南京航空航天大學
Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Exploring data bias: from the known to the unknown

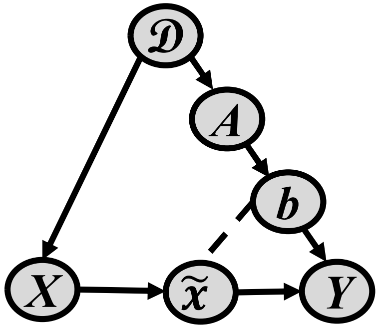
Prelude

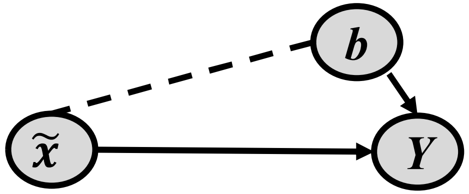
- Bias is a reflection of **real-world structure**, but they affect the fairness of the model and may also lead to a decrease in the model's generalizability in real-world applications.

- Format：
  - ✓ spurious correlation
  - ✓ distribution shift
  - ✓ Shortcut learning — Texture bias, background bias scene bias ....
- Way：
  - ✓ Inductive Bias [**No free lunch theorem**]
- Aspect：
  - Known bias
  - Unknown bias

[1] Geirhos R et al. Shortcut learning in deep neural networks. (Nature Machine Intelligence, 2020)

Statistical aspects





$$P\,(y\mid do(x)) = \sum_{A} P\,(y\mid do(x), a)P(a\mid do(x)) = \sum_{A} P\,(y\mid x, a)P(a)$$

If b is a complete bias from X, then the goal is to <u>make b and Y as independent as possible</u>

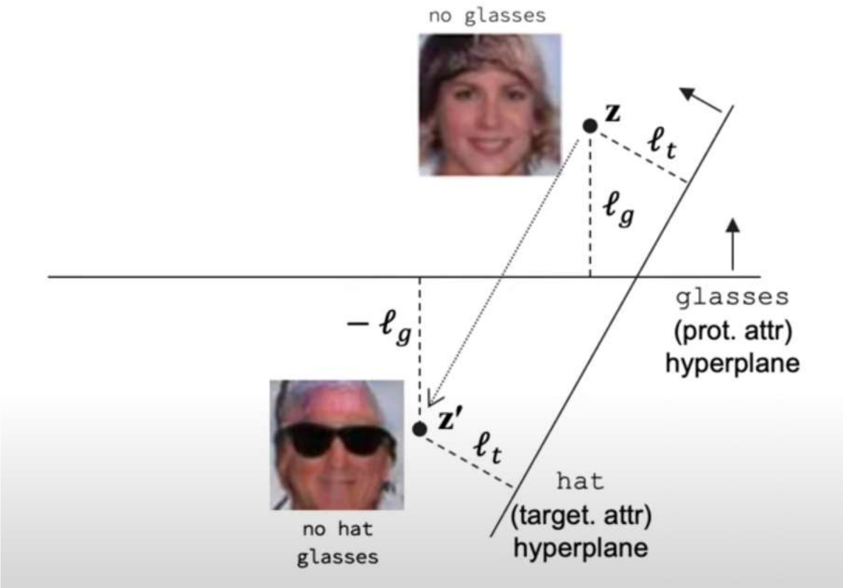
- Random experiment
- $P(Y|B) = P(Y)$
- $\min I(\tilde{X}; B)\quad i.e. \min D_{KL}(p(\tilde{x}, b)||P(\tilde{x})P(b))$

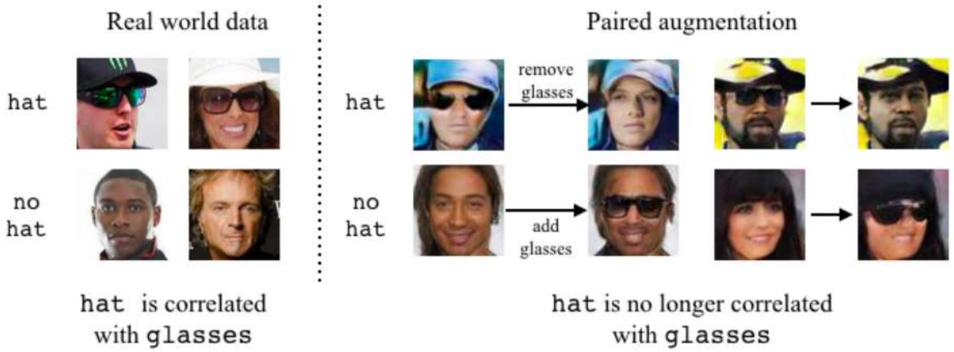
Fair data space

Fair Attribute Classification through Latent Space De-biasing

Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky
Princeton University
{vr23, suhk, olgarus}@cs.princeton.edu

- Constructing a fair dataset by GAN-based data augmentation
  - Assume: linearly differentiable in semantic properties
  - Target: Learning interpretable operational directions, constructing complementary vectors





[2] Ramaswamy V V et al. Fair attribute classification through latent space de-biasing(CVPR2022).

Fair sampling distribution

Bias Mimicking: A Simple Sampling Approach for Bias Mitigation

Maan Qraitem[1], Kate Saenko[1,2], Bryan A. Plummer[1]
[1]Boston University [2]MIT-IBM Watson AI Lab
{mqraitem, saenko, bplum}@bu.edu

- We want to guarantee $P_D(Y|B) = P_D(Y)$

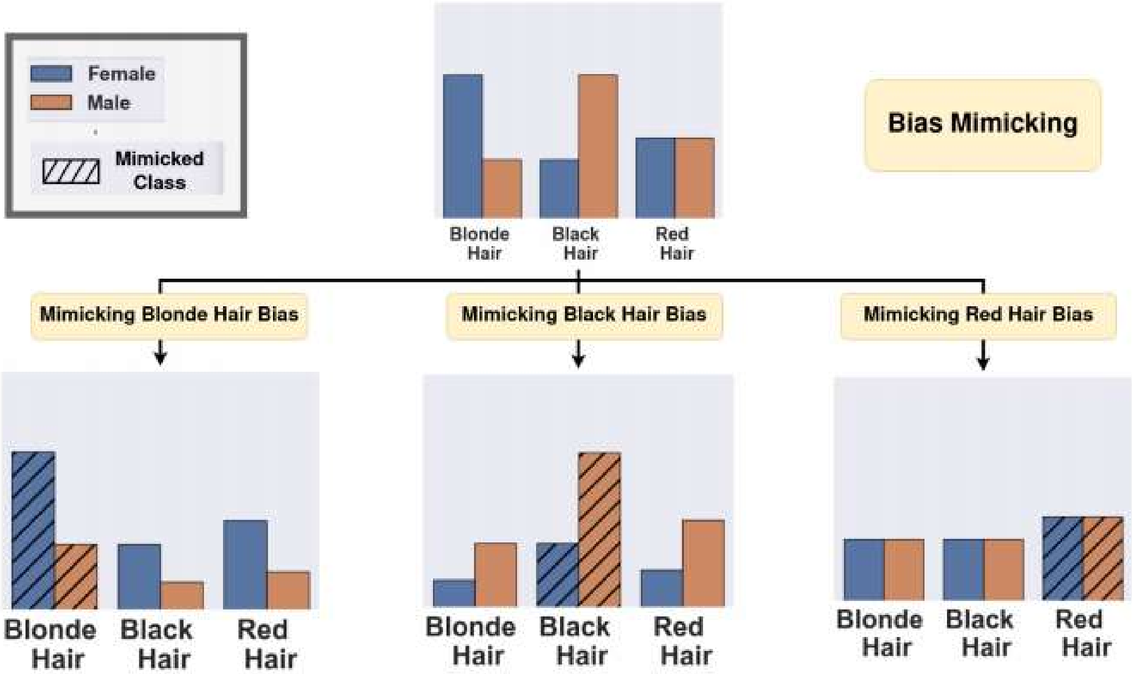
- Bias Mimicking: Given class "c" :
  Ensure that is $P_D(B|Y = c)$
  "mimicked" in each other class

$$P_{d_c}(B = s|Y = c) = P_{d_c}(B = s|Y = c') \quad \forall s \in S.$$

$$P_D(B = s) = \sum_{c \in C} P_D(B = s|Y = c) P_D(Y = c)$$

$$P_D(B = s) = P_D(B = s|Y = c) \sum_{c' \in C} P_D(Y = c')$$

$$= P_D(B = s|Y = c)$$



[3] Qraitem M et al. Bias mimicking: A simple sampling approach for bias mitigation.(CVPR2023).

Fair sampling distribution

Bias Mimicking: A Simple Sampling Approach for Bias Mitigation

Maan Qraitem[1], Kate Saenko[1,2], Bryan A. Plummer[1]
[1]Boston University [2]MIT-IBM Watson AI Lab
{mqraitem,saenko,bplum}@bu.edu

How to Bias Mimick?

- constrain the solution space such that the solution retains **the most number** of samples.
- obtain the set of solutions using a linear program.

c: preserved class
c′: mimicked class
s: bias
l: count

Set of biases

$$\max \sum_s l_s^{c'}$$

$$\text{s.t.} \quad l_s^{c'} \leq |D_{c',s}| \qquad s \in S$$

$$\frac{l_s^{c'}}{\sum_s l_s^{c'}} = P_D(B = s | Y = c) \qquad s \in S$$

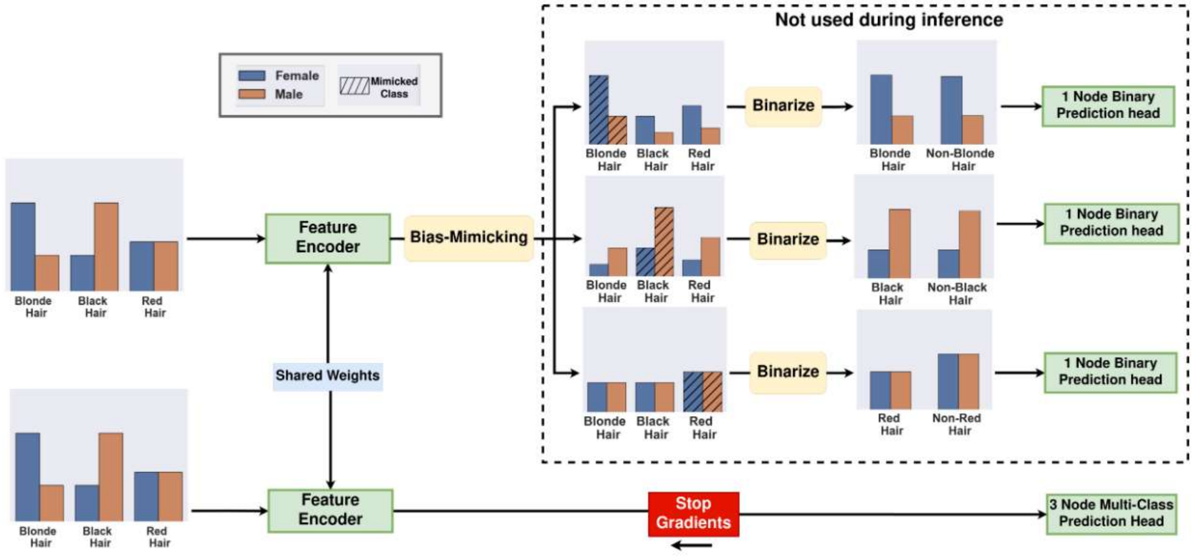
[3] Qraitem M et al. Bias mimicking: A simple sampling approach for bias mitigation.(CVPR2023).

Fair sampling distribution

Bias Mimicking: A Simple Sampling Approach for Bias Mitigation

Maan Qraitem[1], Kate Saenko[1,2], Bryan A. Plummer[1]
[1]Boston University [2]MIT-IBM Watson AI Lab
{mqraitem,saenko,bplum}@bu.edu

- How to train & inference?



✓ <u>Too many</u> additional parameters

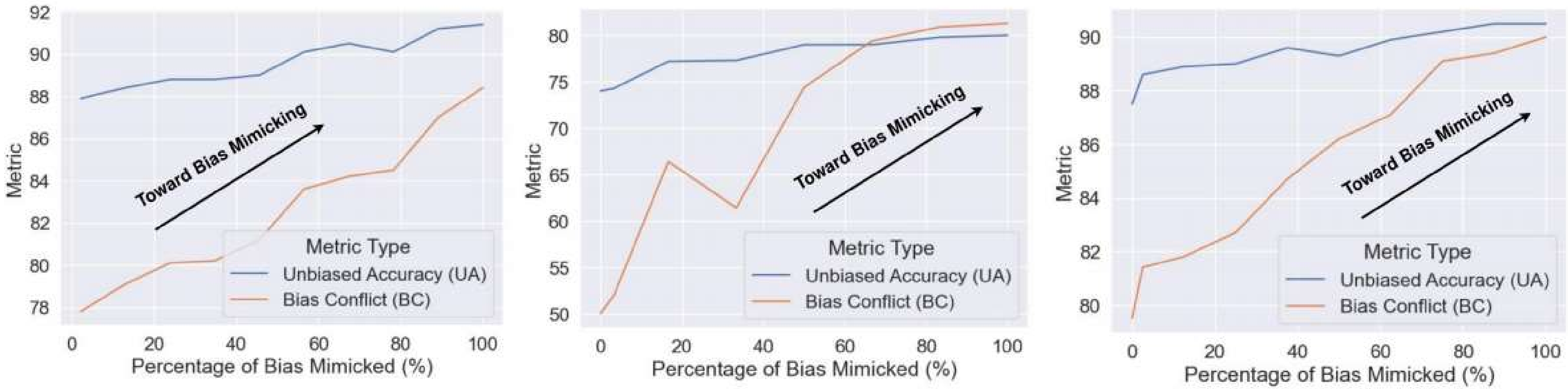
✓ The scores may <u>not be calibrated</u> with respect to each other

Results

Bias Mimicking: A Simple Sampling Approach for Bias Mitigation

Maan Qraitem[1], Kate Saenko[1,2], Bryan A. Plummer[1]
[1]Boston University [2]MIT-IBM Watson AI Lab
{mqraitem, saenko, bplum}@bu.edu

- How sensitive is the model to the mimicking condition?
  - 0%: distribution remains the same
  - 100%: Complete bias mimicking



a) CelebA/Blonde b) Utk-Face/Age a) Utk-Face/Race

The performance of the model **is highly dependent** on the degree of bias mimicry, particularly <u>the bias conflict groups</u>

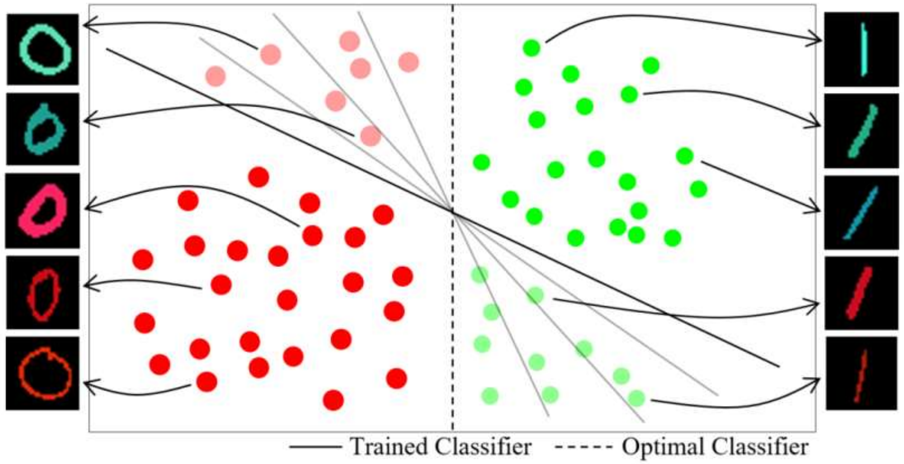
Fair network

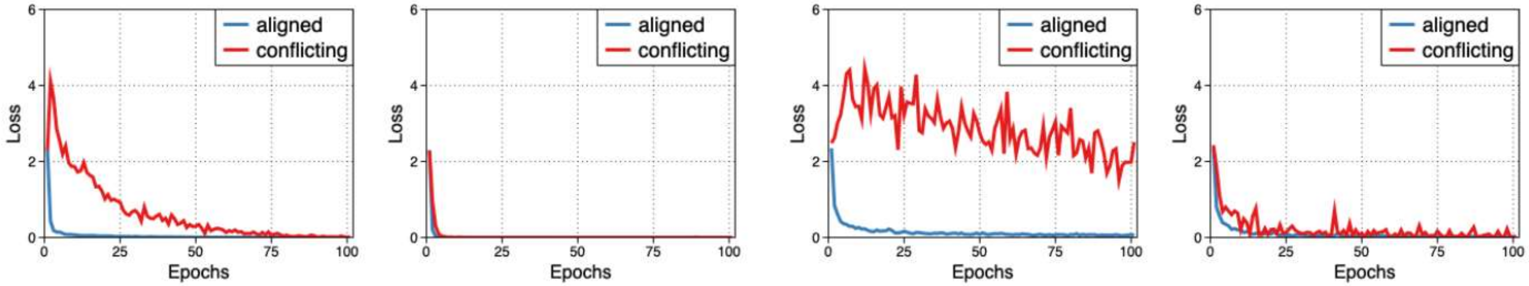
Learning Not to Learn: Training Deep Neural Networks with Biased Data

Byungju Kim[1] Hyunwoo Kim[2] Kyungsu Kim[3] Sungjin Kim[3] Junmo Kim[1]
School of Electrical Engineering, KAIST, South Korea[1]
Beijing Institute of Technology[2]
Samsung Research[3]

- How does the bias affect model training?

The **malignant bias** attributes **are easier to learn** than the original task[5]





(a) Colored MNIST, (Digit, Color) (b) Corrupted CIFAR-10[1], (Object, Corruption)

[4] Nam J et al. Learning from failure: De-biasing classifier from biased classifier.(NeurIPS2020).
[5] Kim B et al. Learning not to learn: Training deep neural networks with biased data.(CVPR2020).

Fair network

Learning Not to Learn: Training Deep Neural Networks with Biased Data

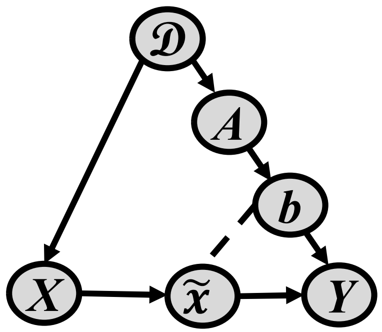
Byungju Kim[1] Hyunwoo Kim[2] Kyungsu Kim[3] Sungjin Kim[3] Junmo Kim[1]
School of Electrical Engineering, KAIST, South Korea[1]
Beijing Institute of Technology[2]
Samsung Research[3]

- How to unlearn the malignant bias?
  - **Minimize the mutual information** between feature embedding and target bias

    $\min I(\tilde{X}; B) \quad i.e. \min D_{KL}(p(\tilde{x}, b)||P(\tilde{x})P(b))$

  - malignant bias $\mathcal{I}(b(X^{train}); Y) \gg \mathcal{I}(b(X^{test}); Y) \approx 0.$
  - Ultimate optimization goal

$$\min_{\theta_f, \theta_g} \mathbb{E}_{\tilde{x} \sim P_X(\cdot)}[\mathcal{L}_c(y_{\tilde{x}}, g(f(\tilde{x})))] + \lambda \mathcal{I}(b(X); f(X))$$



- Issues: How to capture **the distribution of bias** in a network?

[5] Kim B et al. Learning not to learn: Training deep neural networks with biased data.(CVPR2020).

Fair network

Learning Not to Learn: Training Deep Neural Networks with Biased Data

Byungju Kim[1] Hyunwoo Kim[2] Kyungsu Kim[3] Sungjin Kim[3] Junmo Kim[1]
School of Electrical Engineering, KAIST, South Korea[1]
Beijing Institute of Technology[2]
Samsung Research[3]

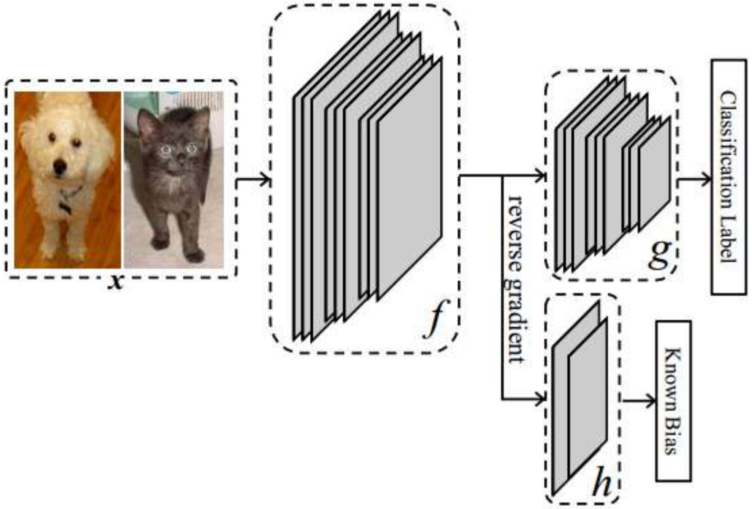
$$\min_{\theta_f,\theta_g} \boxed{\mathbb{E}_{\tilde{x}\sim P_X(\cdot)}[\mathcal{L}_c(y_{\tilde{x}}, g(f(\tilde{x})))] + \lambda\mathcal{I}(b(X); f(X))}$$

$$\mathcal{I}(b(X); f(X)) = H(b(X)) - H(b(X)|f(X)).$$

Q

$$\min_{\theta_f} \mathbb{E}_{\tilde{x}\sim P_X(\cdot)}[\mathbb{E}_{\tilde{b}\sim Q(\cdot|f(\tilde{x}))}[\log Q(\tilde{b}|f(\tilde{x}))]]$$

$$s.t.\ Q(b(X)|f(X)) = P(b(X)|f(X)).$$



$$\mu D_{KL}(P(b(X)|f(X))||Q(b(X)|f(X)))$$

$$\mathcal{L}_{\mathcal{B}}(\theta_f,\theta_h) = \mathbb{E}_{\tilde{x}\sim P_X(\cdot)}[\mathcal{L}_c(b(\tilde{x}), h(f(\tilde{x})))]$$

$$\min_{\theta_f,\theta_g} \max_{\theta_h} \mathbb{E}_{\tilde{x}\sim P_X(\cdot)}[\mathcal{L}_c(y_{\tilde{x}}, g(f(\tilde{x})))$$
$$+ \lambda\mathbb{E}_{\tilde{b}\sim Q(\cdot|f(\tilde{x}))}[\log Q(\tilde{b}|f(\tilde{x}))]]$$
$$- \mu\mathcal{L}_{\mathcal{B}}(\theta_f,\theta_h)$$

[5] Kim B et al. Learning not to learn: Training deep neural networks with biased data.(CVPR2020).

What's more? ———Uncertainty Modeling and Optimization

The data distribution is p. Optimize θ to find the model $M(\theta)$ that makes the evaluation function l optimal

- **Ignoring uncertainty**
- **Stochastic optimization (Known** about p completely**)**

  Optimize θ to maximize the expected value of l under p
- **Robust optimization (Only basic information** about p**)**

  Find the optimal solution when p is the worst case
- **Distributionally robust optimization (**Extraordinarily **known** distributional **characteristi**

  Finding the worst distribution function that satisfies the uncertain parameter features

ERM: $$\hat{\theta}_{\mathrm{ERM}} := \arg\min_{\theta\in\Theta} \mathbb{E}_{(x,y)\sim\hat{P}}[\ell(\theta;(x,y))]$$

DRO: $$\min_{\theta\in\Theta}\left\{\mathcal{R}(\theta) := \sup_{Q\in\mathcal{Q}} \mathbb{E}_{(x,y)\sim Q}[\ell(\theta;(x,y))]\right\}$$

e.g. f-divergence based

What's more? ———Uncertainty Modeling and Optimization

ERM: $$\hat{\theta}_{\text{ERM}} := \arg\min_{\theta\in\Theta} \mathbb{E}_{(x,y)\sim\hat{P}}[\ell(\theta;(x,y))]$$

DRO: $$\min_{\theta\in\Theta}\left\{\mathcal{R}(\theta) := \sup_{Q\in\mathcal{Q}} \mathbb{E}_{(x,y)\sim Q}[\ell(\theta;(x,y))]\right\}$$

Δ^m is an $(m-1)$-dimensional probabilistic simplex

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

$$\mathcal{Q} := \{\textstyle\sum_{g=1}^{m} q_g P_g : q \in \Delta_m\}$$

$$\min_{\theta\in\Theta}\sup_{q\in\Delta_m}\sum_{g=1}^{m} q_g \mathbb{E}_{(x,y)\sim P_g}[\ell(\theta;(x,y))]$$

Algorithm 1: Online optimization algorithm for group DRO

Input: Step sizes η_q, η_θ; P_g for each $g \in \mathcal{G}$

Initialize $\theta^{(0)}$ and $q^{(0)}$

for $t = 1, \ldots, T$ **do**

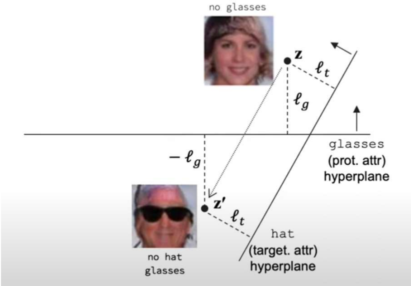
- $g \sim \text{Uniform}(1, \ldots, m)$ // Choose a group g at random
- $x, y \sim P_g$ // Sample x, y from group g
- $q' \leftarrow q^{(t-1)}$; $q'_g \leftarrow q'_g \exp(\eta_q \ell(\theta^{(t-1)}; (x,y)))$ // Update weights for group g
- $q^{(t)} \leftarrow q' / \sum_{g'} q'_{g'}$ // Renormalize q
- $\theta^{(t)} \leftarrow \theta^{(t-1)} - \eta_\theta q_g^{(t)} \nabla \ell(\theta^{(t-1)}; (x,y))$ // Use q to update θ

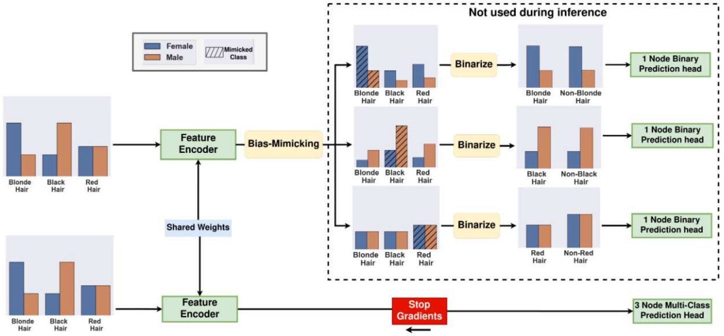
end

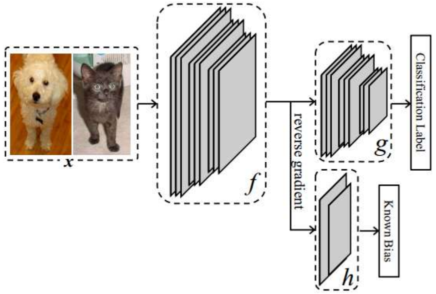
[6] Sagawa S et al. Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization.(ICLR2020).

From known to unknown

- Still need to know a priori information about the bias







- How to trick unknown bias?
  - Unsupervised learning
  - Self-supervised learning
  - MLLM-based

Aspect 1 ——Clustering

Unsupervised Learning of Debiased Representations with Pseudo-Attributes

Seonguk Seo[1] Joon-Young Lee[3] Bohyung Han[1,2]
[1]ECE & [1]ASRI & [1,2]IPAI, Seoul National University [3]Adobe Research
{seonguk, bhhan}@snu.ac.kr jolee@adobe.com

➢ **Observation**:
For a particular attribute (other than the target attribute), samples with **the same label** tend to have **similar representations** in the feature space of a fully trained model

➢ **Motivation** :
Define groups using **biased pseudo-attribute** information obtained through any **clustering algorithm** in the feature embedding space.

➢ **Optimization goals** :

$$\min_{\theta}\left\{\mathbb{R}_{\mathcal{K}}(\theta) := \mathbb{E}_{(\mathbf{x},y)\sim P}\left[\omega_{h((\mathbf{x},y);\tilde{\theta})}\ell((\mathbf{x},y);\theta)\right]\right\}$$

cluster membership

[7] Seo S, Lee J Y, Han B. Unsupervised learning of debiased representations with pseudo-attributes.(CVPR2022).

Aspect 1——Clustering

Unsupervised Learning of Debiased Representations with Pseudo-Attributes

Seonguk Seo[1] Joon-Young Lee[3] Bohyung Han[1,2]
[1]ECE & [1]ASRI & [1,2]IPAI, Seoul National University [3]Adobe Research
{seonguk, bhhan}@snu.ac.kr jolee@adobe.com

Group DRO: $$\min_{\theta \in \Theta} \sup_{q \in \Delta_m} \sum_{g=1}^{m} q_g \mathbb{E}_{(x,y)\sim P_g}[\ell(\theta; (x,y))]$$

Revised version $$\min_{\theta} \left\{ \mathbb{R}_{\mathcal{K}}(\theta) := \mathbb{E}_{(\mathbf{x},y)\sim P} \left[\omega_{h((\mathbf{x},y);\tilde{\theta})} \ell((\mathbf{x},y);\theta) \right] \right\}$$

$$\omega_k = \frac{\mathbb{E}_{(\mathbf{x},y)\sim P_k}[\ell((\mathbf{x},y);\theta)]}{N_k} = \frac{\mathbb{E}_{(\mathbf{x},y)\sim P}[\ell((\mathbf{x},y);\theta) \mid h((\mathbf{x},y);\tilde{\theta}) = k]}{\sum_i \mathbb{1}(h((\mathbf{x}_i,y_i);\tilde{\theta}) = k)}$$

Step1： **Clustering** the training samples in the feature embedding space of the fully optimized base model, assuming that each cluster corresponds to a bias pseudo-attribute

Step2： **Weighting** between groups considering the size and average difficulty of each cluster

Algorithm 1: Debiasing with bias pseudo-attribute

Require: step size η_θ, momentum m, training steps T, batch size B, the number of clusters K
Base model:
Initialize $\tilde{\theta}$
for $t = 1, ..., T$ **do**
  Sample $(\mathbf{x}_i, y_i) \sim P$ for $i = 1, ..., B$;
  $\tilde{\theta} \leftarrow \tilde{\theta} - \eta_\theta \sum_{i=1}^{B} \nabla \ell((\mathbf{x}_i, y_i); \tilde{\theta})$;
end
for $k = 1, ..., K$ **do**
  $P_k = \{(\mathbf{x}_n, y_n) \mid h((\mathbf{x}_n, y_n); \tilde{\theta}) = k$ for all $n\}$;
  $N_k = |P_k|$;
end
Target model:
Initialize θ and ω_k for $k = 1, ..., K$
for $t = 1, ..., T$ **do**
  $\omega_k \leftarrow (1-m)\omega_k + \frac{m}{N_k} \mathbb{E}_{(\mathbf{x},y)\sim P_k}[\ell((\mathbf{x},y);\theta)]$ for $k = 1, ..., K$;
  Sample $(\mathbf{x}_i, y_i) \sim P$ for $i = 1, ..., B$;
  $\alpha_i = \omega_{h((\mathbf{x}_i,y_i);\tilde{\theta})}$;
  $\overline{\alpha}_i = \alpha_i / \sum_{i=1}^{B} \alpha_i$;
  $\theta \leftarrow \theta - \eta_\theta \sum_{i=1}^{B} \overline{\alpha}_i \nabla \ell((\mathbf{x}_i, y_i); \theta)$;
end

[7] Seo S, Lee J Y, Han B. Unsupervised learning of debiased representations with pseudo-attributes.(CVPR2022).

Aspect 2—Contrastive Approach

Correct-N-Contrast: A Contrastive Approach for Improving Robustness to Spurious Correlations

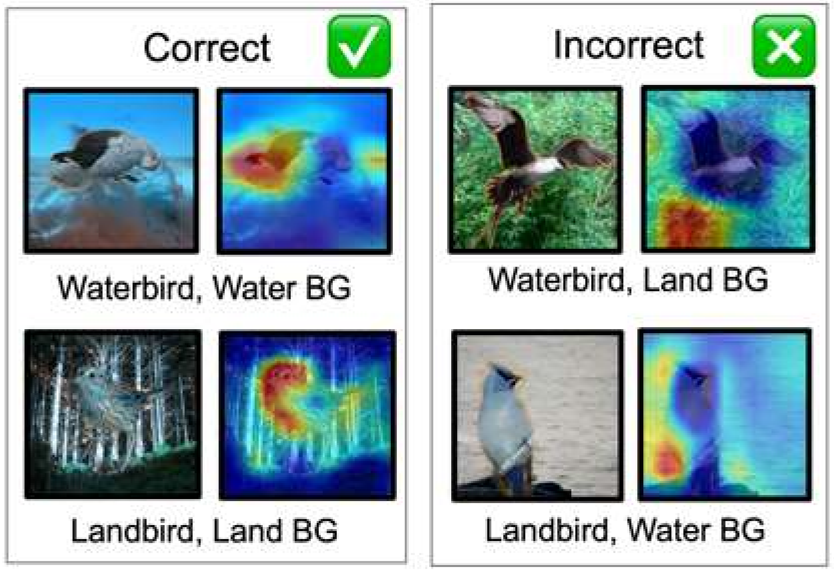
Michael Zhang†, Nimit S. Sohoni†, Hongyang R. Zhang‡, Chelsea Finn† & Christopher Ré†
†Stanford University, ‡Northeastern University
{mzhang,nims,hongyang,cbfinn,chrismre}@cs.stanford.edu
Corresponding author

➢ **Observation**:
The **worst population accuracy** of the neural network in the false correlation case is closely related to its representation - i.e., the output of its last hidden layer - only to **the extent that it relies on true labels** rather than false attributes

➢ **Motivation** :
By **improving alignment** while keeping the class mean error low, we can help improve the worst group error for the class

➢ **New sampling scheme**

Samples <u>with the same category but different spurious attributes</u> as **different "views"** (anchors and **positive samples**) of the same category, and samples **negative samples** of data points with the <u>same inferred spurious attributes</u> but <u>different categories</u>



[8] Zhang M et al. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations.(CVPR2023).

Aspect 2—Contrastive Approach

Correct-N-Contrast: A Contrastive Approach for Improving Robustness to Spurious Correlations

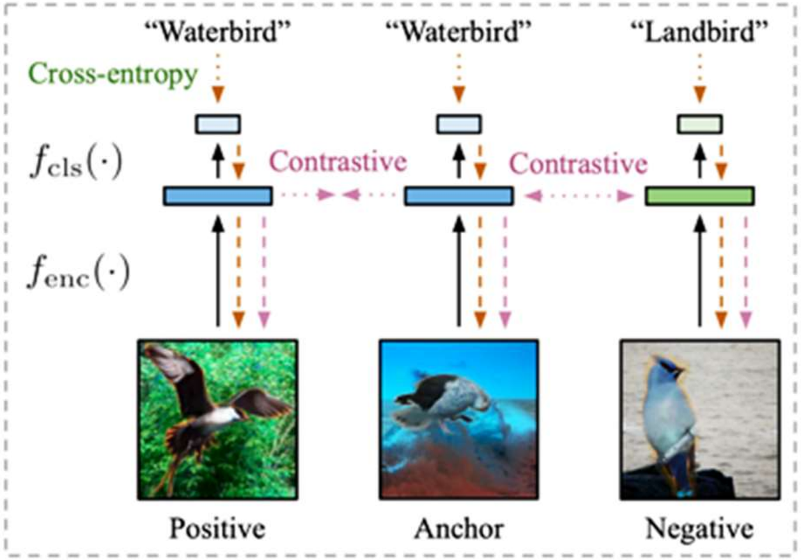
Michael Zhang†, Nimit S. Sohoni†, Hongyang R. Zhang‡, Chelsea Finn† & Christopher Ré†
†Stanford University, ‡Northeastern University
{mzhang,nims,hongyang,cbfinn,chrismre}@cs.stanford.edu
Corresponding author

Stage 1: ERM Training

- Train model with ERM to predict ground-truth labels from data
- Collect predictions of the trained ERM model on the *training data*
- Get contrastive batches from predictions:

Stage 2: Supervised Contrastive Learning

- Train a new model with supervised contrastive loss + classification loss using contrastive batches



$$\hat{\mathcal{L}}(f_\theta; x, y) = \lambda \hat{\mathcal{L}}^{\text{sup}}_{\text{con}}(f_{\text{enc}}; x, y) + (1-\lambda)\hat{\mathcal{L}}_{\text{cross}}(f_\theta; x, y)$$

$$\hat{\mathcal{L}}^{\text{sup}}_{\text{con}}(f_{\text{enc}}) = \boxed{\hat{\mathcal{L}}^{\text{sup}}_{\text{con}}}\left(x_1, \{x_m^+\}_{m=1}^M, \{x_n^-\}_{n=1}^N; f_{\text{enc}}\right) + \hat{\mathcal{L}}^{\text{sup}}_{\text{con}}\left(x_1^+, \{x_i\}_{i=1}^M, \{x_n'^-\}_{n=1}^N; f_{\text{enc}}\right)$$

$$-\frac{1}{M}\sum_{m=1}^{M} \log \frac{\exp(z_1^\top z_m^+/\tau)}{\sum_{m=1}^{M}\exp(z_1^\top z_m^+/\tau) + \sum_{n=1}^{N}\exp(z_1^\top z_n^+/\tau)}$$

[8] Zhang M et al. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations.(CVPR2023).

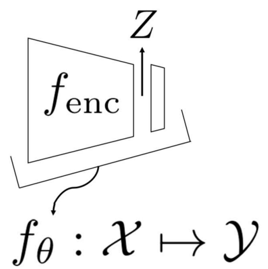
Aspect 2—Theoretical proof

Correct-N-Contrast: A Contrastive Approach for Improving Robustness to Spurious Correlations

Michael Zhang[†], Nimit S. Sohoni[†], Hongyang R. Zhang[‡], Chelsea Finn[†] & Christopher Ré[†]
[†]Stanford University, [‡]Northeastern University
{mzhang,nims,hongyang,cbfinn,chrismre}@cs.stanford.edu
Corresponding author

- **Motivation**:

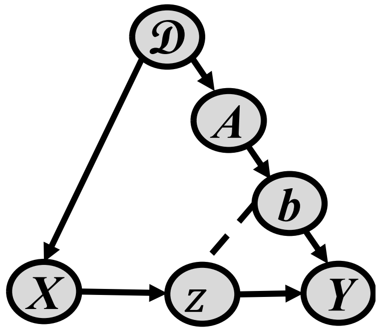
  Reducing alignment losses can **narrow the gap** between the worst and average group losses



- **Metrics design**:
  - Alignment loss

$$\hat{\mathcal{L}}_{\text{align}}(f_{\text{enc}}; g, g') = \frac{1}{|G|}\frac{1}{|G'|}\sum_{(x,y,a)\in G}\sum_{(x',y,a')\in G'} \|f_{\text{enc}}(x) - f_{\text{enc}}(x')\|_2$$

Degree to which samples with the same class but different spurious attributes map to neighborhood vectors



  - Mutual information

$$\hat{I}(Y; Z) = \frac{1}{|Z|}\sum_{z\in Z}\sum_{y\in Y} p(y \mid z)\log\frac{p(y \mid z)}{p(y)}$$

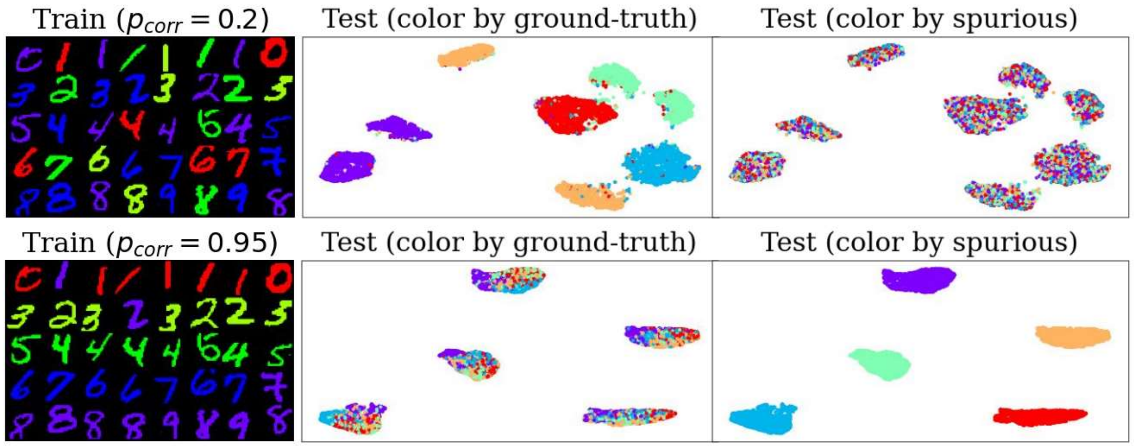
$$D_{KL}\{P(y|z)||P(y)\}$$

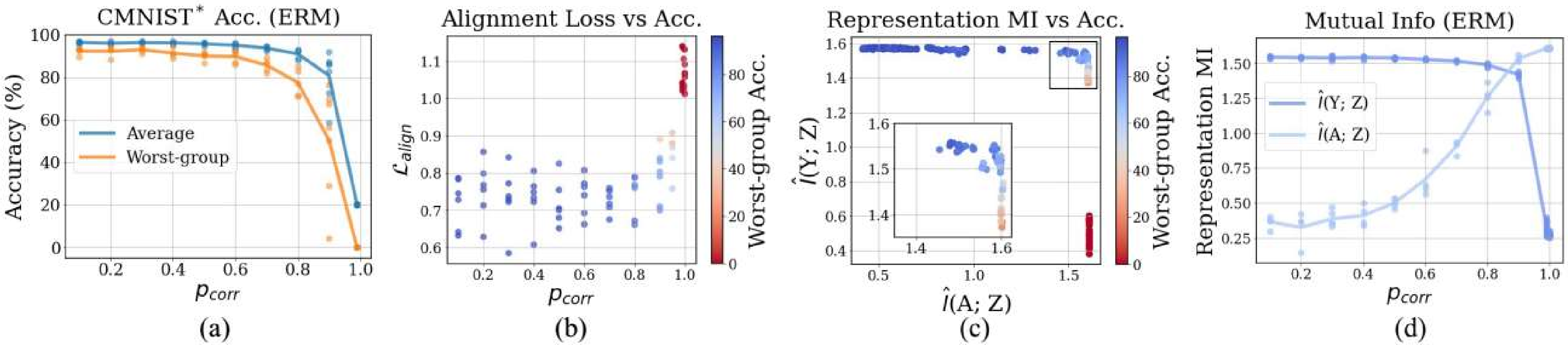
Approximate MI between model-learned representations and labels

Aspect 2—Theoretical proof

Correct-N-Contrast: A Contrastive Approach for Improving Robustness to Spurious Correlations

Michael Zhang[†], Nimit S. Sohoni[†], Hongyang R. Zhang[‡], Chelsea Finn[†] & Christopher Ré[†]
[†]Stanford University, [‡]Northeastern University
{mzhang,nims,hongyang,cbfinn,chrismre}@cs.stanford.edu
Corresponding author





[8] Zhang M et al. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations.(CVPR2023).

Aspect 2—Theoretical proof

Let $\mathcal{L}_{\text{wg}}(f_\theta; y)$ be the worst-group loss among groups in $\mathcal{G}_y$:

$$\mathcal{L}_{\text{wg}}(f_\theta; y) := \max_{g \in \mathcal{G}_y} \mathop{\mathbb{E}}_{(x,\tilde{y},a)\sim P_g} \left[\ell(f_\theta(x), \tilde{y})\right].$$

Let $\mathcal{L}_{\text{avg}}(f_\theta; y)$ be the average loss among groups in $\mathcal{G}_y$:

$$\mathcal{L}_{\text{avg}}(f_\theta; y) := \mathop{\mathbb{E}}_{(x,\tilde{y},a)\sim P:\forall a \in \mathcal{A}} \left[\ell(f_\theta(x), \tilde{y})\right].$$

Additionally, let $\mathcal{L}_{\text{align}}(f_\theta; y)$ be the largest cross-group alignment loss among groups in $\mathcal{G}_y$:

$$\hat{\mathcal{L}}_{\text{align}}(f_\theta; y) := \max_{g \in \mathcal{G}_y, g' \in \mathcal{G}_y:\, g \neq g'} \hat{\mathcal{L}}_{\text{align}}(f_{\text{enc}}; g, g').$$

Correct-N-Contrast: A Contrastive Approach for Improving Robustness to Spurious Correlations

Michael Zhang[†], Nimit S. Sohoni[†], Hongyang R. Zhang[‡], Chelsea Finn[†] & Christopher Ré[†]
[†]Stanford University, [‡]Northeastern University
{mzhang,nims,hongyang,cbfinn,chrismre}@cs.stanford.edu
Corresponding author

Theorem 3.1. *In the setting described above, suppose the weight matrix of the linear classification layer W satisfies $\|W\|_2 \leq B$, for some $B > 0$. Suppose the loss function $\ell(x, y)$ is C_1-Lipschitz in x and bounded from above by C_2, for some $C_1 >$ and $C_2 > 0$. Let n_g be the size of any group $g \in \mathcal{G}$ in the training set. Then, for any $\delta > 0$, with probability $1 - \delta$, the following holds for any $y \in \mathcal{Y}$:*

$$\mathcal{L}_{wg}(f_\theta; y) - \mathcal{L}_{avg}(f_\theta; y) \leq BC_1 \cdot \hat{\mathcal{L}}_{align}(f_\theta; y) + \max_{g \in \mathcal{G}_y} C_2 \sqrt{8 \log(|\mathcal{G}_y|/\delta)/n_g}.$$

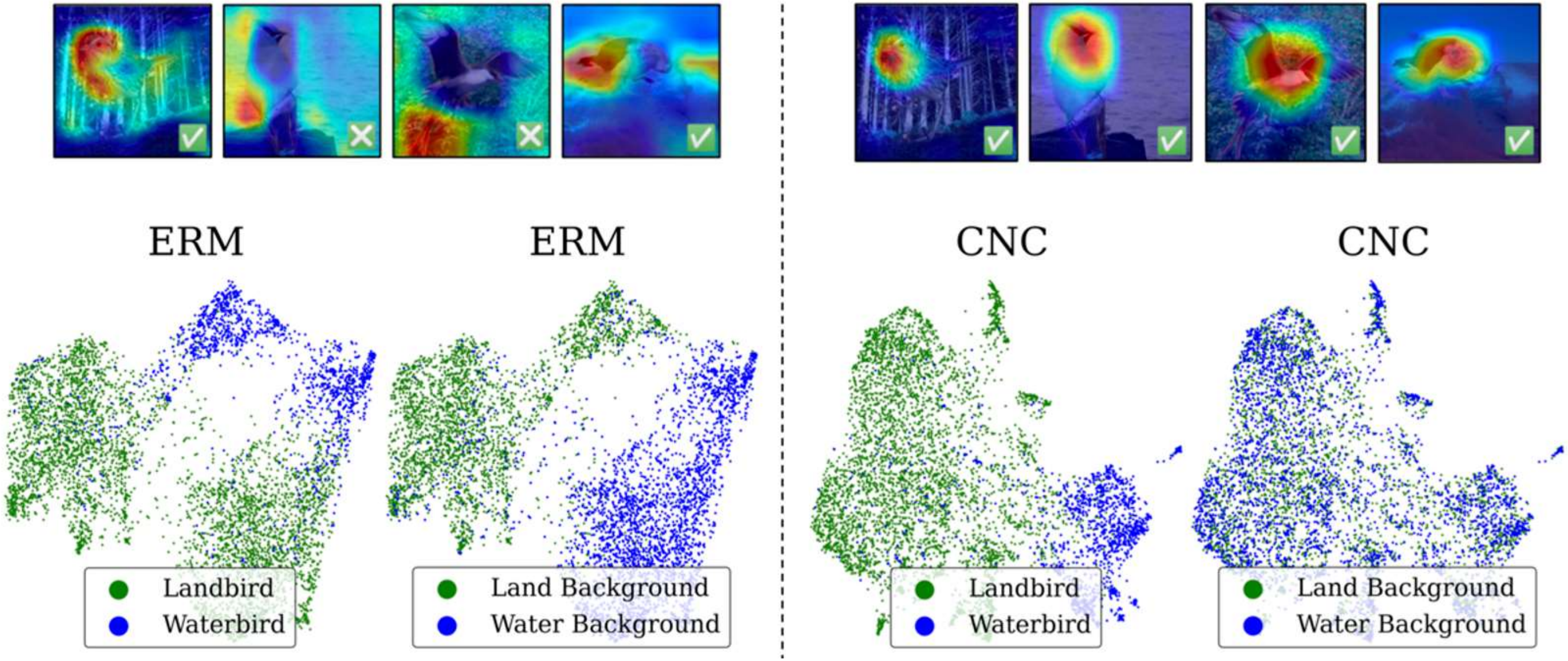
The **upper bound** of the loss gap between the worst and average groups is **linearly and positively** correlated with the largest cross-group alignment loss

[8] Zhang M et al. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations.(CVPR2023).

Aspect 2—Results

Correct-N-Contrast: A Contrastive Approach for Improving Robustness to Spurious Correlations

Michael Zhang[†], Nimit S. Sohoni[†], Hongyang R. Zhang[‡], Chelsea Finn[†] & Christopher Ré[†]
[†]Stanford University, [‡]Northeastern University
{mzhang,nims,hongyang,cbfinn,chrismre}@cs.stanford.edu
Corresponding author



[8] Zhang M et al. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations.(CVPR2023).

Discovering and Mitigating Visual Biases through Keyword Explanation

Younghyun Kim*[1] Sangwoo Mo*[2] Minkyu Kim[3] Kyungmin Lee[1] Jaeho Lee[4] Jinwoo Shin
[1]KAIST [2]University of Michigan [3]KRAFTON [4]POSTECH
younghyun.kim@kaist.ac.kr swmo@umich.edu

Aspect 3—Keyword Explanation

➢ **Motivation**: Unknown visual bias can't be interpreted

- Visualized spurious features that are **not human-readable**
- Thus, they are hard to be directly utilized for debiasing

(a) **class:** band aid, **spurious feature:** fingers, **-41.54%** (b) **class:** space bar, **spurious feature:** keys, **-46.15%** (c) **class:** plate, **spurious feature:** food, **-32.31%** (d) **class:** butterfly, **spurious feature:** flowers, **-21.54%** (e) **class:** potter's wheel, **spurious feature:** vase, **-21.54%**

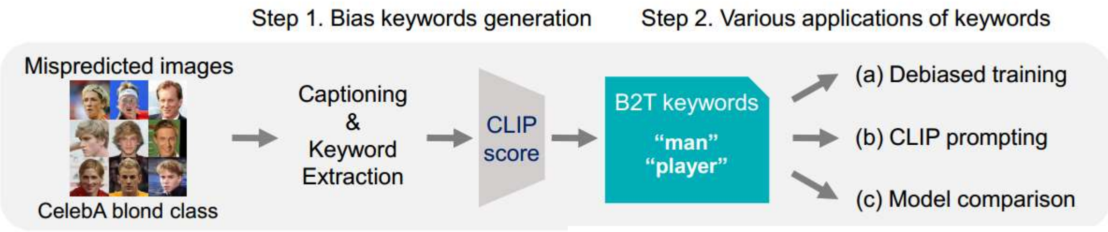
[9] Kim Y et al. Discovering and Mitigating Visual Biases through Keyword Explanation.(CVPR2024).

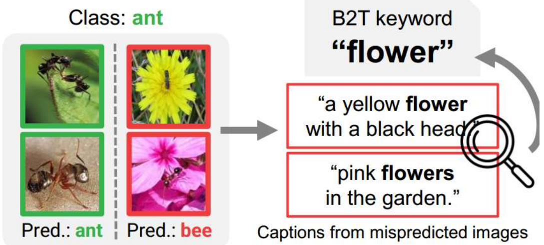
Aspect 3—Keyword Explanation

Discovering and Mitigating Visual Biases through Keyword Explanation

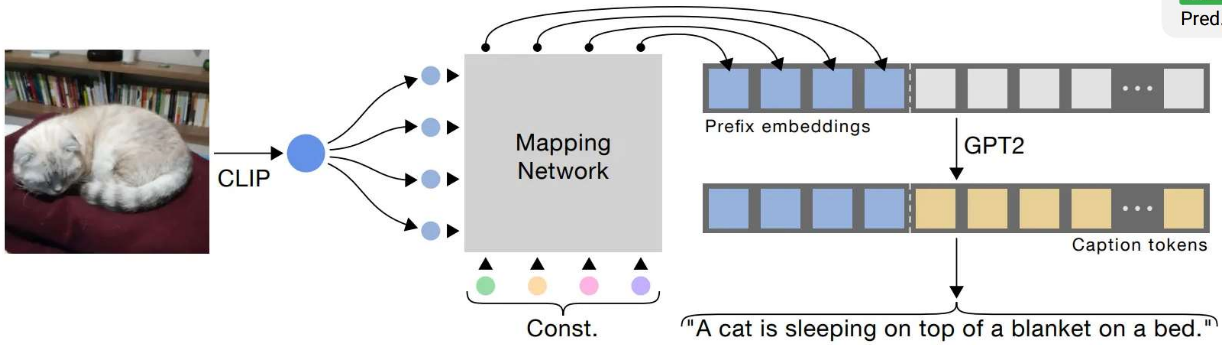
Younghyun Kim*[1] Sangwoo Mo*[2] Minkyu Kim[3] Kyungmin Lee[1] Jaeho Lee[4] Jinwoo Shin
[1]KAIST [2]University of Michigan [3]KRAFTON [4]POSTECH
younghyun.kim@kaist.ac.kr swmo@umich.edu

- **Method**: <u>B2T: Bias-to-text</u>





Step1: use ClipCap as our default **captioning model**



Step2: apply the YAKE algorithm to **extract keywords**

Text Preprocessing (Segmentation) --> Feature Extraction --> Individual Word Weight Calculation --> Candidate Keyword Generation

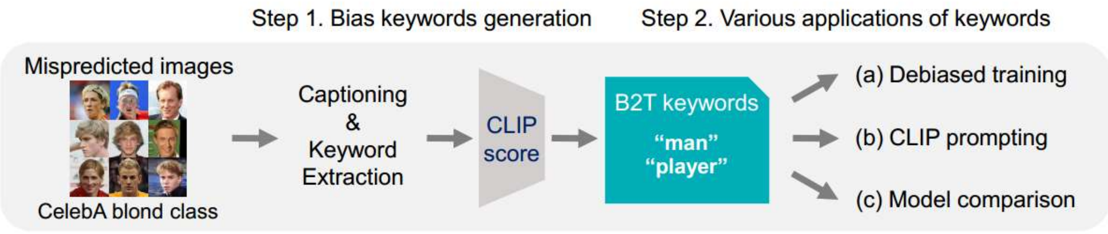
[9] Kim Y et al. Discovering and Mitigating Visual Biases through Keyword Explanation.(CVPR2024).

Aspect 3—Keyword Explanation

Discovering and Mitigating Visual Biases through Keyword Explanation

Younghyun Kim*[1] Sangwoo Mo*[2] Minkyu Kim[3] Kyungmin Lee[1] Jaeho Lee[4] Jinwoo Shin
[1]KAIST [2]University of Michigan [3]KRAFTON [4]POSTECH
younghyun.kim@kaist.ac.kr swmo@umich.edu

- **Method**: <u>B2T: Bias-to-text</u>

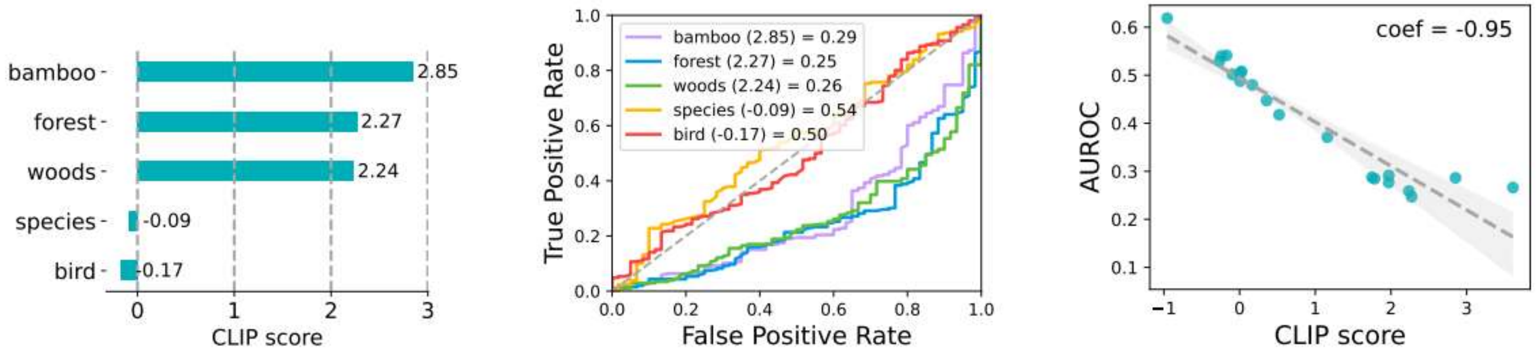


Step3: **verify** that keywords represent bias by CLIP score

To **measures the similarity** between the <u>keywords</u> and the <u>incorrectly predicted images</u>

$$s_{\text{CLIP}}(a;\mathcal{D}) := \text{sim}(a, \mathcal{D}_{\text{wrong}}) - \text{sim}(a, \mathcal{D}_{\text{correct}})$$

$$\boxed{\text{sim}(a, \mathcal{D}) := \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} f_{\text{image}}(x)\, f_{\text{text}}(a).}$$



(a) CLIP score (b) ROC curve of subgroup accuracy (c) Correlation of CLIP score and AUROC

Effect of the CLIP score (waterbird class)

[9] Kim Y et al. Discovering and Mitigating Visual Biases through Keyword Explanation.(CVPR2024).

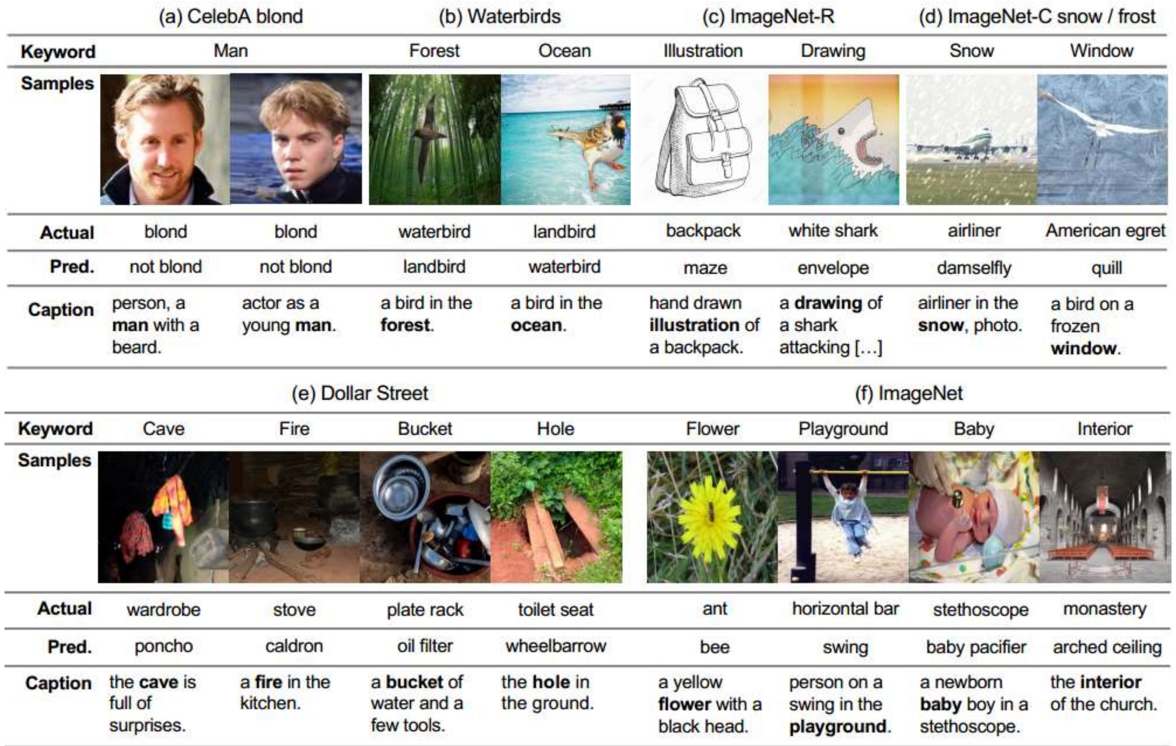
Aspect 3—Keyword Explanation

Discovering and Mitigating Visual Biases through Keyword Explanation

Younghyun Kim*[1] Sangwoo Mo*[2] Minkyu Kim[3] Kyungmin Lee[1] Jaeho Lee[4] Jinwoo Shin
[1]KAIST [2]University of Michigan [3]KRAFTON [4]POSTECH
younghyun.kim@kaist.ac.kr swmo@umich.edu

We first extract B2T keywords, then use them to various applications:

- ➤ **Debiased training**
- ➤ Model comparison
- ➤ CLIP prompting
- ➤ Label diagnosis



- B2T discovers spurious correlations and distributions shifts

- e.g.) "man" for CelebA blond、"forest" and "ocean" for Waterbirds、"illustration" and "drawing" for IN-R、"snow" and "window" for IN-C

[9] Kim Y et al. Discovering and Mitigating Visual Biases through Keyword Explanation.(CVPR2024).

Aspect 3—Keyword Explanation

Discovering and Mitigating Visual Biases through Keyword Explanation

Younghyun Kim*[1] Sangwoo Mo*[2] Minkyu Kim[3] Kyungmin Lee[1] Jaeho Lee[4] Jinwoo Shin
[1]KAIST [2]University of Michigan [3]KRAFTON [4]POSTECH
younghyun.kim@kaist.ac.kr swmo@umich.edu

We first extract B2T keywords, then use them to various applications:

- Debiased training
- Model comparison
- **CLIP prompting**
- Label diagnosis

Table 8. Prompt designs for debiaisng zero-shot classifiers.

Dataset	Dataset-wise Template	Class Name
CelebA	• [class name] • [class name] man • [class name] player • [class name] person • [class name] artist • [class name] comedy • [class name] film • [class name] actor • [class name] face	1. Blond • blond hair • celebrity of blond hair 2. Non blond • non blond hair • celebrity of non blond hair
Waterbirds	• [class name] • [class name] on the forest • [class name] with woods • [class name] on a tree • [class name] on a branch • [class name] in the forest • [class name] on the tree • [class name] on the ocean • [class name] on a beach • [class name] on the lake • [class name] with a surfer • [class name] on the water • [class name] on a boat • [class name] on the dock • [class name] on the rocks • [class name] in the sunset • [class name] with a kite • [class name] on the sky • [class name] is on flight • [class name] is on flies	1. Landbird • landbird 2. Waterbird • waterbird

- **Modify** the cue by adding a keyword, e.g., “[class]'s photo” in [group], where the keyword represents the name of the group

- **Obtaining** the average prompts embedding for a class in all groups
- **Comparing** broader class embeddings for image classification

[9] Kim Y et al. Discovering and Mitigating Visual Biases through Keyword Explanation.(CVPR2024).

Aspect 3—Keyword Explanation

Discovering and Mitigating Visual Biases through Keyword Explanation

Younghyun Kim*[1] Sangwoo Mo*[2] Minkyu Kim[3] Kyungmin Lee[1] Jaeho Lee[4] Jinwoo Shin
[1]KAIST [2]University of Michigan [3]KRAFTON [4]POSTECH
younghyun.kim@kaist.ac.kr swmo@umich.edu

We first extract B2T keywords, then use them to various applications:

- Debiased training
- CLIP prompting
- **Model comparison**
- Label diagnosis

Keyword	Work		Supermarket	
Samples				
ViT-B	O	O	O	O
RN50	O	X	O	X
Actual (RN50)	dumbbell	dumbbell	shopping basket	shopping basket
Pred (RN50)	dumbbell	horizontal bar	shopping basket	grocery store
Caption	a set of dumbbells with weights.	person **works** out in the gym.	a basket full of food.	woman shopping in a **supermarket**.

- Bias keywords can be used to <u>analyze and compare different classifiers</u> based on their keywords

e.g.) architecture: ResNet vs. ViT

[9] Kim Y et al. Discovering and Mitigating Visual Biases through Keyword Explanation.(CVPR2024).

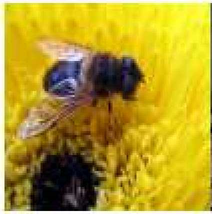
Discovering and Mitigating Visual Biases through Keyword Explanation

Younghyun Kim*[1] Sangwoo Mo*[2] Minkyu Kim[3] Kyungmin Lee[1] Jaeho Lee[4] Jinwoo Shin
[1]KAIST [2]University of Michigan [3]KRAFTON [4]POSTECH
younghyun.kim@kaist.ac.kr swmo@umich.edu

Aspect 3—Keyword Explanation

We first extract B2T keywords, then use them to various applications:

- Debiased training
- CLIP prompting
- Model comparison
- **Label diagnosis**

Keyword	Bee	Boar	Desk	Market
Samples				
Label	fly	pig	computer mouse	custard apple
Pred.	bee	wild boar	desktop computer	grocery store
Caption	a **bee** on a yellow flower.	wild **boar** in the forest.	the **desk** in the office.	fruit and vegetables at the **market**.

- B2T can diagnose common labeling errors, such as mislabeling and label ambiguities

[9] Kim Y et al. Discovering and Mitigating Visual Biases through Keyword Explanation.(CVPR2024).

THANKS