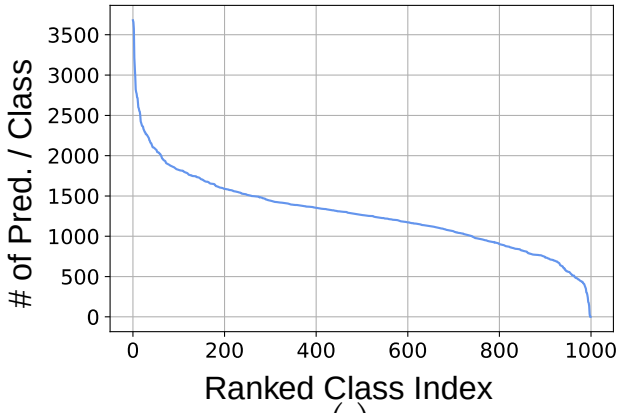
Debiasing Vision-Language Models for Vision Tasks: A Survey

Beier ZHU, Hanwang Zhang

Frontiers of Computer Science, DOI: 10.1007/s11704-024-40051-3

Biases in VLMs

- Despite the impressive capabilities, it is concerning that the VLMs are prone to inheriting biases from the uncurated datasets scraped from the Internet
  - Label bias, certain classes (words) appear more frequently in the pre-training data.
  - Spurious correlation, non-target features, e.g., image background, that are correlated with labels, resulting in poor group robustness.
  - Social bias, which is a special form of spurious correlation, focuses on societal harm.



(a)

(b.1)

	WG	AVG	Gap
Acc.	36.6	92.2	55.6

(b.2)

(c)

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.

Main Contributions

- Summary of Debiasing Methods

Method	Debiasing type	Training data?	Retraining?
ZPE [6]	Label bias	PT data	No, post-hoc
GLA [5]	Label bias	DS data	No, post-hoc
REAL [15]	Label bias	DS&PT	Yes, linear classifier
ProReg [12]	Spurious corr.	DS data	Yes, fine-tuning
C-Adapter [13]	Spurious corr.	DS data	No, adapter
Orth-Cali [14]	Spurious corr. Social bias	No	No, adjust prompt
FairSampling [7]	Gender bias	DS data	Yes, fine-tuning
FeatureClip [7]	Gender bias	DS data	No, post-hoc
AdvDebias [16]	Social bias	DS data	No, prompt tuning
DeAR [4]	Social bias	DS data	No, adapter

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