南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

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]
[1]KAIST [2]University of Michigan [3]KRAFTON [4]POSTECH
younghyun.kim@kaist.ac.kr swmo@umich.edu

CVPR 2024

Background

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

- Previous research has attempted to identify visual biases by analyzing problematic samples or problematic attributes . However, these methods define biases indirectly, often relying on visualization or sample groups with specific statistics, and they require human supervision to express them in an explainable form.

- To address this issue, recent research aimed at interpreting biases using vision-language models .Nonetheless, these studies have limitations in discovering and mitigating novel biases.

- Some studies retrieve the closest word from a predefined vocabulary, limiting their discovery to known biases.

- Others analyze neurons or images synthesized by generative models to comprehend biases. However, they focus on generating detailed captions explaining activated neurons or failure examples, which can help understand individual cases but hard to utilize for debiasing.

Methods

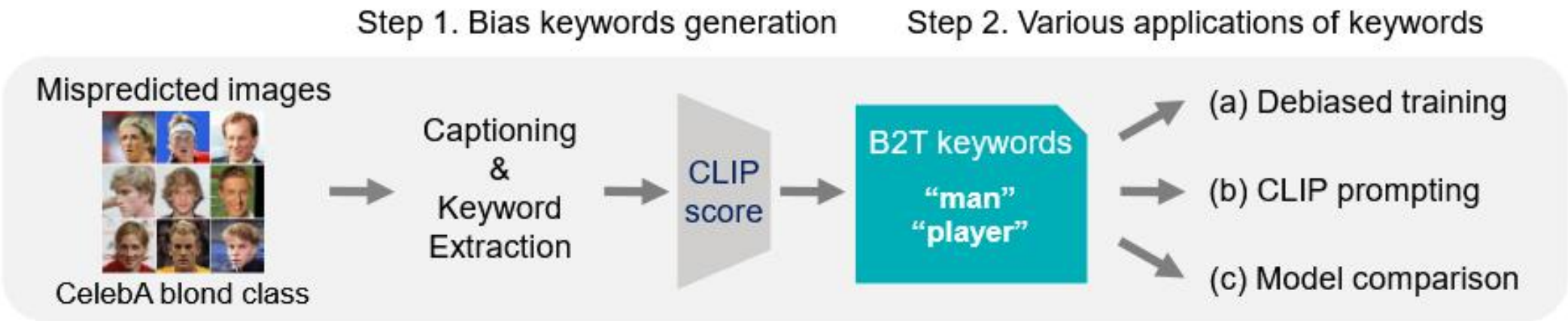


Figure 2. **Method.** (Step 1) B2T generates language descriptions from mispredicted images and extracts common keywords. We then verify whether these keywords indicate bias by measuring their similarity to the mispredicted images using a vision-language model like CLIP [59]. (Step 2) The discovered keywords have various applications, including debiased training, CLIP prompting, and model comparison.

1、Bias keywords

Our core idea is to **extract keywords that represent biases**. To achieve this, we extract common keywords from the language descriptions of class-wise mispredicted images.

Table 4. **Ablation on different captioning models.** B2T keywords discovered by different captioning models. We report the average inference time to extract a caption from a single image (in seconds on an RTX 3090 GPU) alongside the model names. The models consistently capture highly biased keywords such as "man," "forest," and "beach," while different models may find diverse fine-grained keywords such as "rainforest" or "lake."

		ClipCap	BLIP	CoCa	BLIP-2	LLaVA
	Inference time	0.13 sec	0.20 sec	0.34 sec	0.56 sec	1.90 sec
CelebA blond	man	O	O	O	O	O
Waterbird	forest	O	O	O	O	O
	bamboo	O	O	O	O	O
	woods	O	-	-	O	-
	rainforest	O	-	-	-	-
Landbird	beach	O	O	O	O	O
	ocean	-	O	O	O	-
	boat	-	O	O	O	O
	lake	O	-	-	-	-

2、CLIP score

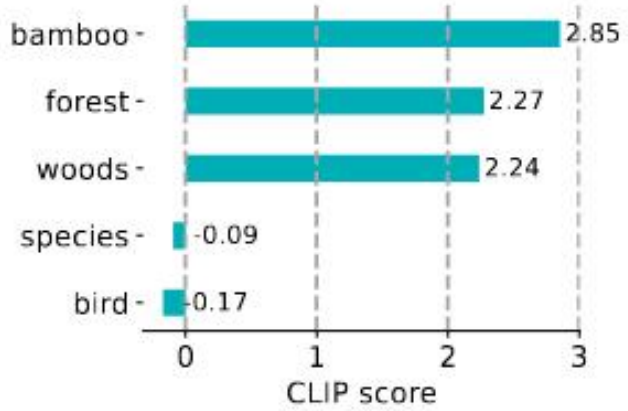
We **validate whether the keywords represent bias**. To do this, we use a vision-language scoring model like CLIP that measures the similarities between keywords and the mispredicted images.

$$s_{\text{CLIP}}(a;\mathcal{D}) := \text{sim}(a, \mathcal{D}_{\text{wrong}}) - \text{sim}(a, \mathcal{D}_{\text{correct}}). \quad (1)$$

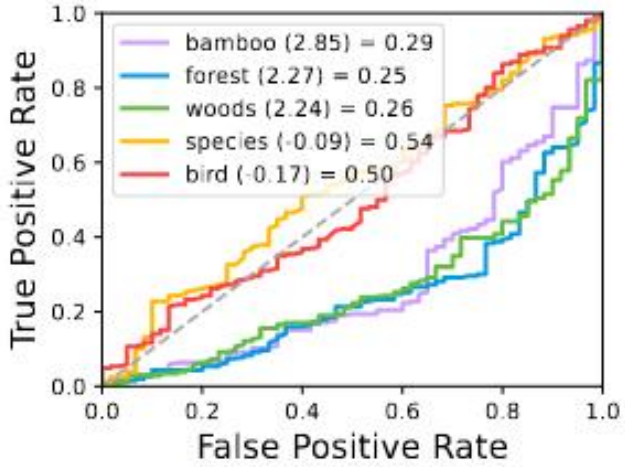
Here, $\text{sim}(a, \mathcal{D})$ is the similarity between the keyword a and the dataset $\mathcal{D}$, computed as the average cosine similarity between normalized embeddings of a word $f_{\text{text}}(a)$ and images $f_{\text{image}}(x)$ for $x \in \mathcal{D}$, where

$$\text{sim}(a, \mathcal{D}) := \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} f_{\text{image}}(x)\, f_{\text{text}}(a). \quad (2)$$

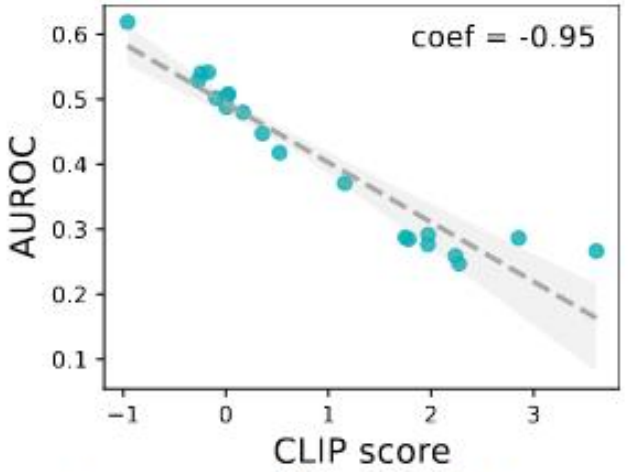
3、Validation of the CLIP score



(a) CLIP score



(b) ROC curve of subgroup accuracy



(c) Correlation of CLIP score and AUROC

Figure 3. **Effect of the CLIP score (waterbird class).** (a) The CLIP score can identify incorrect bias keywords, showing low CLIP scores near zero for non-bias keywords like "species." (b) The ROC curve represents subgroup accuracy, which defines the subgroup based on images with high CLIP similarity to specific keywords while varying the thresholds. The legend displays the B2T keywords alongside their corresponding CLIP scores in parentheses, with the AUROC of their respective curves denoted after the equal sign. Keywords with high CLIP scores tend to exhibit low subgroup accuracies, indicating they are biases. (c) Colored dots illustrate the negative correlation between the CLIP score and AUROC of subgroup accuracy over B2T keywords, indicating that a higher CLIP score implies stronger bias.

Experiments

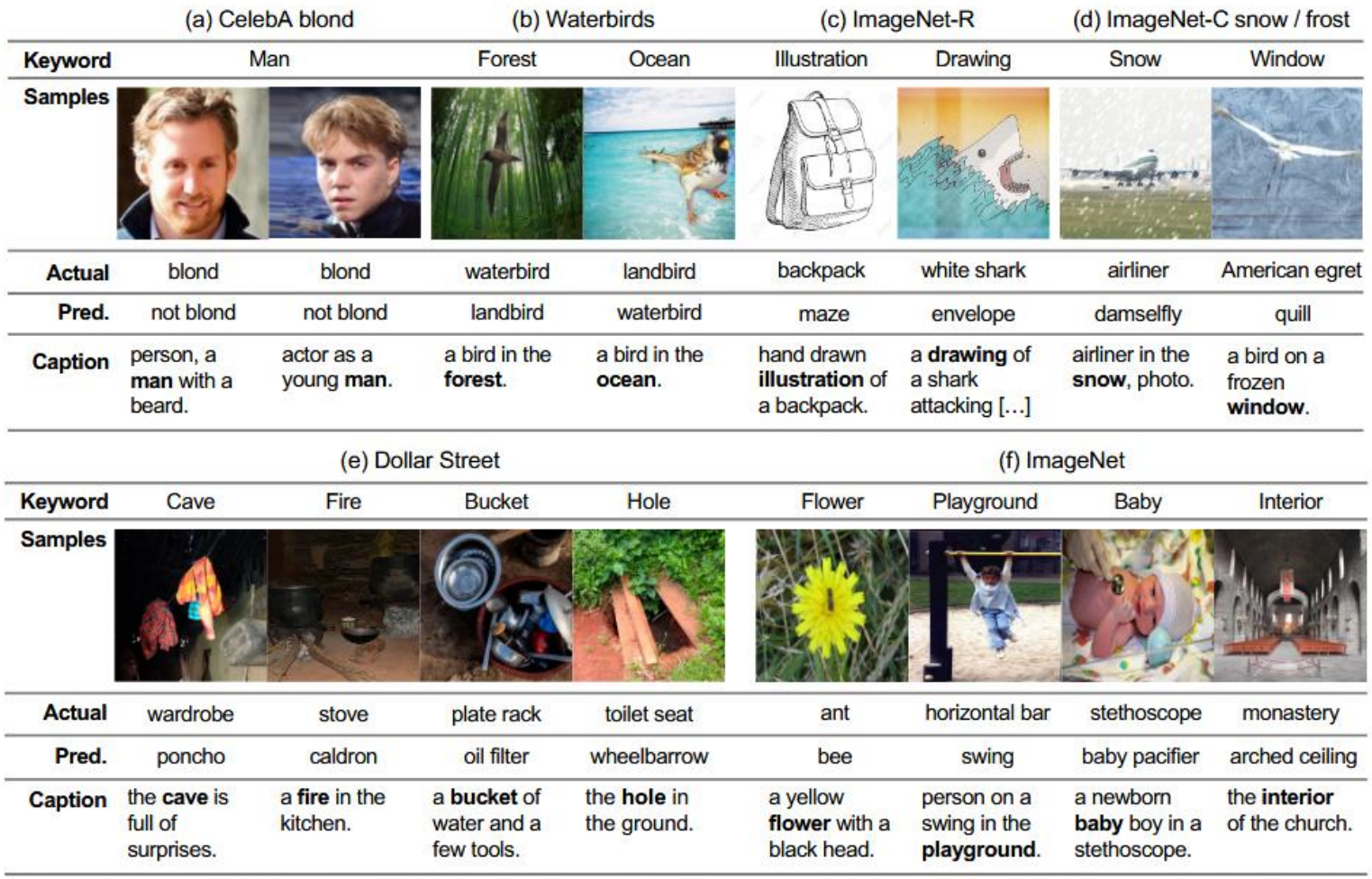


Figure 4. **Discovered biases in image classifiers.** Visual examples of mispredicted images, along with their corresponding bias keywords, captions, actual classes, and predicted classes. B2T successfully identified known biases, such as (a) gender bias in CelebA blond, (b) background bias in Waterbirds, and distribution shifts in (c) ImageNet-R with different styles, and (d) ImageNet-C with natural corruptions. B2T also uncovered novel biases in larger datasets, such as the spurious correlations between (e) the keyword "cave" and the wardrobe class, indicating geographical bias in Dollar Street, and (f) the keyword "flower" and the ant class, indicating contextual bias in ImageNet.

Applications

①Debiased DRO training

DRO 是 Distributionally Robust Optimization（分布鲁棒优化）的缩写 。在训练模型时，训练数据和测试数据的分布可能存在差异（如域适应场景），传统的经验风险最小化（ERM）方法只关注在训练数据上的性能，可能导致在测试数据上表现不佳。而 DRO 方法会考虑训练数据分布周围可能的分布情况，通过在一个分布集合上进行优化，使得模型具有更好的泛化能力，能够应对测试数据分布的变化。

Table 1. **Debiased DRO training.** Worst-group and average accuracies (%) of our debiased classifier (DRO-B2T) and prior works. GT denotes the usage of ground-truth bias labels for training, and bold denotes the best worst-group accuracy. B2T keywords enable accurate bias label prediction, facilitating effective DRO training.

		CelebA blond		Waterbirds	
Method	GT	Worst	Avg.	Worst	Avg.
ERM	-	47.7±2.1	94.9	62.6±0.3	97.3
LfF [55]	-	77.2	85.1	78.0	91.2
GEORGE [74]	-	54.9±1.9	94.6	76.2±2.0	95.7
JTT [44]	-	81.5±1.7	88.1	83.8±1.2	89.3
CNC [86]	-	88.8±0.9	89.9	88.5±0.3	90.9
DRO-B2T (ours)	-	**90.4**±0.9	93.2	**90.7**±0.3	92.1
DRO [66]	✓	90.0±1.5	93.3	89.9±1.3	91.5

Applications

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

②CLIP zero-shot prompting

Table 2. **CLIP zero-shot prompting.** Worst-group and average accuracies (%) of the CLIP zero-shot classifier using the base prompt or augmented ones: with the base group names (group) or B2T keywords with positive (B2T-pos) or negative (B2T-neg) CLIP scores. Bold indicates the best worst-group accuracy. B2T-pos improves worst-group accuracy, while B2T-neg harms. This implies that augmenting proper keywords to the prompts enhances the debiased accuracy of CLIP zero-shot inference.

	CelebA blond		Waterbirds	
	Worst	Avg.	Worst	Avg.
CLIP zero-shot	76.2	85.2	50.3	72.7
+ Group prompt [85]	76.7	87.0	53.7	78.0
+ B2T-neg prompt	72.9	88.0	45.4	70.8
+ B2T-pos prompt (ours)	**80.0**	87.2	**61.7**	76.9

Applications

③Label 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**.

Figure 7. **Label diagnosis.** We identify labeling errors, such as mislabeling and label ambiguities, in ImageNet using bias keywords. For example, the keyword "bee" implies that the images labeled as "fly" class are actually mislabeled. On the other hand, the keyword "desk" indicates that the images contain multiple objects, including both a "computer mouse" and a "desktop computer" on the desk, making it difficult to assign the appropriate class.

Ablation Study

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**.

Figure 6. **Model comparison: ResNet vs. ViT.** We compare the predictions made by ResNet and ViT, both trained and evaluated on ImageNet. We report their predicted labels and B2T keywords from ResNet. ViT excels at understanding global contexts and handling fine-grained classes than ResNet. For example, ResNet struggles with complex images whose B2T keywords represent abstract contexts like "work out" and "supermarket."

Table 5. **Ablation on different scoring models.** B2T keywords alongside their scores using different scoring models. The models provide consistent rankings, with high scores for keywords like "man" or "bamboo forest," supporting their reliability.

		CLIP	OpenCLIP	BLIP	BLIP-2
CelebA blond	man	1.06	2.23	1.19	4.04
	player	0.35	1.30	0.74	2.67
	face	-0.28	0.44	0.49	1.46
	actress	-1.63	-2.48	-1.68	-4.25
Waterbird	bamboo forest	3.61	4.68	5.22	9.85
	woods	2.24	4.43	3.47	7.08
	bird	-0.09	0.67	-0.03	-0.70
	pond	-0.27	-0.63	-0.92	-1.69

due to limitations in their training data. Nevertheless, both models perform well in various scenarios, highlighting the practical merits of our B2T framework. Further discussions of limitations can be found in Appendix F.

Thanks