A Prompt Array Keeps the Bias Away: Debiasing Vision-Language Models with Adversarial Learning

Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat,
Hannah Rose Kirk, Aleksandar Shtedritski and Max Bain

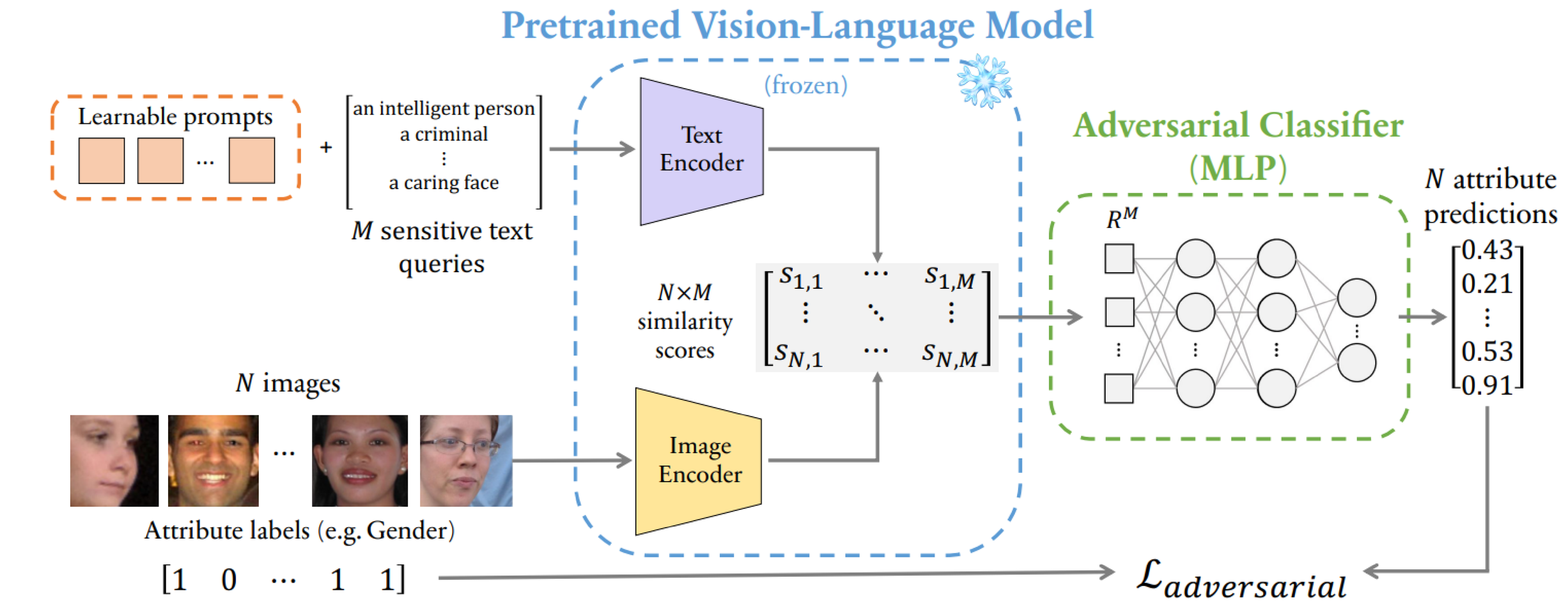

Figure 1: Our proposed debiasing method for pre-trained vision-language models

Motivation

Large-scale vision-language (VL) are becoming increasingly pervasive in our lives, mainly due to their superficially impressive performance on downstream tasks (e.g., semantic search queries) with minimal fine-tuning. However, these models are trained on large-scale internet datasets that are too large to be manually audited for their ingrained representational biases and require infeasible compute to retrain. This runs the risk of entrenching social and cultural biases with 'snapshots in time'. Therefore, there is a need to develop cheap and efficient methods, which don't rely on access to the original datasets nor excessive compute resources, for measuring and mitigating bias in these models.

Aim

Having established a baseline for evaluating bias in the model; the aim is to learn unbiased joint image-text representations such that:

1. The model outputs a similarity score image-text pairs
2. The model is unbiased (defined as outputting similar distributions of scores across attributes for a given text query)

Methods

Investigating and evaluating different measures of bias:

1. Word Embedding Association Test (WEAT): found to be too sensitive to changes in model architecture, syntactic changes in inputs and datasets, despite promising use in the pure NLP domain
2. Ranking metrics (*MaxSkew* and *NDKL*): these have a natural application in the VL domain due to their increased use for semantic image search. These were found be useful baselines for bias measurement

Debiasing:

1. Fairness objective with adversarial learning:
   The aim of the adversarial classifier θ_{adv} is to predict the attribute label A of image I given only its similarity logits from the set of sensitive text queries. The adversary is trained to minimise the cross-entropy loss between the predicted attributed labels and the ground truth labels
2. Various optimisation methods are deployed to preserve feature representations and downstream performance: regularisation, prepending learnable text tokens and joint training (jointly optimising for unbiasedness and image-text contrastive loss)

Key contributions

1. Evaluation of suitable methods for measuring bias in VL models (*MaxSkew* and *NDKL*)
2. Evaluate gender and racial bias on state-of-the-art pretrained VL models on standardised face datasets
3. Provide a framework for debiasing VL models requiring only sensitive attribute labels as supervision. We show that jointly optimising for unbiasedness and image-text contrastive loss (ITC) via prepending an array of learnable prompt tokens to text queries reduces bias without decreasing the quality of the image-text representations

Figure 2: Bias measures (for gender) across combinations of syntactic changes, models (RN50, ViTB/16, ViTB/32), and datasets (FairFace validation set & UTKFace). We use WEAT pairwise adjectives concept sets (Caliskan et al., 2017). Standard deviation is taken over all combinations of model architecture and datasets (other results we use ViTB/32)

Table 1: Measuring the effect on gender bias and performance of prepending prompt tokens; adversarial debiasing on FairFace; and ITC training on Flickr30k-train

Model	Bias↓		Performance↑	
	$MaxSkew@1K$	*NDKL*	$flickr_{R@5}$	$IN1K_{acc}$
CLIP	0.233	0.104	85.9	68.1
CLIP-clip ($m = 400$)	0.073(-69%)	0.023(-78%)	78.5(-9%)	64.6(-5%)
CLIP-clip ($m = 256$)	**0.056(-76%)**	0.023(-78%)	63.7(-26%)	55.8(-18%)
$CLIP_{+prompt}$ (debias)	0.073(-69%)	**0.021(-80%)**	64.2(-25%)	54.9(-19%)
$CLIP_{+prompt}$ (itc)	0.247(+6%)	0.104(+0%)	**90.6(+5%)**	**68.4(+0%)**
$CLIP_{+prompt}$ (debias+itc)	0.113(-52%)	0.036(-65%)	88.5(+3%)	67.6(-1%)

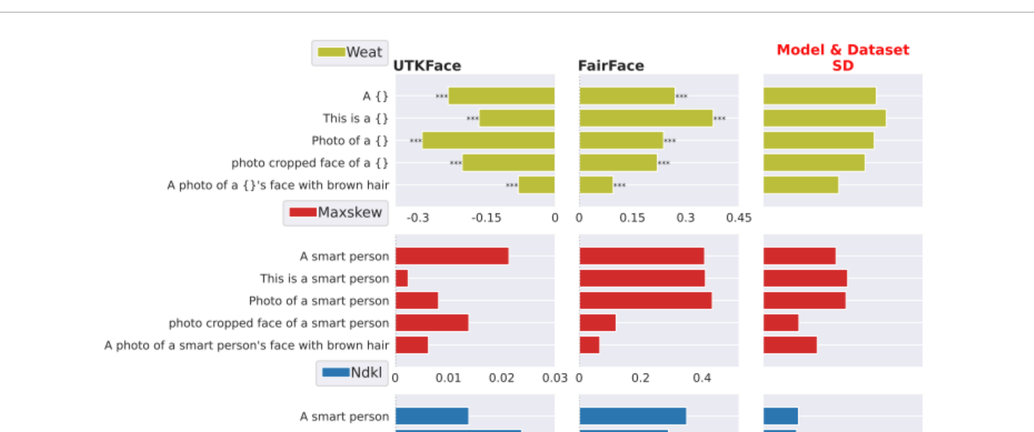

Figure 3: Text query - "A photo of a *smart* person"