Front. Comput. Sci., 2025, 19(1): 191321
https://doi.org/10.1007/s11704-024-40051-3

LETTER

Debiasing vision-language models for vision tasks: a survey

Beier ZHU (✉), Hanwang ZHANG

School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore

1 Introduction

In recent years, foundation Vision-Language Models (VLMs), such as CLIP [1], which empower zero-shot transfer to a wide variety of domains without fine-tuning, have led to a significant shift in machine learning systems. Despite the impressive capabilities, it is concerning that the VLMs are prone to inheriting biases from the uncurated datasets scraped from the Internet [2–5]. We examine these biases from three perspectives. (1) Label bias, certain classes (words) appear more frequently in the pre-training data. (2) Spurious correlation, non-target features, e.g., image background, that are correlated with labels, resulting in poor group robustness. (3) Social bias, which is a special form of spurious correlation, focuses on societal harm. Unaudited image-text pairs might contain human prejudice, e.g., gender, ethnicity, and age, that are correlated with targets. These biases are subsequently propagated to downstream tasks, leading to biased predictions.

In this survey, we provide an overview of the three biases prevalent in visual classification within the area of VLMs, along with strategies to mitigate these biases. By doing the survey, we hope to provide a useful resource for the debiasing and VLMs community.

2 Biases in VLMs

Label bias Real-world large-scale datasets often exhibit long-tailed distributions, wherein most labels are associated with only few samples. Formally, we consider a dataset in which each input $\mathbf{x} \in \mathcal{X}$ is associated with a class label $y \in \mathcal{Y}$ (the class name in the natural language can be regarded as a surrogate of class label). Label bias exists when $\mathbb{P}(y)$ is highly skewed. Naive learning on such data develops an undesirable bias towards dominant labels. Recently, label bias in vision-language models have gained significant attention [3,4,6]. They found that the predictions of VLMs are skewed toward frequent semantics. As illustrated in Fig. 1(a), a zero-shot CLIP model yields disproportionately skewed predictions on the ImageNet-1K dataset. An example of this imbalance is the over-prediction of more than 3500 instances as belonging to class 0, which is three times the actual number of samples in that class.[1)]

Spurious correlation Consider a dataset in which each input $\mathbf{x} \in \mathcal{X}$ is associated with a class label $y \in \mathcal{Y}$ and a spurious attribute $a \in \mathcal{A}$. Spurious correlation occurs when non-target attributes a such as image background that are correlated with labels y. The spurious correlation might come from many biases: (1) geography bias, data are collected from some specific locations, (2) sub-population bias, training data only cover a part of the sub-categories of the class, and (3) style bias, data have a certain style (e.g., sketches). For example, a model might inaccurately classify cows (y) on a beach, as it's trained to recognize cows in typical settings like grassy fields, focusing on background cues (a) rather than the actual animal. These spurious correlations result in *poor group robustness*. Existing researches [8,9] illustrate the presence of spurious correlation in the predictions of VLMs. Figure 1(b.2) demonstrates the worst-group (WG) and average accuracies with zero-shot CLIP on Waterbirds dataset. The groups in Waterbirds are presented in Fig. 1(b.1), where groups are based on bird type (waterbird or landbird) and their background (water or land). Zhang et al. [9] find that the zero-shot VLMs can achieve 55.6 gap between worst-group and average accuracy, indicating the existence of spurious correlations.

Social bias Social bias is a special form of spurious correlation, where the associations learned by a model reflect societal stereotypes or prejudices, e.g., gender, ethnicity and age. The web-scraped datasets are too large to be manually audited to filter out the image-text pairs with stereotypes, limiting the usefulness of VLMs in real-world high-stakes scenarios. Many works found social biases in VLMs [2,5,10]. For example, as illustrated in Fig. 1(c), greater similarity between the text "doctor" and male images than female ones in VLMs could undermine trust in models using these biased representations.

3 Debiasing methods

In Table 1, we summarize the properties of the existing

Received January 11, 2024; accepted July 8, 2024

E-mail: beier002@e.ntu.edu.sg

Special Issue—Excellent Young Computer Scientists Vision on Foundation Models

[1)] Note that some unbiased predictions might come from representation biases as discussed in [7]

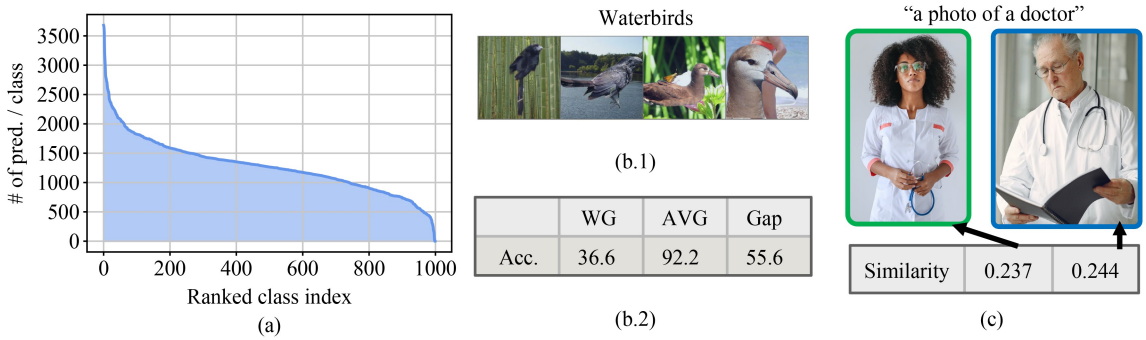

Fig. 1 Example of label bias, spurious correlation and social bias in VLMs. (a) The predictions of zero-shot VLMs on ImageNet-1K are highly imbalanced. (b.1) Examples of different groups in Waterbirds. (b.2) Worst-group, average accuracy and their gap with zero-shot classification. (c) Zero-shot VLMs exhibit gender stereotypes, with a bias towards predicting doctors as male

Table 1 Summary of debiasing methods of VLMs. PT and DS stands for pre-training and downstream respectively

Method	Debiasing type	Training data?	Retraining?
ZPE [4]	Label bias	PT data	No, post-hoc
GLA [3]	Label bias	DS data	No, post-hoc
REAL [11]	Label bias	DS&PT	Yes, linear classifier
ProReg [8]	Spurious corr.	DS data	Yes, fine-tuning
C-Adapter [9]	Spurious corr.	DS data	No, adapter
Orth-Cali [10]	Spurious corr. Social bias	No	No, adjust prompt
FairSampling [5]	Gender bias	DS data	Yes, fine-tuning
FeatureClip [5]	Gender bias	DS data	No, post-hoc
AdvDebias [12]	Social bias	DS data	No, prompt tuning
DeAR [2]	Social bias	DS data	No, adapter

debiasing methods, categorizing them from the perspectives of the type of biases they can address, whether training data is required, and whether retraining the VLMs is necessary.

Debiasing label bias The core of debiasing label bias is to estimate the pre-training label prior. Allingham et al. [4] propose ZPE which assumes that there is an access of the pre-training data, i.e., $\mathcal{D}_{\text{pt}}$. The label bias is solved by subtracting the expected logits $\log \mathbb{P}_{\text{pt}}(y)$ based on the pre-training data, i.e., $\log \mathbb{P}_{\text{pt}}(y) = \mathbb{E}_{\mathbf{v}\sim\mathcal{D}_{\text{pt}}}[s(\mathbf{v}, \mathbf{w}_y)]$. However, the pretraining data are often inaccessible because of the copyright or privacy concerns. Instead, GLA [3] proposes to only use the downstream data, i.e., $\mathcal{D}_{\text{ds}}$, to estimate the label bias. In specific, GLA adjusts the margin $\log \mathbf{q}$ of VLM classifier to achieve the lowest error in downstream data, i.e., $\min_{\mathbf{q}} \mathbb{E}_{(\mathbf{v},y)\sim\mathcal{D}_{\text{ds}}}[\ell(s(\mathbf{v}, \mathbf{w}) - \log \mathbf{q}, y)]$. The label bias is mitigated by subtracting the margin. Recently, REAL [11] replaces the original class names with the synonyms found in pre-training text to mitigate the imbalanced performance.

Improving group robustness Zhu et al. [8] found that when adapting VLMs to downstream tasks, pre-training knowledge and downstream knowledge have different biases. They propose ProReg loss to enable unbiased learning from both biased knowledge. Concretely, during fine-tuning, ProReg calculates the Kullback-Leibler divergence and cross-entropy loss for each training sample, then merge these using an adaptive sample-specific weight. Zhang et al. [9] found that the poor group robustness comes with poor similarity between sample embeddings from different groups but the same class. They propose to train an adapter, which enforces similarities between "hard" samples that zero-shot VLM predicts incorrectly and the "positive" samples that the model predicts correctly within the same class, and makes the "hard" samples far away from the nearest samples in different classes. In [10], the authors propose to debiase foundation VLMs by projecting out the biased direction in the text embedding without the additional data. Concretely, the authors first prepare a series of irrelevant features through natural language, i.e., "a photo of a [irrelevant attribute]", then construct a projection matrix to make the space of classifier weights orthogonal to these irrelevant features. This approach, which can be expressed as a closed-form solution, is training-free and effectively addresses social bias by designating protected attributes.

Debiasing social bias Wang et al. [5] target at mitigating gender bias in image searching. They propose FairSampling to sample each gender pairs equiprobably during contrastive learning. To avoid re-training of the VLMs, [5] proposes to remove feature embedding dimensions that are strongly associated with gender label. The correlation is computed based on mutual information. Berg et al. [12] employ adversarial classifier during prompt tuning for bias mitigation. Specifically, the adversarial classifier is train to predict protected attributes, e.g., gender, ethnicity and age, while the proposed method attempts to maximize the adversarial loss to achieve a representation that blind to protected attributes. Similarly, DeAR [2] learns a residual visual representation that, when added to the original, prevents the prediction of protected attributes, while remaining close to the original. Unlike [12] that jointly trains an adversarial classifier, DeAR trains the classifier separately.

4 Conclusion

While VLMs showcase remarkable capabilities, it's crucial to recognize the potential biases these models might carry into downstream applications. In this survey article, we have reviewed several current work of debiasing VLMs, focusing on label bias, spurious correlation, and social bias. Currently, most VLMs debiasing methods focus on discriminative models, such as image classification, while generative tasks like image captioning and image generation receive little attention in terms of debiasing. This could become a significant research direction in the future.

Competing interests The authors declare that they have no competing interests or financial conflicts to disclose.

References

1. Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. 2021, 8748–8763
2. Seth A, Hemani M, Agarwal C. DeAR: debiasing vision-language models with additive residuals. In: Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023, 6820–6829
3. Zhu B, Tang K, Sun Q, Zhang H. Generalized logit adjustment: Calibrating fine-tuned models by removing label bias in foundation models. In: Proceedings of the 37th Conference on Neural Information Processing Systems. 2023, 64663–64680
4. Allingham J U, Ren J, Dusenberry M W, Gu X, Cui Y, Tran D, Liu J Z, Lakshminarayanan B. A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models. In: Proceedings of the 40th International Conference on Machine Learning. 2023, 26
5. Wang J, Liu Y, Wang X. Are gender-neutral queries really gender-neutral? Mitigating gender bias in image search. In: Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. 2021, 1995–2008
6. Wang X, Wu Z, Lian L, Yu S X. Debiased learning from naturally imbalanced pseudo-labels. In: Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, 14627–14637
7. Cui J, Zhu B, Wen X, Qi X, Yu B, Zhang H. Classes are not equal: an empirical study on image recognition fairness. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2024, 23283–23292
8. Zhu B, Niu Y, Lee S, Hur M, Zhang H. Debiased fine-tuning for vision-language models by prompt regularization. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. 2023, 3834–3842
9. Zhang M, Ré C. Contrastive adapters for foundation model group robustness. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022, 1576
10. Chuang C Y, Jampani V, Li Y, Torralba A, Jegelka S. Debiasing vision-language models via biased prompts. 2023, arXiv preprint arXiv: 2302.00070
11. Parashar S, Lin Z, Liu T, Dong X, Li Y, Ramanan D, Caverlee J, Kong S. The neglected tails in vision-language models. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024, 12988–12997
12. Berg H, Hall S, Bhalgat Y, Kirk H, Shtedritski A, Bain M. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. 2022, 806–822