Classifier-to-Bias: Toward Unsupervised Automatic Bias Detection for Visual Classifiers

Quentin Guimard[1] Moreno D'Incà[1] Massimiliano Mancini[1] Elisa Ricci[1,2]

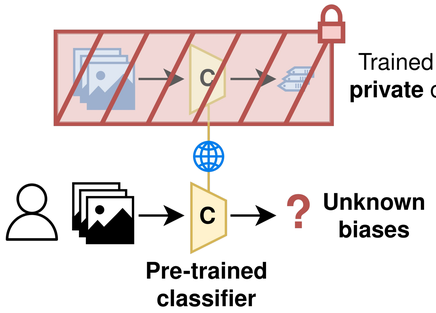
[1]University of Trento [2]Fondazione Bruno Kessler

UNIVERSITY OF TRENTO

CVPR Nashville JUNE 11-15, 2025

Motivation

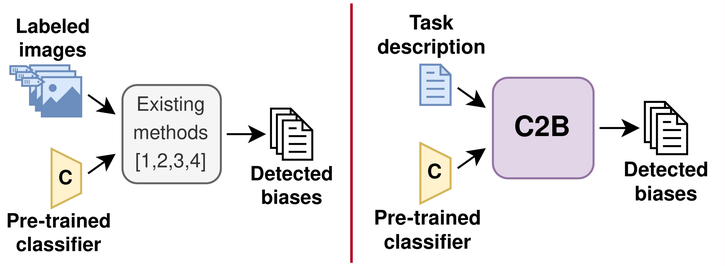
Publicly available pre-trained models may be trained on private data with **unknown biases** and **spurious correlations**.



Existing bias detection methods rely on labeled data to identify biases, which may not be available.

Setting

C2B eliminates the need for labeled data, enabling broader and more flexible bias detection.



Existing methods [1-4]: **supervised** bias detection

Classifier-to-Bias (C2B): **unsupervised** bias detection

[1] Kim et al. Discovering and mitigating visual biases through keyword explanation. CVPR 2024.
[2] Rezaei et al. PRIME: Prioritizing Interpretability in Failure Mode Extraction. ICLR 2024.
[3] Zhao et al. Language-guided Detection and Mitigation of Unknown Dataset Bias. arXiv 2024.
[4] Ciranni et al. Say My My Name: a Model's Bias Discovery Framework. arXiv 2024.

*How can we **detect biases** in pre-trained visual classifiers **without any labeled data**?*

Paper

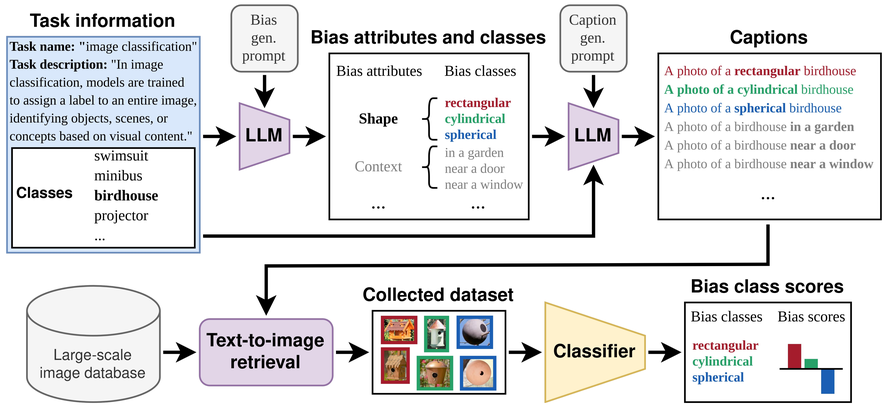
Classifier-to-bias (C2B)

Code

Method

1. Input: Task description and a pre-trained classifier.

2. Bias and caption generation: LLM generates bias attributes and classes, then creates captions that combine these biases with target classes.



3. Image retrieval: VLM-based text-to-image retrieval collects images representing each bias class.

4. Bias measurement: The classifier is tested on the retrieved images to calculate bias scores.

Evaluation

Why is C2B hard to evaluate?

Dataset coverage
Biases may not be present in the dataset

Annotation gaps
Even if present, biases may not be annotated

Representativeness
Annotated datasets may not represent real-world distribution

Ground-truth evaluation: compare detected biases with annotations

Face attribute classification on CelebA

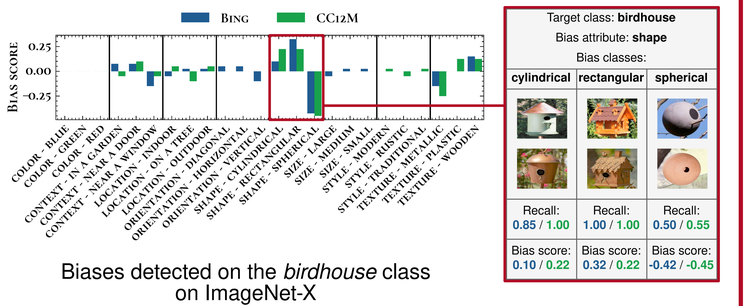
Method	GT → Detected Hit(↑)	GT → Detected FH(↓)	GT → Detected Miss (↓)	Detected → GT Hit(↑)	Detected → GT FH (↓)	Detected → GT Miss(↓)
FaceXFormer						
B2T [33]	4.53	**3.14**	92.32	6.59	**4.10**	89.29
C2B (Bing)	**12.29**	6.88	**80.83**	**14.18**	7.34	**78.48**
C2B (CC12M)	10.76	7.75	81.49	12.75	8.15	79.09

VQA-based evaluation: pseudo-label biases with a VQA model

Image classification on ImageNet-X

Model	B2T [33]	C2B (Bing)	C2B (CC12M)
ResNet50_V2	0.20	**0.26**	0.20
ResNet101_V2	0.19	**0.28**	0.20
ResNet152_V2	0.20	**0.28**	0.20
ViT_B_16_SWAG	0.20	**0.32**	0.21

Examples



Biases detected on the *birdhouse* class on ImageNet-X