Debiasing Convolutional Neural Networks via Meta Orthogonalization

Kurtis David[†], Qiang Liu[†], Ruth Fong[‡]

[†]UT Austin [‡] Oxford University

Motivation

While deep learning models often achieve strong task performance, their successes are hampered by their inability to disentangle spurious correlations from causative factors, such as **protected attributes** (e.g. race, gender) to make decisions. In this work, we tackle the problem of debiasing **convolutional neural networks** (CNNs).

Inspired by advances in interpretability literature, we show that it is possible to apply previous techniques in debiasing word embeddings [1,4] to visual systems. Through a suite of datasets simulating varying levels of bias, we show that our method, **Meta Orthogonalization**, can be competitive to the state of the art adversarial debiasing methods.

Visual Concepts in Networks

Image concepts such as *color, objects, textures* can be represented as learned **embedding vectors** given a trained CNN with parameters θ [3,5,8]. See Figure 1 as an example.

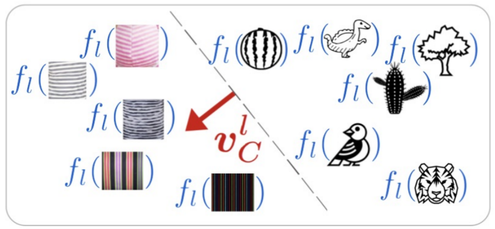

Figure 1: Learning “stripes” concept vector [5].

Naturally these are learned *post-hoc*, or after training; however we learn concepts **simultaneously** with θ to obtain the most “current" representations of concepts while applying our framework.

Debiasing Word Embeddings

Because image concepts have vector representations like word embeddings, notice that we can apply methods of debiasing word embeddings [1] that induce gender neutral words to be **orthogonal** *w.r.t.* a bias direction e.g. gender.

Because our image concept embeddings are learned simultaneously, we follow [4] that augments training with the following penalty function:

$$\mathscr{L}_{\text{debias}} = \sum_{\beta} (\beta^\top \nu)^2,$$

where ν represents the bias direction in the word vector space, and β represents a desired word embedding to be debiased w.r.t. ν.

Meta Orthogonalization

To do this, we can jointly optimize for three losses (visualized in Figure 2):

1. $\mathscr{L}_{\text{class}}$ — original task loss to train the CNN
2. $\mathscr{L}_{\text{concept}}$ — logistic loss to learn image concepts
3. $\mathscr{L}_{\text{debias}}$ — penalty term forcing task concepts to be orthogonal to bias

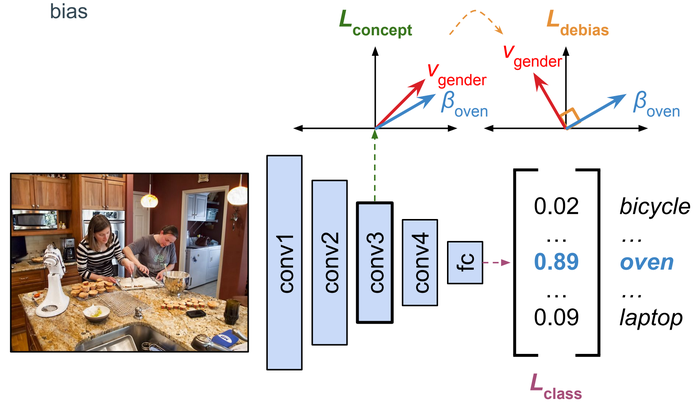

Figure 2: Method overview.

Given that β_c denotes a concept embedding, and ν the bias concept direction, one way to apply (3) is to penalize the projections of each β_c on ν. However, this does not affect the model parameters θ; to do so, we note that if we learn β_c using SGD, each update is of the form:

$$\beta_c' = \beta_c - \alpha \nabla_{\beta_c} \mathscr{L}_{\text{concept}}(c, \theta).$$

Thus, we can instead use β_c' in the learning update to regularize θ. This is sometimes called the “meta step” [2,6].

Datasets

We first test our method against adversarial debiasing on the Benchmarking Attribution Methods (**BAM**) dataset [7]. Its unique setup allows us to control **object-class** co-occurrences, as well as simulate a protected attribute as a pair of objects (Figure 3). Refer to main paper for all situations in Figure 4.

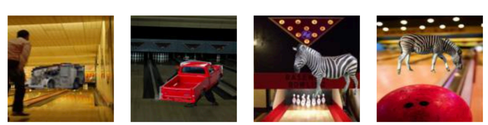

Figure 3: Example images from BAM, with existence of truck” and “zebra” objects as the protected attribute.

Lastly, we use **COCO** to also show our results on a real world dataset, with gender annotation labels as the protected attribute.

Experimental Results

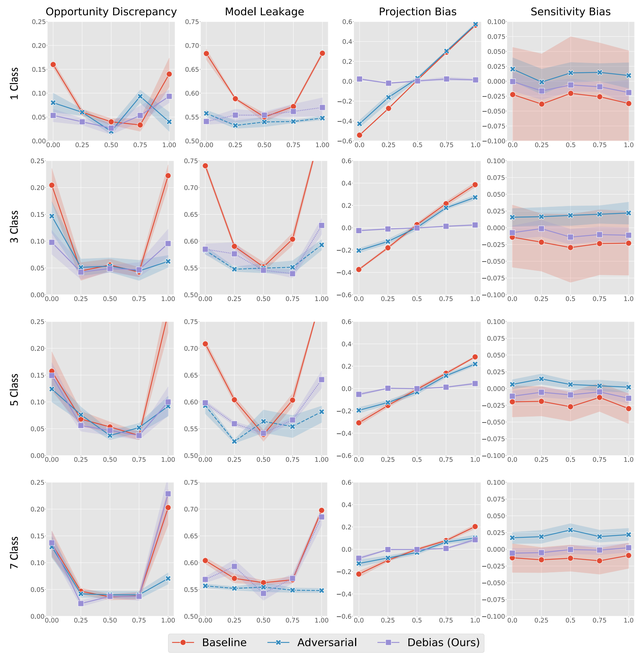

Figure 4: BAM Results. First two columns lower is better. Last two columns, closer to 0 is better.

We see that as bias is strengthened by the co-occurrence ratio (x-axis) or # Classes, the baseline ResNet50 exhibits stronger signs of bias, confirming our experimental setup. Regardless, our method matches results from adversarial debiasing, if not better. The worst outcomes are in the “7 Class” case when the co-occurrence ratio is at its extremes.

Model	Discrepancy	Leakage	Projection Bias	Sensitivity Bias
Baseline	0.1314 ±0.0217	70.46*	−0.4086 ±0.0243	0.0172 ±0.0185
adv@conv5	—	64.92*	—	—
adv@image[†]	0.1400 ±0.0200	68.49*	−0.4225 ±0.0275	0.0165 ±0.0141
Debias(Ours)	**0.1245** ±0.0189	**61.07**	**0.0003** ±0.0416	**0.0066** ±0.0067

Figure 5: COCO Results. First two columns lower is better. Last two columns, closer to 0 is better.

In COCO, we show significant gains across **all metrics** against adversarial debiasing, even though we only train to minimize **projection bias.** Theoretical connections between orthogonality of concepts and fairness may exist, and are targets for future work.

Paper URL:

References: [1] Bolukbasi et al., *NIPS* 2016 [2] Finn et al., *ICML* 2017 [3] Fong and Vedaldi, *CVPR* 2018 [4] Kaneko and Bollegala, *ACL* 2019 [5] Kim et al., *ICML* 2018 [6] Rebuffi et al., *CVPR* 2020 [7] Yang and Kim, 2019 [8] Zhou et al., *ECCV* 2018