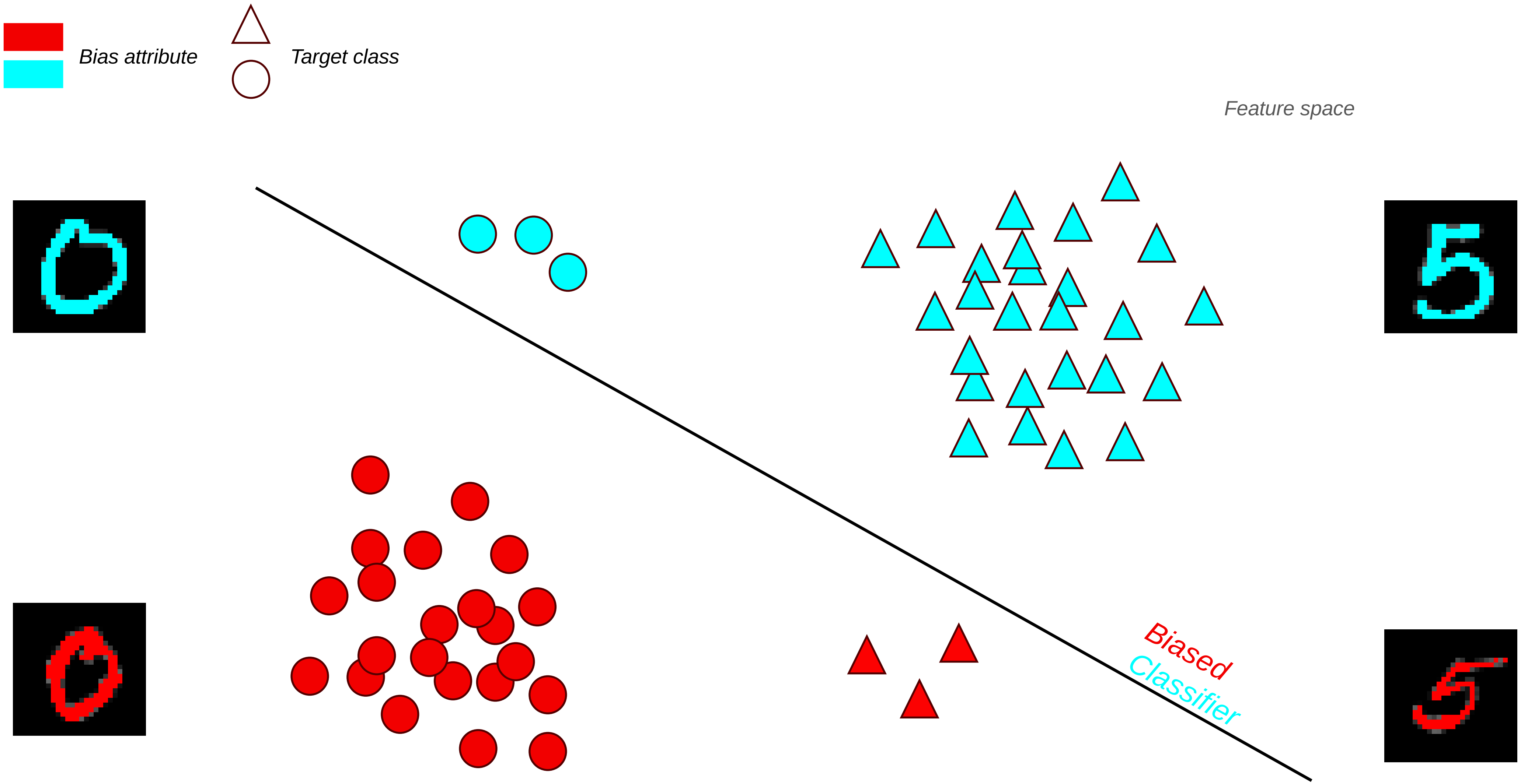


Bias-conflicting/unbiased

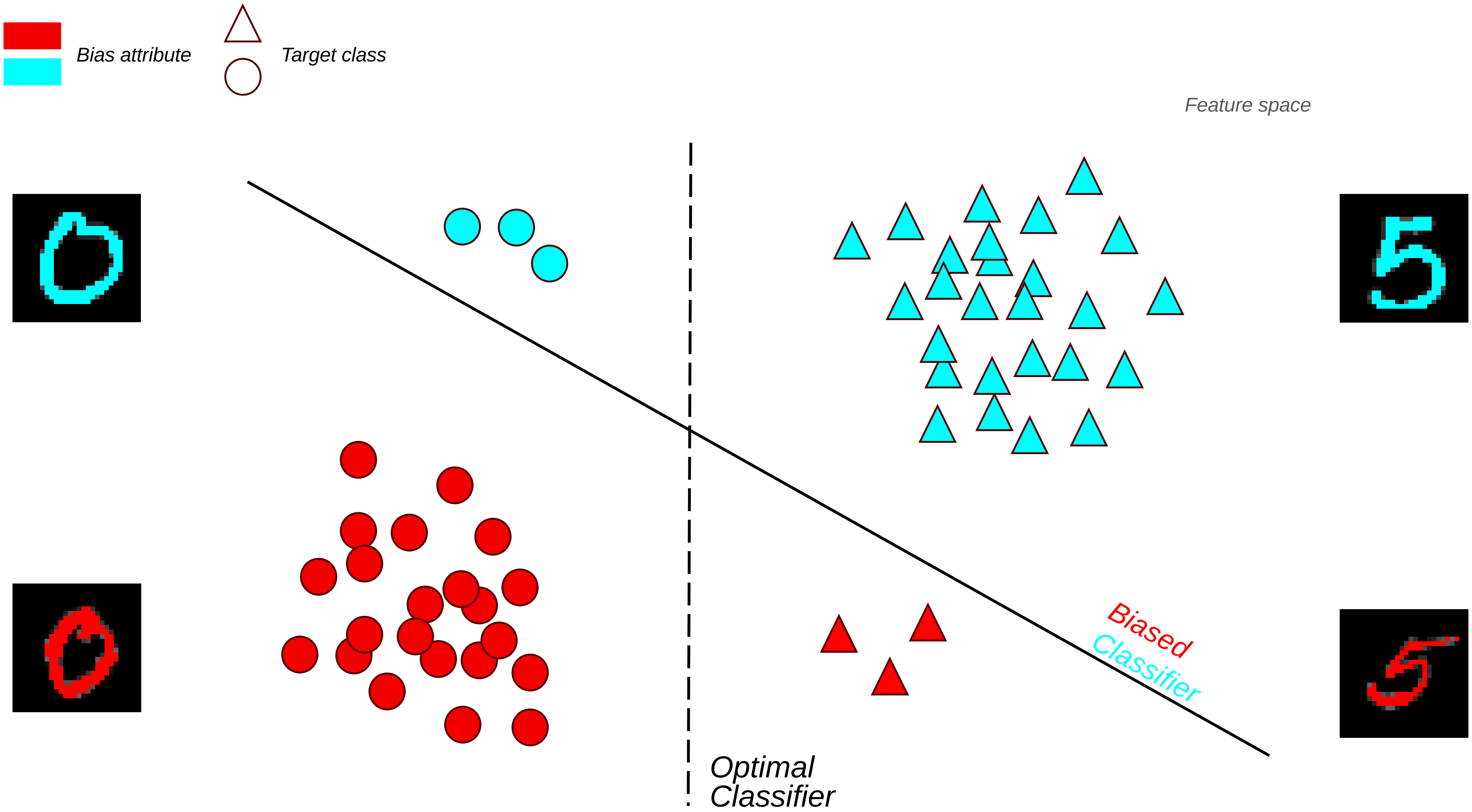
What does it mean to debias a model?



What does it mean to debias a model?



What does it mean to debias a model?



Outline

Introduction

Common Datasets

Supervised approaches

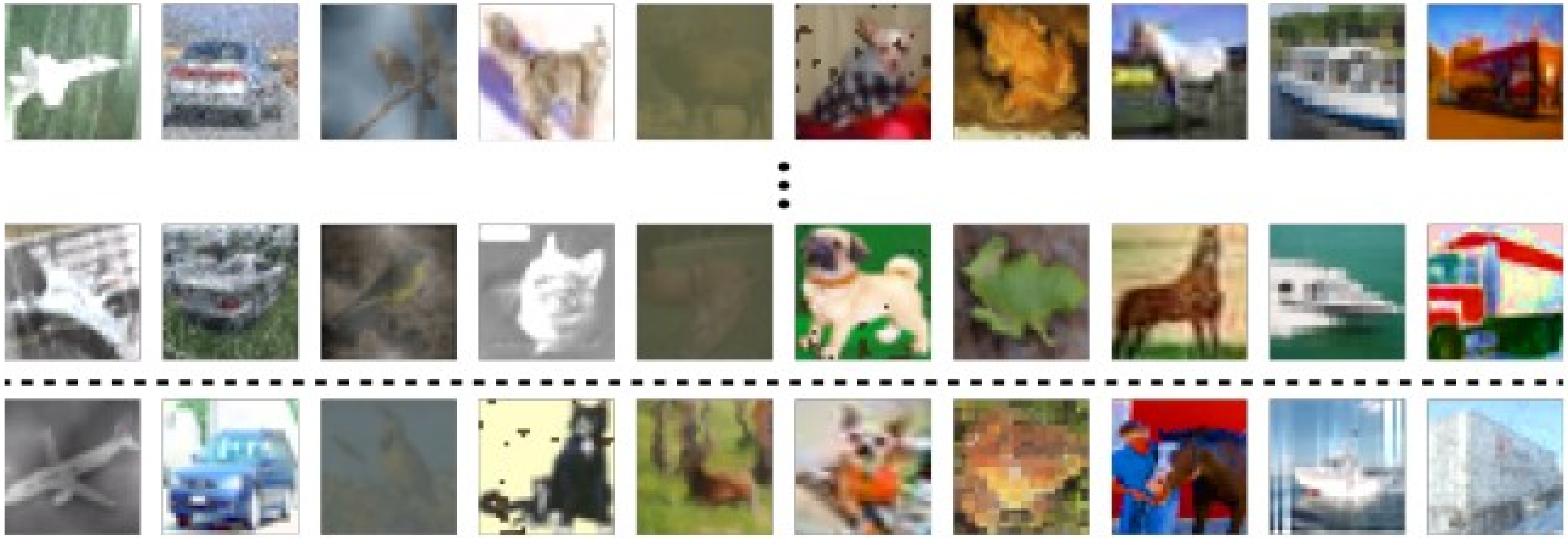
Unsupervised approaches

Recent trends

conclusions

Synthetic dataset

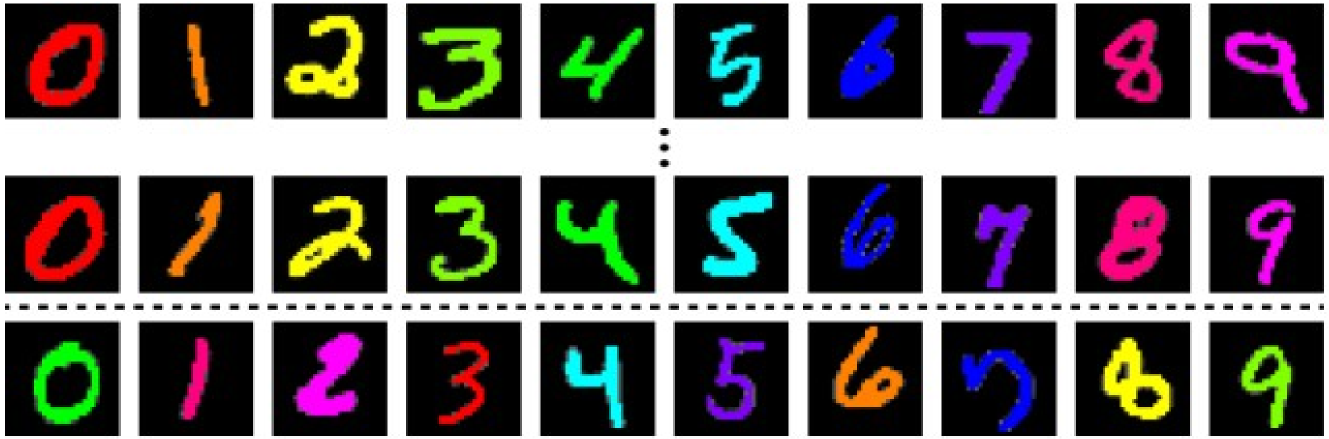
Corrupted Cifar-10



- 60.000 images
- 32x32 pixels
- Texture biases (*Brightness, Contrast, Gaussian Noise, Frost, Elastic Transform, Gaussian Blur, Defocus Blur, Impulse Noise, Saturate*)
- Training set with different rho
- Test set: 90% Bias-Conflicting and 10% aligned

Colored MNIST

Bias-aligned/biased



Bias-conflicting/unbiased

- 60.000 images
- 28x28 pixels
- Digit correlates with its color
- Training set with different rho
- Test set: 90% Bias-Conflicting and 10% aligned

Real-world datasets (1)

BAR (Biased Action Recognition)

BAR

- 2.595 images
- 224x224 pixels
- Bias: Setting in which an action is perfomed
- No bias annotation

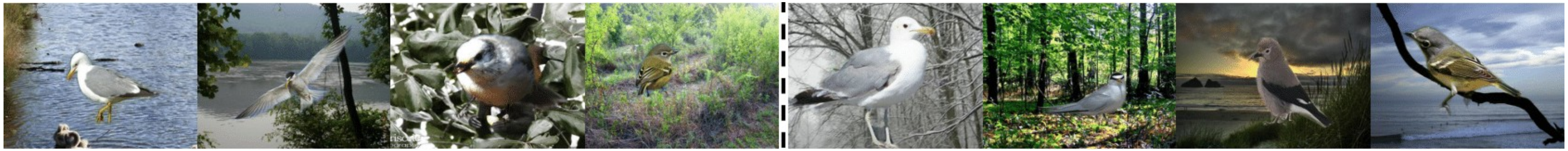
BFFHQ

- 21.200 images
- 224x224 pixels
- Bias: Gender
- Training set: 95% bias aligned
- Test set : Balanced

WATERBIRDS

- 11.968 images
- 224x224 pixels
- CUB + Places
- Bias: Background
- Training set: 95% bias aligned
- Test set : Balanced

BFFHQ from Flickr-Faces-HQ



Waterbirds

10

Real-world datasets

CelebA

- 202,599 images
- 224x224 pixels
- 40 annotated attributes;
- Several biases (e.g., color hair, gender, make-up)

CelebA
(Blond / Not Blond)

Bias mitigation approaches

- Intuitively, methods for mitigating the model's prediction dependency on bias;
- Increase the generalization and robustness of a trained model.

Landbird from waterbirds

Bias mitigation approaches

- Intuitively, methods for mitigating the model's prediction dependency on bias;
- Increase the generalization and robustness of a trained model.

Biased model

Bias mitigation approaches

- Intuitively, methods for mitigating the model's prediction dependency on bias;
- Increase the generalization and robustness of a trained model.

Biased model

Debiasing method

Debiased model

Outline

Introduction

Datasets

Supervised approaches

Unsupervised approaches

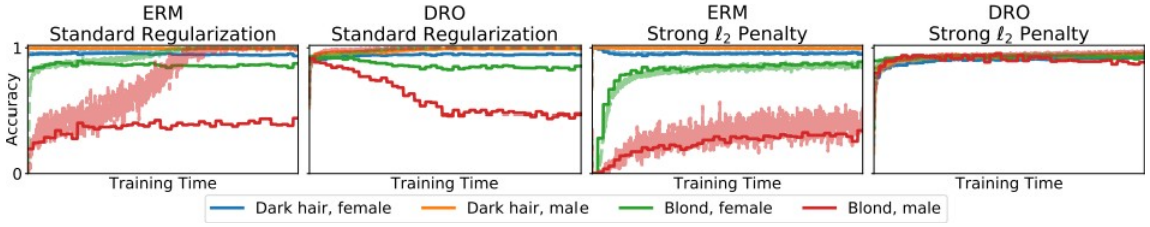
Recent trends

Conclusions

Model debiasing in Image Classification

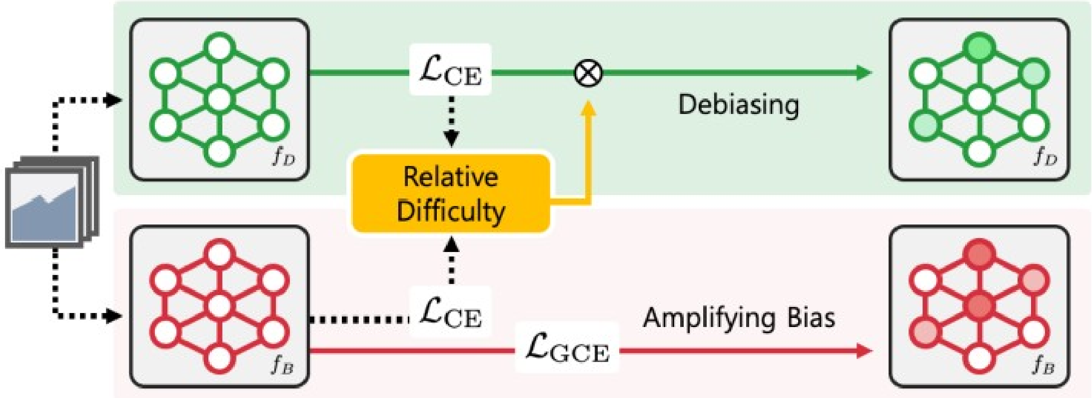
- **Supervised does not refer to target labels**
- Supervised indicates approaches relying on bias information for mitigation;
- Unsupervised debiasing do not assume any prior information on bias

Supervised (Bias label required)



Sagawa, Shiori, et al. "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization." *arXiv preprint arXiv*:1911.08731 (2019).

Unsupervised (No bias information)

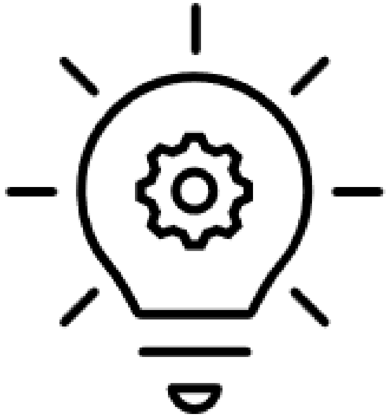


Nam, Junhyun, et al. “Learning from failure: De-biasing classifier from biased classifier.” *Advances in Neural Information Processing Systems* 33 (2020): 20673-20684.

Supervised approaches

Basic intuition

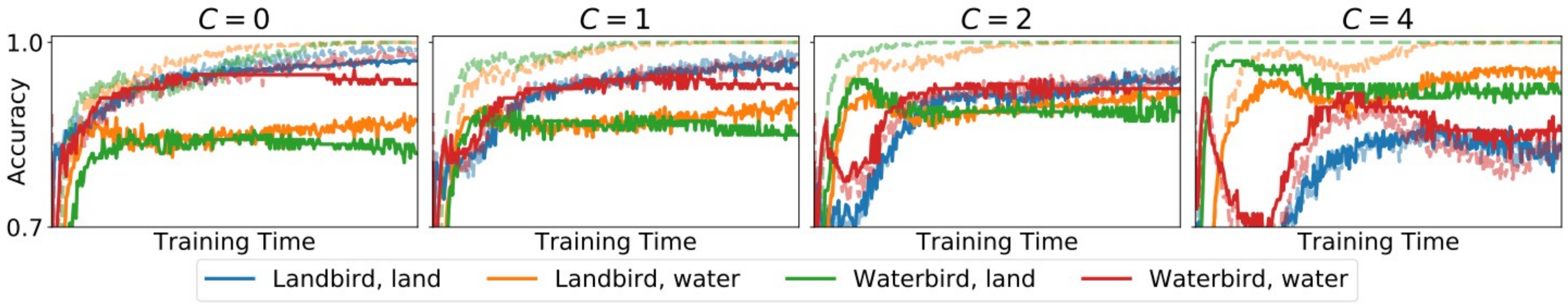
- If bias labels are known, it is possible to reweight or augment the training samples;
- The model can be forced to focus more on bias-conflicting samples;
- Debiasing can happen at the level of features or predictions



Fare clic per inserire note

Supervised approaches

Group optimization: GroupDRO



$$\hat{\theta}_{\mathrm{DRO}} := \underset{\theta\in\Theta}{\arg\min}\left\{\hat{\mathcal{R}}(\theta) := \max_{g\in\mathcal{G}} \mathbb{E}_{(x,y)\sim\hat{P}_g}[\ell(\theta;(x,y))]\right\}$$

$$\hat{\theta}_{\mathrm{adj}} := \underset{\theta\in\Theta}{\arg\min}\ \max_{g\in\mathcal{G}}\left\{\mathbb{E}_{(x,y)\sim\hat{P}_g}[\ell(\theta;(x,y))] + \frac{C}{\sqrt{n_g}}\right\}$$

Naïve alternative

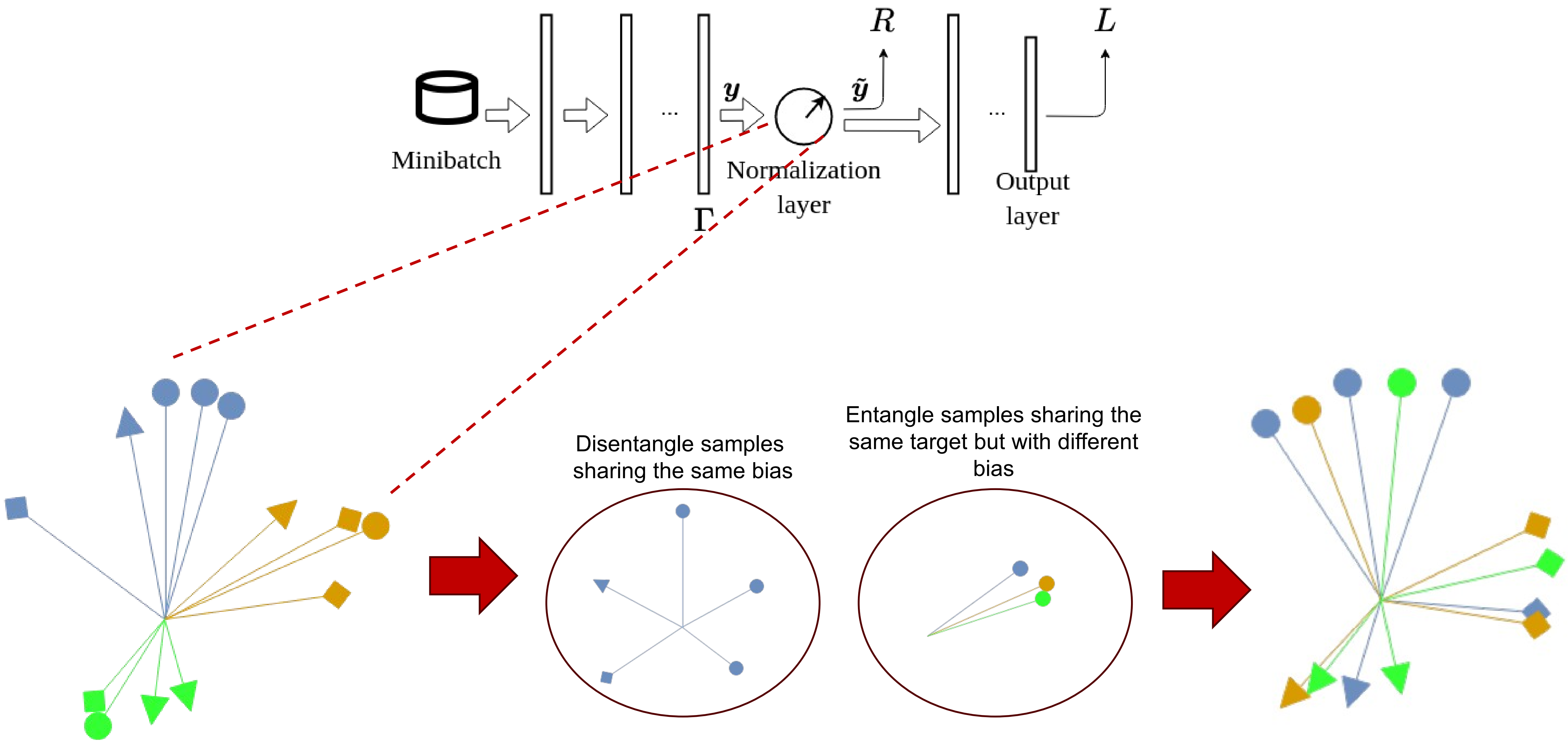
$$\hat{\theta}_w := \underset{\theta\in\Theta}{\arg\min}\ \mathbb{E}_{(x,y,g)\sim\hat{P}}[w_g\,\ell(\theta;(x,y))]$$

$$w_g = 1/\mathbb{E}_{g'\sim\hat{P}}[\mathbb{I}(g' = g)]$$

Sagawa, Shiori, et al. "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization." *arXiv preprint arXiv:1911.08731* (2019).

Supervised approaches

End: Features disentanglement



Tartaglione, E., et al., (2021). End: Entangling and disentangling deep representations for bias correction. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13508-13517).

Supervised approaches

Recap

- Supervised approaches rely on bias information for model debiasing;
- They include dataset cleaning, post-processing or in-model approaches.
- They are usually more accurate than unsupervised counterpart;

Fare clic per inserire note

Outline

Introduction

Datasets

Supervised approaches

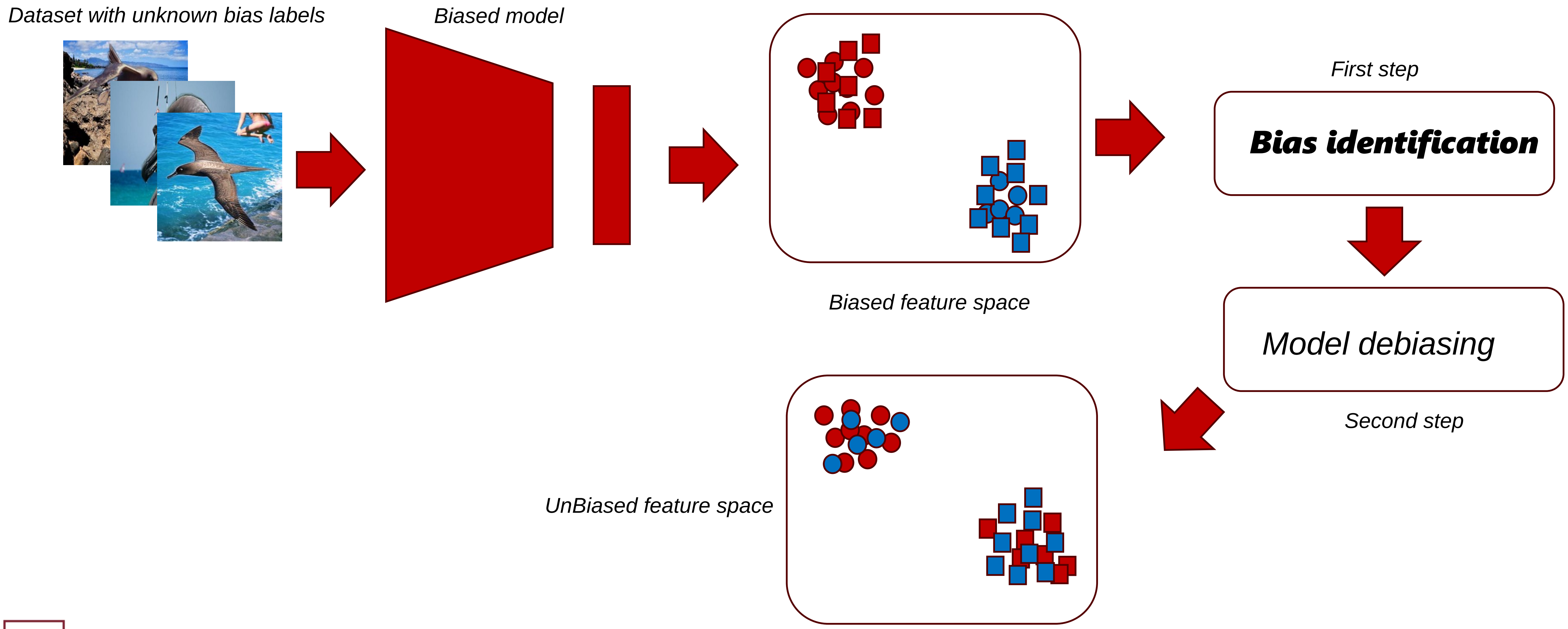
Unsupervised approaches

Recent trends

Conclusions

Towards unsupervised debiasing

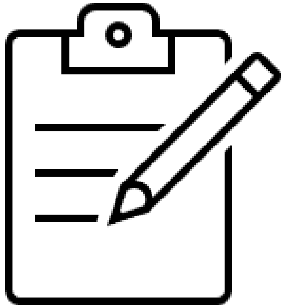
- How can we reweight samples for bias mitigation, if bias is unknown?



Bias identification

- Bias identification allows to produce pseudolabels that can be used for debiasing;
- Methods exploit the feature space (e.g., MoDAD, George) or predictions (e.g., JTT);
- The more precise bias identification, the better the debiasing performance.

Model debiasing



- Bias-conflicting augmentation and upsampling (e.g., MoDAD, Just Train Twice);
- Loss re-weighting (e.g., Learning with a Biased Committee);
- Adversarial debiasing (e.g., BiasAdv).

Two-steps unsupervised approaches

Just train twice

- **A biased model has a higher probability to misclassify a bias-conflicting sample.**
- Two-step method: bias identification + debiasing
- The error set is identified as: $E = \{(x_i, y_i) \text{ s.t. } \hat{f}_{\text{id}}(x_i) \neq y_i\}.$
- ERM up sampling the samples in the error set (predicted bias-conflicting)

$$J_{\text{up-ERM}}(\theta, E) = \left(\lambda_{\text{up}} \sum_{(x,y) \in E} \ell(x, y; \theta) + \sum_{(x,y) \notin E} \ell(x, y; \theta) \right)$$

	Waterbirds worst-group test acc.		CelebA worst-group test acc.	
	Tuned for average	Tuned for worst-group	Tuned for average	Tuned for worst-group
CVaR DRO (Levy et al., 2020)	62.0%	75.9%	36.1%	64.4%
LfF (Nam et al., 2020)	44.1%	78.0%	24.4%	77.2%
JTT (Ours)	62.5%	86.7%	40.6%	81.1%

Fare cli

Liu, Evan Z., et al. "Just train twice: Improving group robustness without training group information." International Conference on Machine Learning. PMLR, 2021.

Two-steps unsupervised approaches

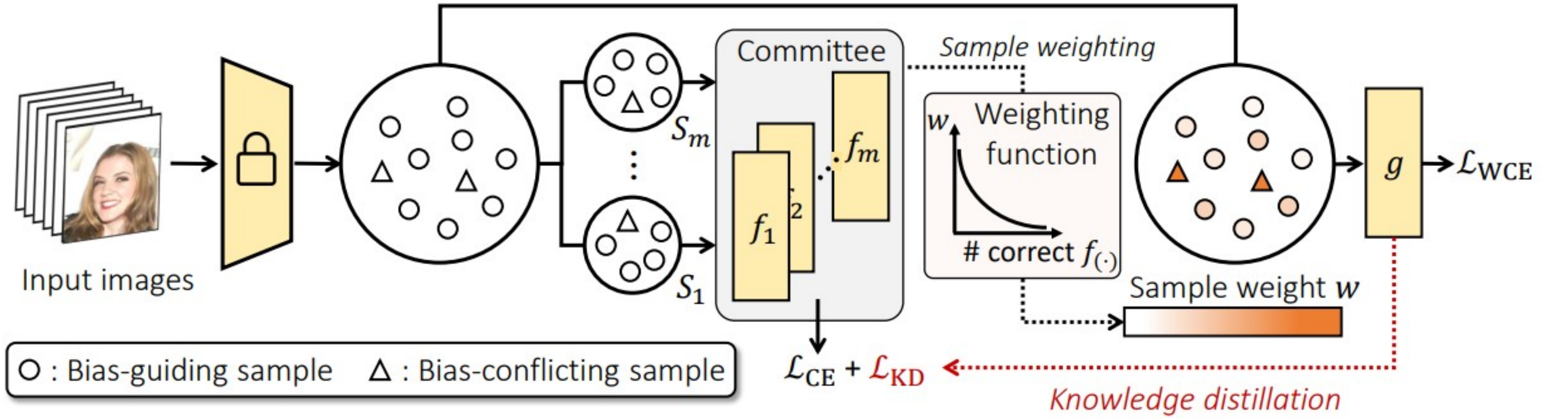
Learning with Bias Committee

- **Bias-Aligned samples are correctly classified by a committee**
- Backbone pre-trained with BYOL
- Random sample of *m* subsets
- Weight based on consensus

$$w(x) = \frac{1}{\sum_{l=1}^{m} \mathbb{1}(f_l(x) = y)/m + \alpha}$$

- Weighted ERM

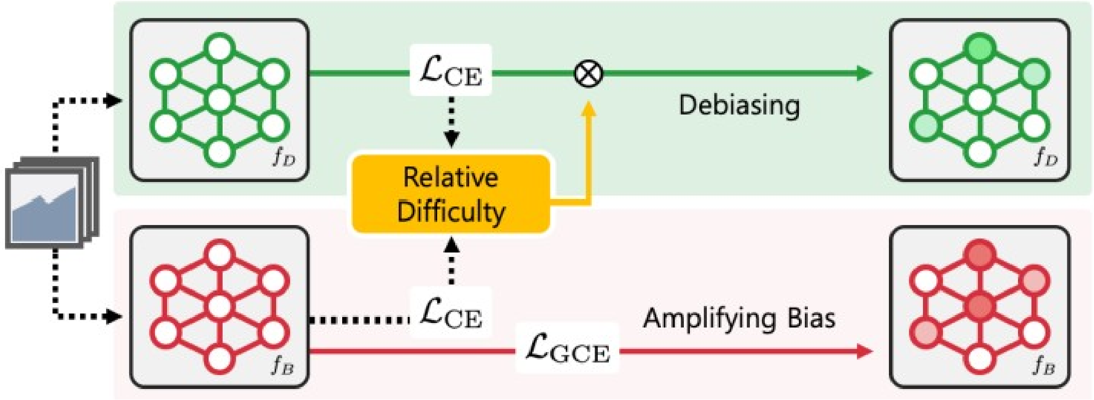
$$\mathcal{L}_{\text{WCE}} = \sum_{(x,y)\in\mathcal{B}} w(x) \cdot \text{CE}(g(x), y),$$



Kim, Nayeong, et al. "Learning debiased classifier with biased committee." *Advances in Neural Information Processing Systems* 35 (2022): 18403-18415.

End-to-end methods

- Model debiasing is performed without requiring bias identification
- They usually still employ an auxiliary models to provide *indirect* information on bias

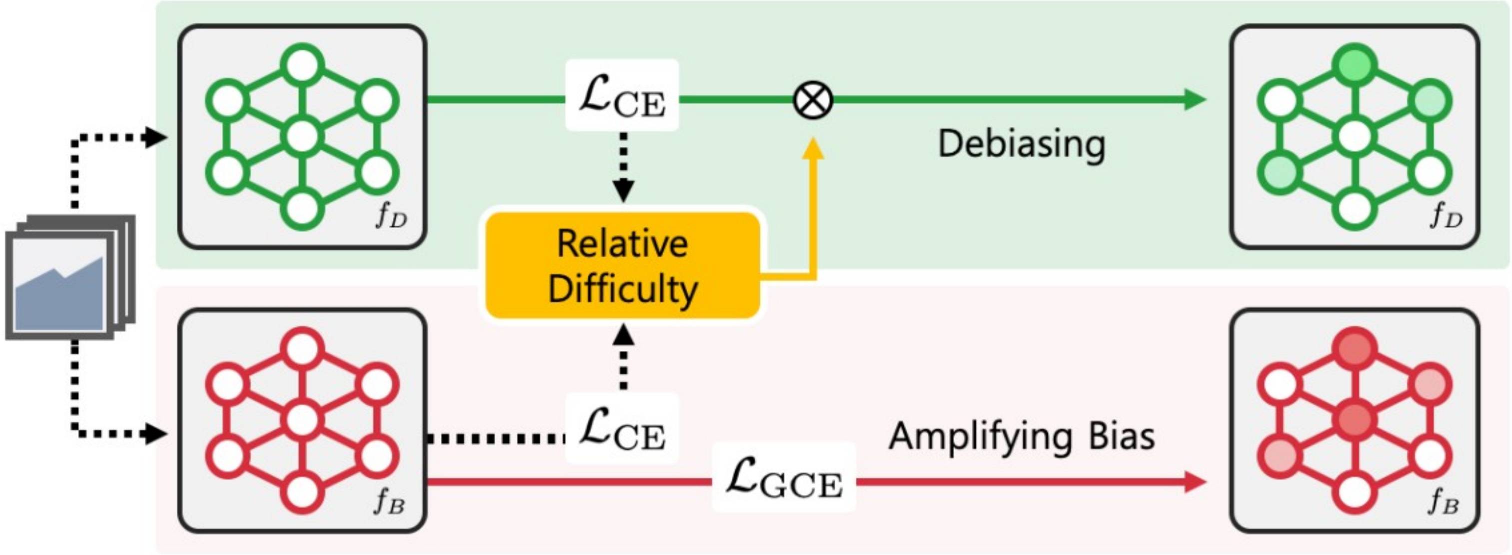


Nam, Junhyun, et al. "Learning from failure: De-biasing classifier from biased classifier." *Advances in Neural Information Processing Systems* 33 (2020): 20673-20684.

Learning from Failure

- **Bias affects the model only if it is easier to learn than the target attribute**;
- GCE loss function to amplify easy samples (bias-aligned);
- Model D (Debiased) is trained with Weighted CE according to:

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



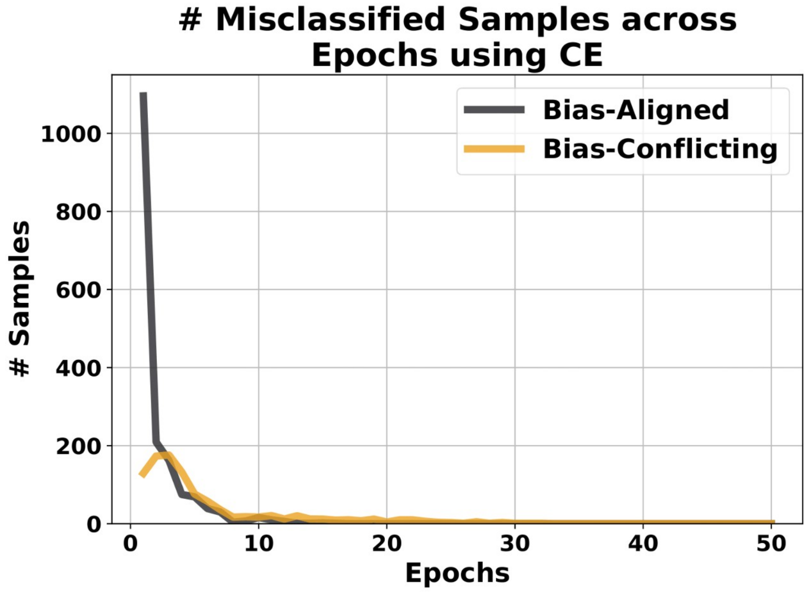
$$\mathrm{GCE}(p(x;\theta), y) = \frac{1 - p_y(x;\theta)^q}{q} \qquad \frac{\partial \mathrm{GCE}(p, y)}{\partial \theta} = p_y^q \frac{\partial \mathrm{CE}(p, y)}{\partial \theta}$$

Fare clic per inserire note

Nam, Junhyun, et al. "Learning from failure: De-biasing classifier from biased classifier." Advances in Neural Information Processing Systems 33 (2020): 20673-20684.

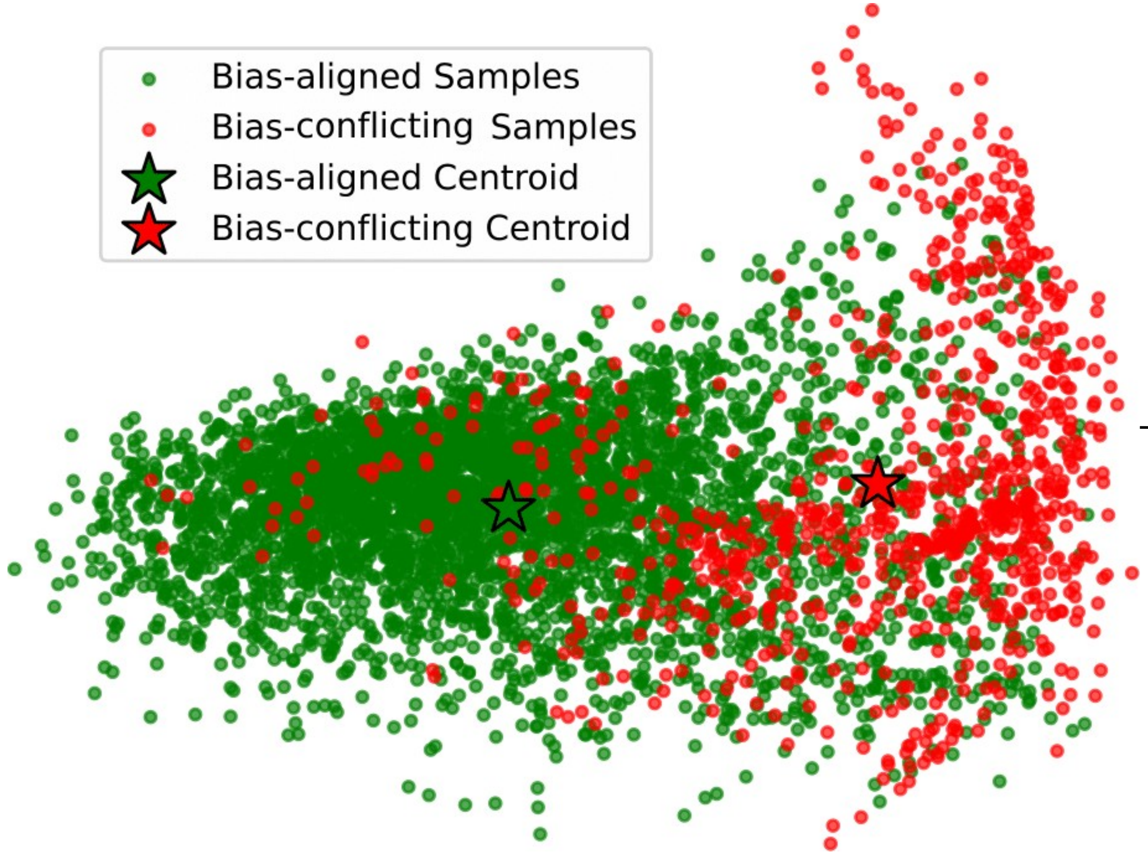
Open challenges in model debiasing

- How to obtain a precise bias identification;
- How to avoid using bias annotated (or not) validation sets;
- How to avoid bias-conflicting memorization;
- How to discover bias in models;
- Real-world datasets for benchmarking.

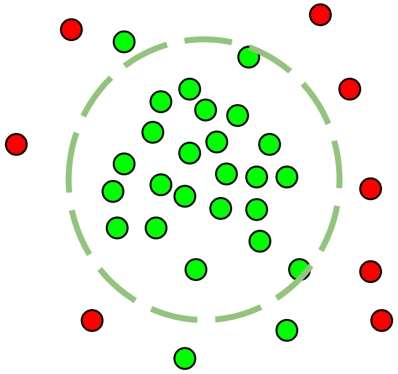


Looking at model debiasing through the lens of anomaly detection (MoDAD)

Principal Component Analysis



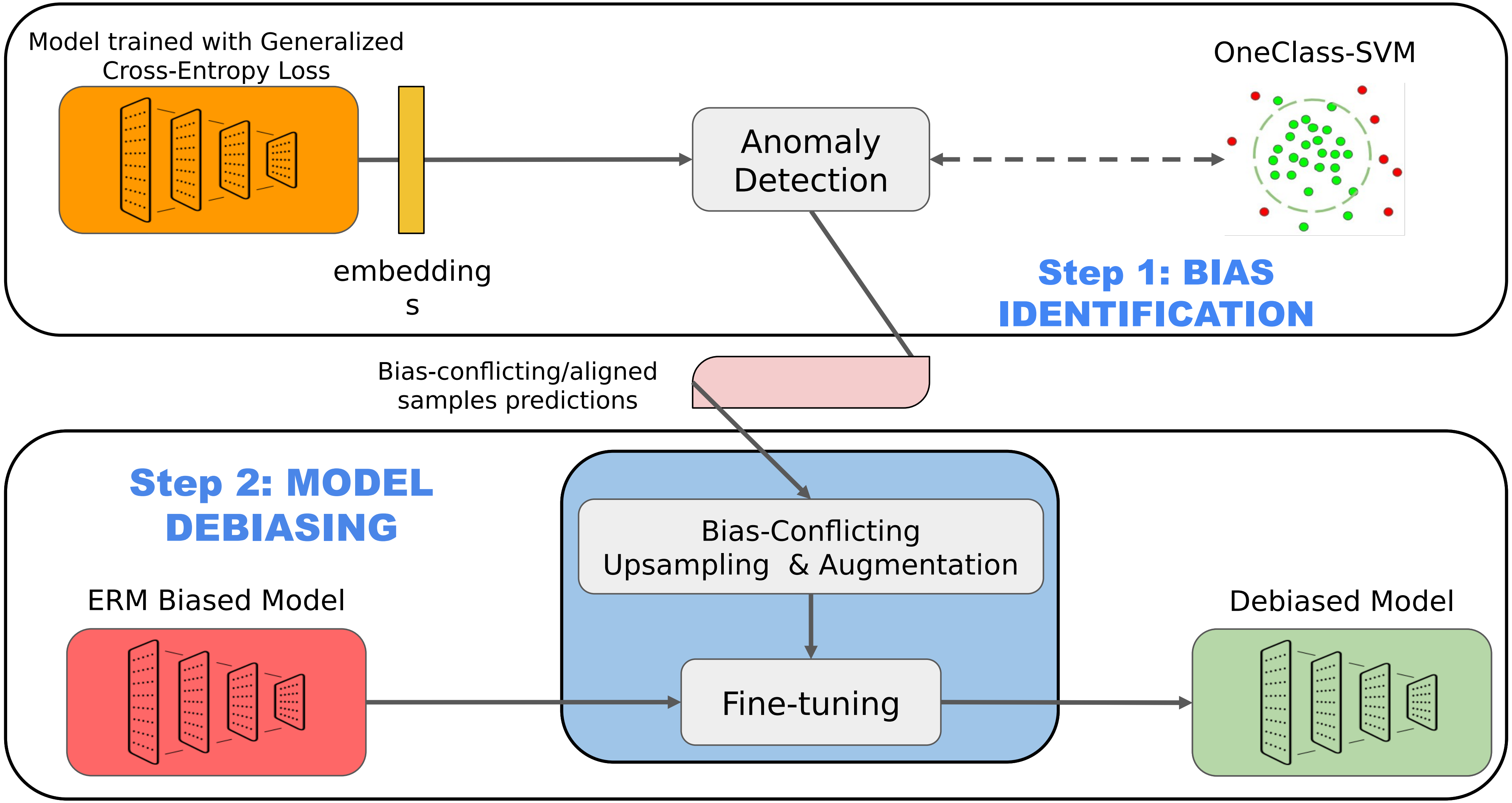
One-Class Support Vector Machine



Anomaly Detection

Pastore, V. P., Ciranni, M., Marinelli, D., Odone, F., & Murino, V. (2025, February). Looking at Model Debiasing through the Lens of Anomaly Detection. In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 2548-2557). IEEE.

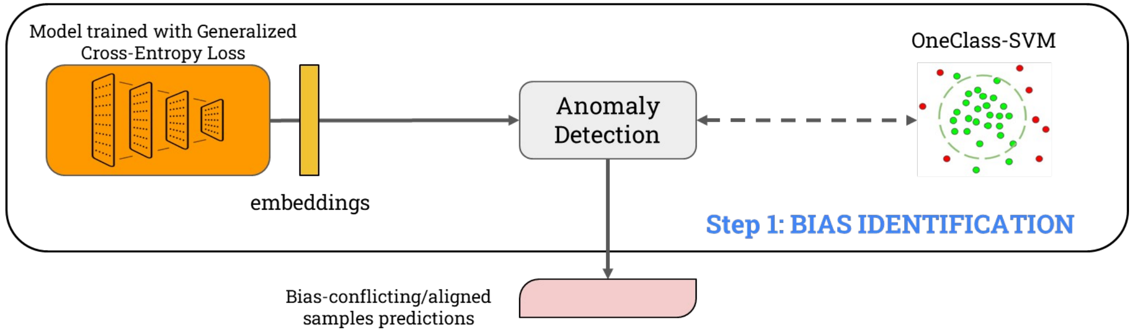
Method overview



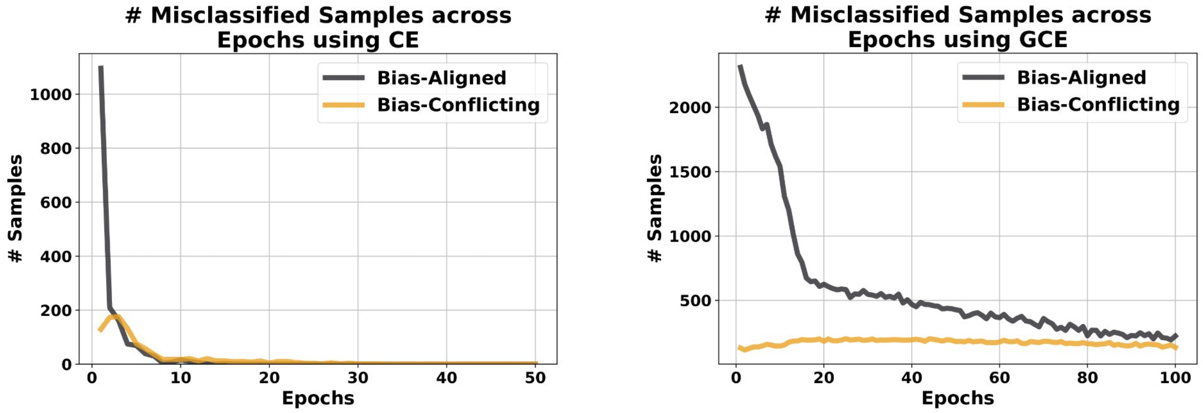
Pastore, V. P., Ciranni, M., Marinelli, D., Odone, F., & Murino, V. (2025, February). Looking at Model Debiasing through the Lens of Anomaly Detection. In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 2548-2557). IEEE.

Bias identification

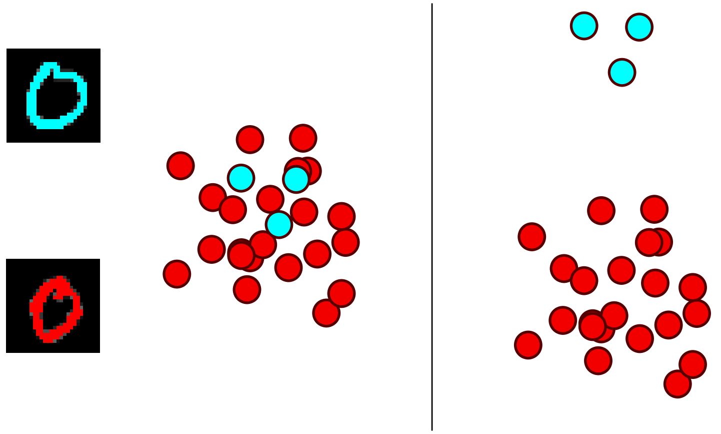
The impact of GCE loss function



- *The more precise bias-identification, the more effective model debiasing*

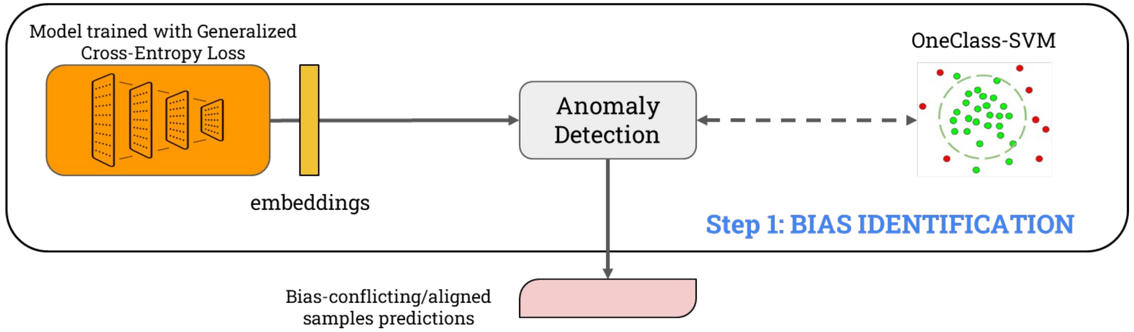


Misclassified training-set samples on Waterbirds dataset



Training Feature space: CE (left) , GCE (right)

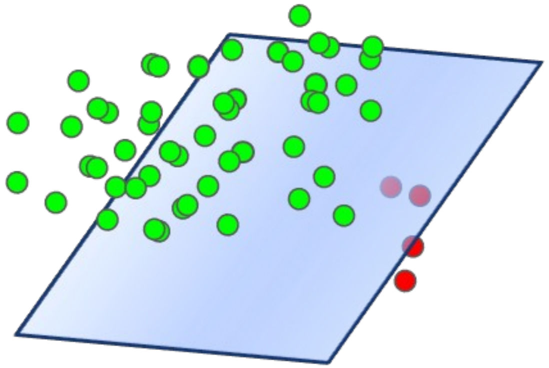
Bias identification



Anomaly detection

Modified One-Class Support Vector Machine

OCSVM

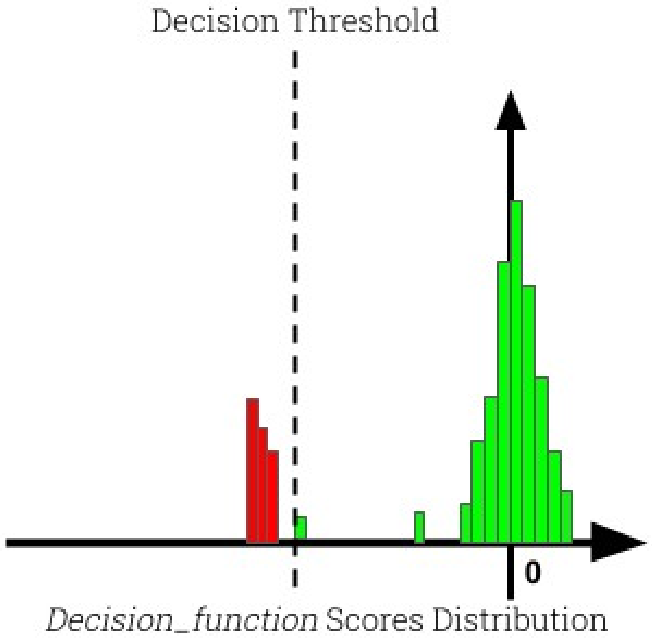


$$f(x) = sign(\sum_{i=1}^{m} \alpha_i \, K(x, x_i) - (\lambda + \ \tau))$$

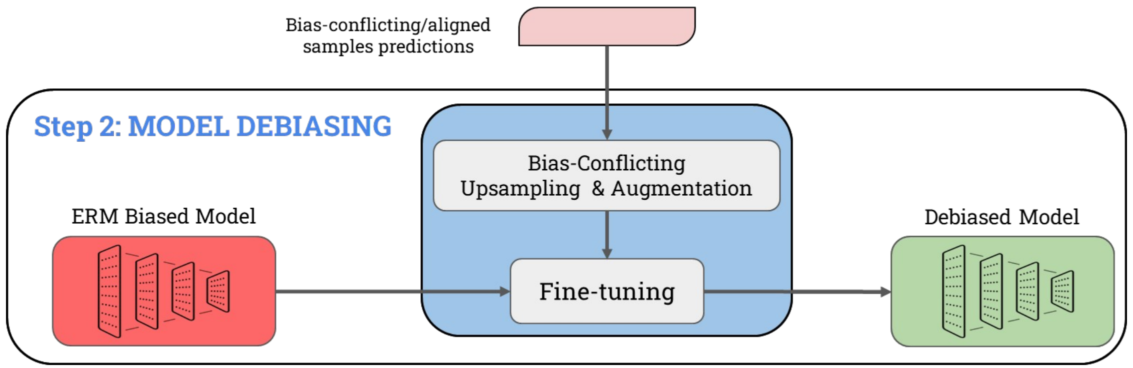
- N_i = |Samples of class *i*|
- C_i = |Correctly classified samples of class *i*|
- $r = 0.5$

$$p_i = \frac{N_i - C_i}{N_i} r \cdot 100$$

$\tau = percentile\ (scores_i\,,\, p_i)$

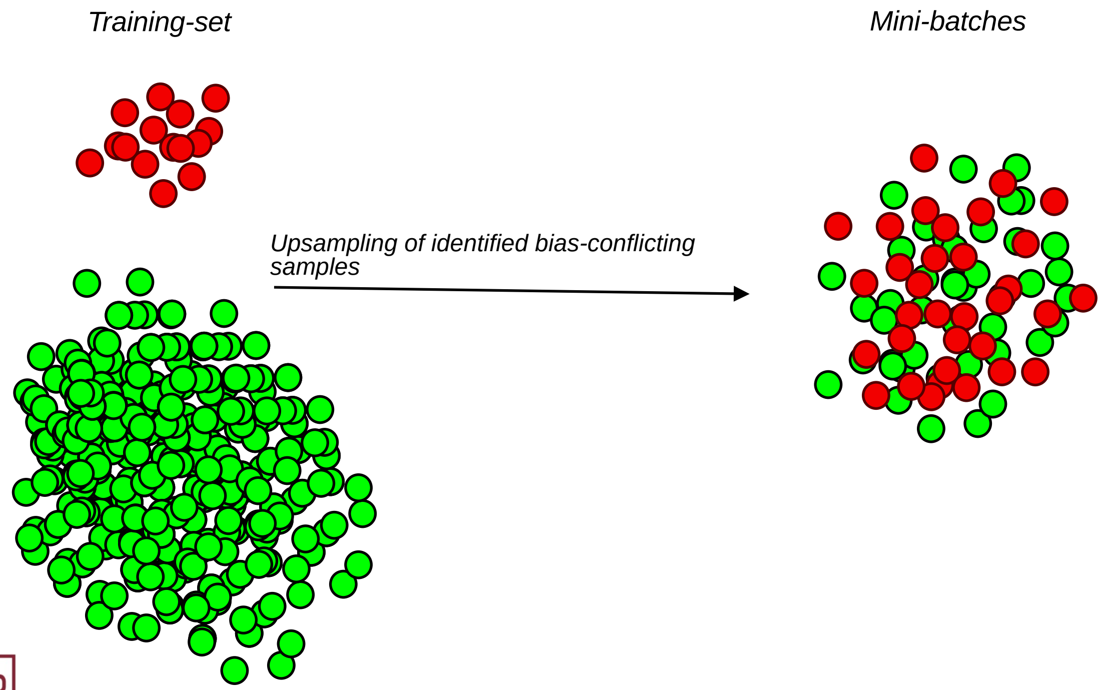


Model debiasing



- *Upsampling bias-conflicting samples (with DA);*
- *Weighted random sampler*
- *A biased model is debiased using this approach.*

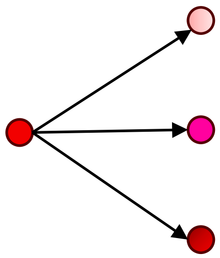
Upsampling



Data Augmentation

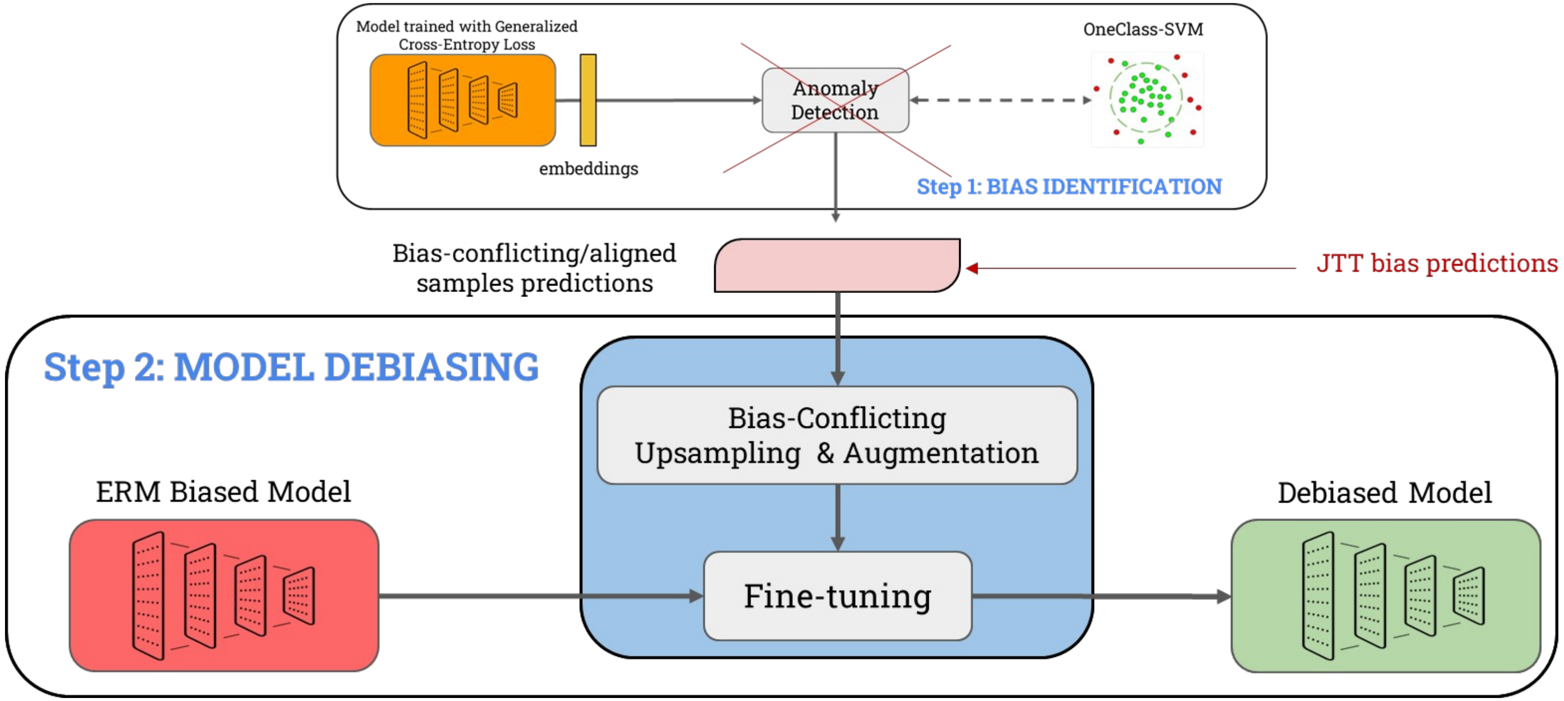
Geometric and color-space transformations

—*Randomverticalflip*
—*RandomRotation*
—*RandomAutoContrast*
—*CenterCrop*



Is a precise bias identification really important?

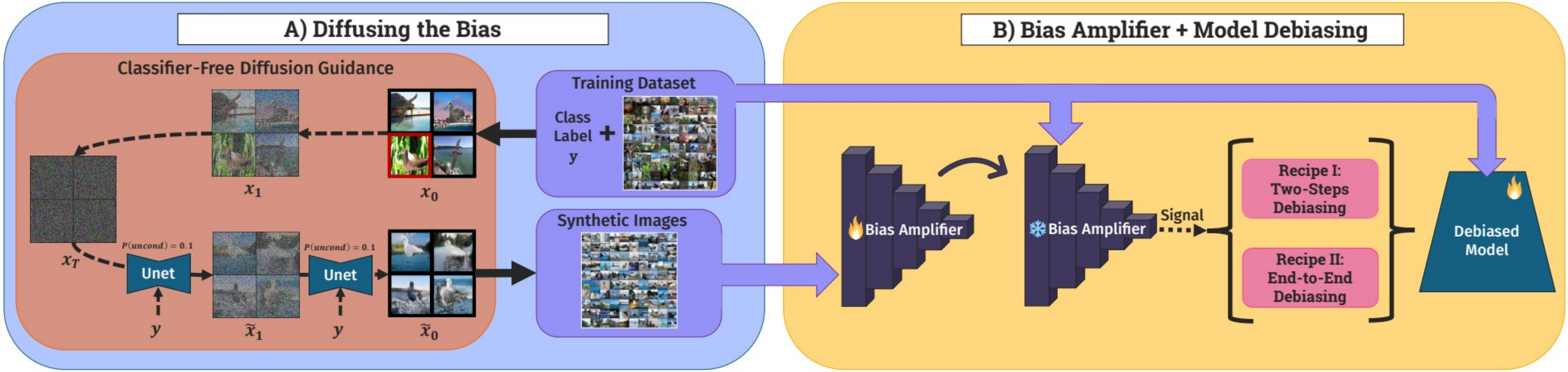
- JTT bias predictions + MoDAD step 2 -> **- 1.63 %** Conflicting accuracy;
- MoDAD bias predictions + JTT debiasing -> **+ 1 %** Conflicting accuracy w.r.t. JTT;



How does the model change in making predictions?

DiffusingDeBias (DDB): solving memorization by construction

General overview

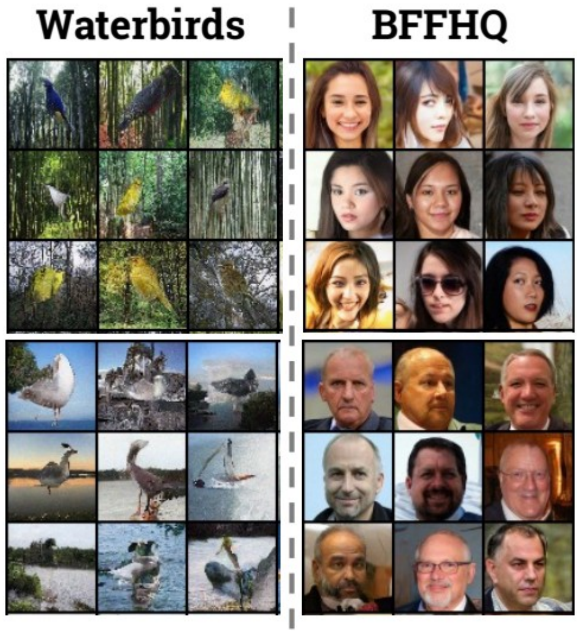


Ciranni, M., Pastore, V. P., Di Via, R., Tartaglione, E., Odone, F., & Murino, V. (2025). Diffusing DeBias: Synthetic Bias Amplification for Model Debiasing. arXiv preprint arXiv:2502.09564.

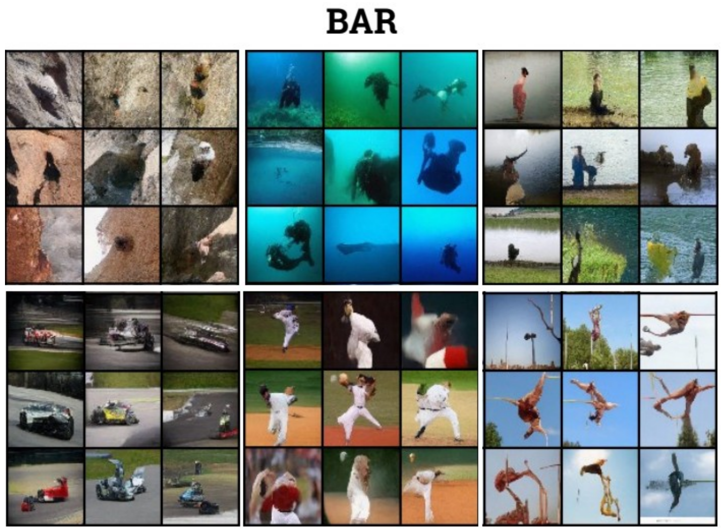
Fare clic per inserire note

DiffusingDeBias (DDB)

Basic concepts



- Diffusion model can amplify bias present in training data;
- Such property allows to obtain the generation of a *purer* bias-aligned distribution;
- This syntethic data can be used for training an auxiliary model;
- Ideally, this can be plugged into any debiasing method;
- Memorization solved by construction;
- Validation set is not employed for training the auxiliary model.



Fare clic per inserire note

Impact on performance

Method	Unsup	Waterbirds	BAR	BFFHQ		ImageNet-A
		WGA	Avg.	Avg.	Confl.	Avg.
ERM	–	62.60 ± 0.30	51.85 ± 5.92	–	60.13 ± 0.46	30.30
LISA [48]	–	89.20	–	–	–	–
G-DRO [42]	–	91.40 ± 1.10	–	–	–	–
George [43]	✓	76.20 ± 2.00	–	–	–	–
JTT [34]	✓	83.80 ± 1.20	68.53 ± 3.29	–	62.20 ± 1.34	–
CNC [51]	✓	88.50 ± 0.30	–	–	–	–
B2T+G-DRO [24]	✓	90.70 ± [illegible]	[illegible]	[illegible]	[illegible]	–
LfF [38]	✓	78.00	62.98 ± 2.76	–	62.97 ± 3.22	–
ETF-Debias [46]	✓	–	–	–	73.60 ± 1.22	–
Park et al. [39]	✓	–	–	71.68	–	–
LWBC [22]	✓	–	62.03 ± 2.76	–	–	35.97 ±0.49
CDvG+LfF [18]	✓	84.80	–	–	62.20 ±0.45	34.60
DebiAN [32]	✓	–	69.88 ± 2.92	–	62.80 ± 0.60	–
MoDAD [40]	✓	89.43 ± 1.69	69.83 ± 0.72	–	68.33 ± 2.89	–
DDB-II (ours)	✓	**91.56** ± 0.15	**72.81** ± 1.02	**83.15** ± 1.76	70.93 ± 0.14	37.53 ±0.82
DDB-I (ours)	✓	90.81 ± 0.68	70.40 ± 1.41	81.27 ± 0.88	**74.67** ± 2.37	**39.80** ±0.50
DDB-I (w/ err. set)	✓	90.34 ± 0.41	70.59 ± 0.19	82.44 ± 0.64	71.40 ± 0.92	38.12 ±0.96

Fare clic per inserire note

Outline

Introduction

Datasets

Supervised approaches

Unsupervised approaches

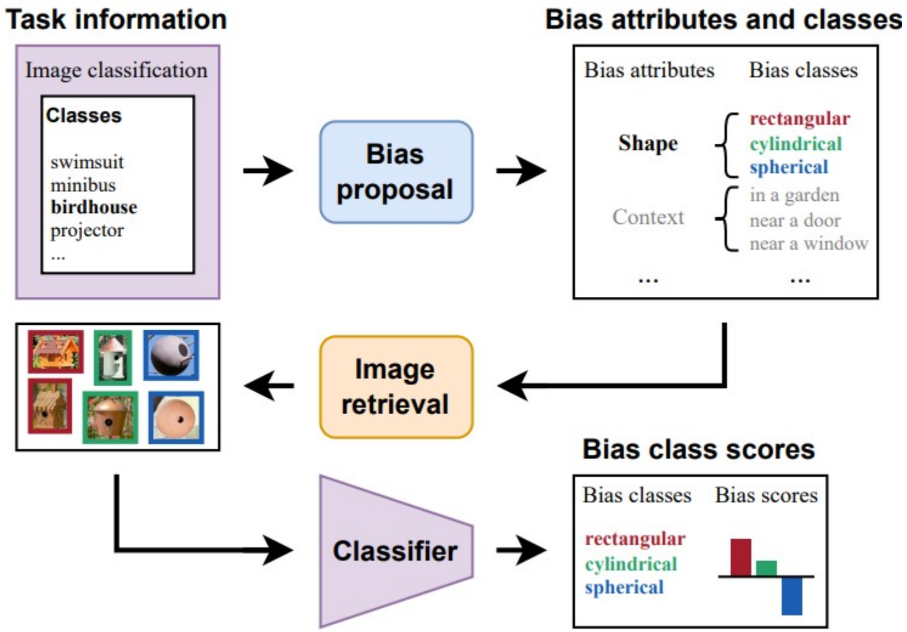
Recent trends

Challenges and conclusions

Bias discovery frameworks

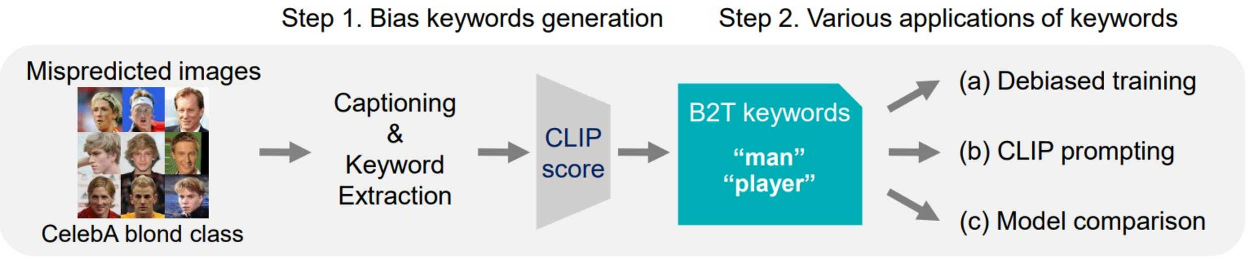
- Methods that inspect trained models to provide information on potential biases;
- Typically, the primary aim is not to debias but to expose bias

Classifier-To-Bias (C2B)



Guimard, Q., D'Incà, M., Mancini, M., & Ricci, E. (2025). Classifier-to-Bias: Toward Unsupervised Automatic Bias Detection for Visual Classifiers. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 15151-15161).

Bias-To-Text (B2T)

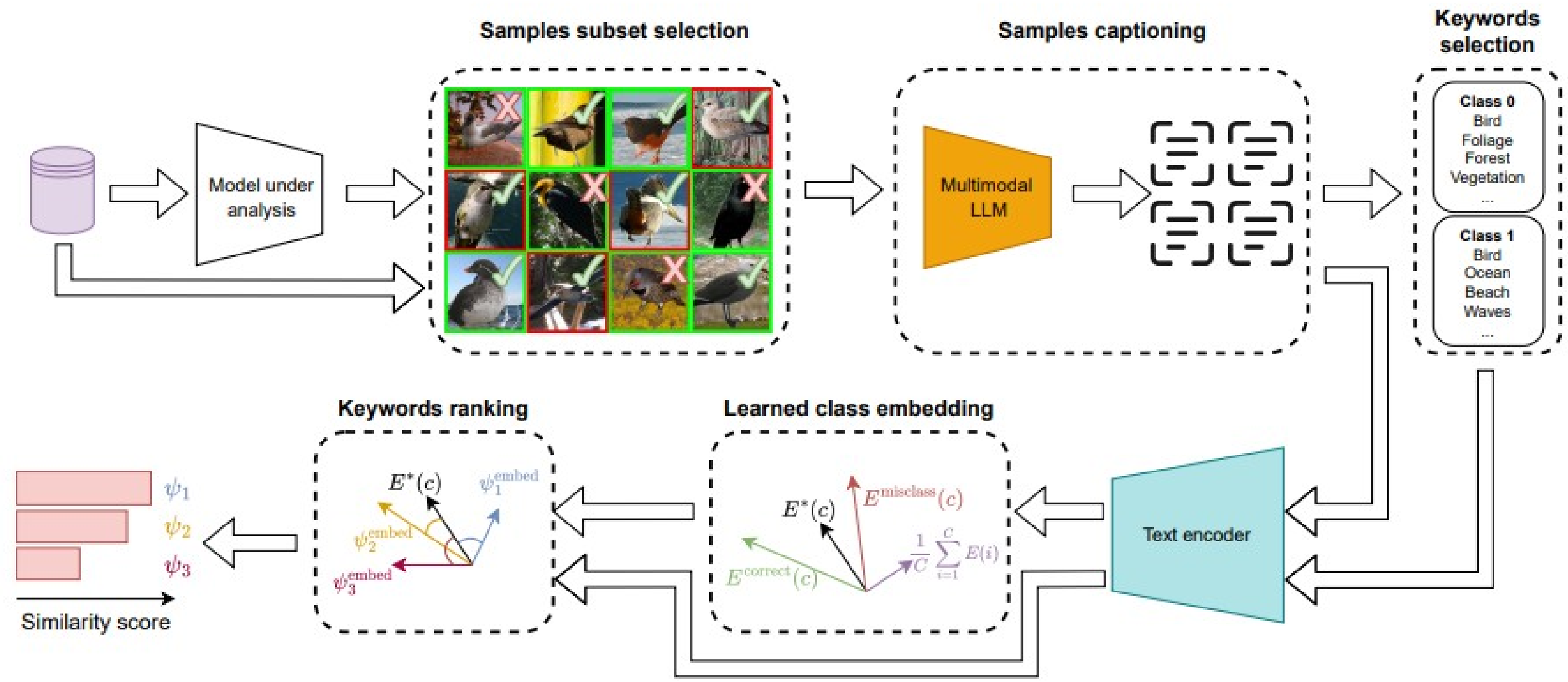


Kim, Y., Mo, S., Kim, M., Lee, K., Lee, J., & Shin, J. (2024). Discovering and mitigating visual biases through keyword explanation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11082-11092).

Bias discovery frameworks

- Methods that inspect trained models to provide information on potential biases;
- Typically, the primary aim is not to debias but to expose bias

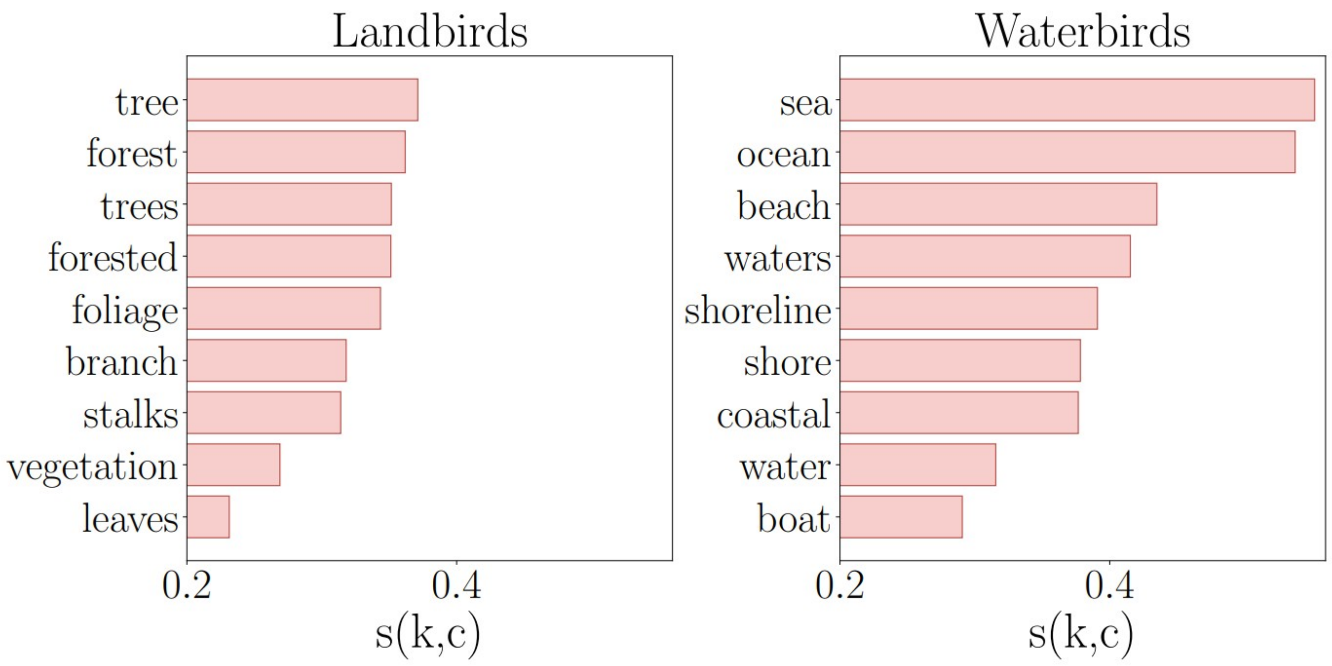
Say My Name (SaMyNa)



Ciranni, M., Molinaro, L., Barbano, C. A., Fiandrotti, A., Murino, V., Pastore, V. P., & Tartaglione, E. (2024). Say My Name: a Model's Bias Discovery Framework. arXiv preprint arXiv:2408.09570.

Bias discovery frameworks

Examples of SaMyNa generated bias keywords on waterbirds



Ciranni, M., Molinaro, L., Barbano, C. A., Fiandrotti, A., Murino, V., Pastore, V. P., & Tartaglione, E. (2024). Say My Name: a Model's Bias Discovery Framework. arXiv preprint arXiv:2408.09570.

Dealing with multiple biases

- In case of multiple biases, many debiasing methods end-up mitigating one attribute while amplifying the dependency on the other one

	I.D. Acc	shortcut reliance		
		BG Gap ↑	CoObj Gap ↑	BG+CoObj Gap ↑
ERM	97.6	-15.3	-11.2	-69.2
Mixup	98.3	-12.6	-9.3	-61.8
CutMix	96.6	-45.0 (×2.94)	-4.8	-86.5
Cutout	97.8	-15.8 (×1.03)	-10.4	-71.4
AugMix	98.2	-10.3	-12.1 (×1.08)	-70.2
SD	97.3	-15.0	-3.6	-36.1
CF+F Aug	96.8	-16.0 (×1.04)	**+0.4**	-19.4
LfF	97.2	-11.6	-18.4 (×1.64)	-63.2
JTT (E=1)	95.9	-8.1	-13.3 (×1.18)	-40.1
EIIL (E=1)	95.5	-4.2	-24.7 (×2.21)	-44.9
JTT (E=2)	94.6	-23.3 (×1.52)	-5.3	-52.1
EIIL (E=2)	95.5	-21.5 (×1.40)	-6.8	-49.6
DebiAN	98.0	-14.9	-10.5	-69.0
LLE (ours)	96.7	**-2.1**	-2.7	**-5.9**

◆ Urban cars: target classes are country cars and urban cars;
◆ bias are backgrounds and co-occuring objects

Li, Z., Evtimov, I., Gordo, A., Hazirbas, C., Hassner, T., Ferrer, C. C., ... & Ibrahim, M. (2023). A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20071-20082).

Outline

Introduction

Bias in image classification

Debiasing Approaches

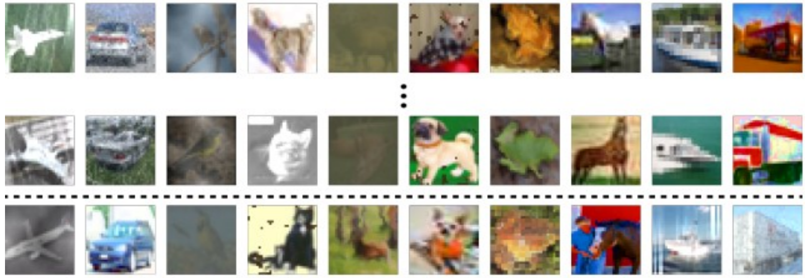
Proposed method

Experiments

Conclusions

Conclusions and takeaways

- Bias is a significant problem harnessing AI’s application to real-world problems;
- Bias is **inherent in the data** as in humans who generate it;
- Shortcuts corresponding to bias learned by a model;
- Methods for model debiasing can be divided into supervised and unsupervised;
- Unsupervised methods can be further categorized as two-step or end-to-end;
- Open challenges include precise bias identification, validation sets, but also unrealistic datasets.
- Bias in specific domain may be hard to discover, and to mitigate.



Contact

vito.paolo.pastore@unige.it

More information on my research on:

vitopaolopastore.github.io