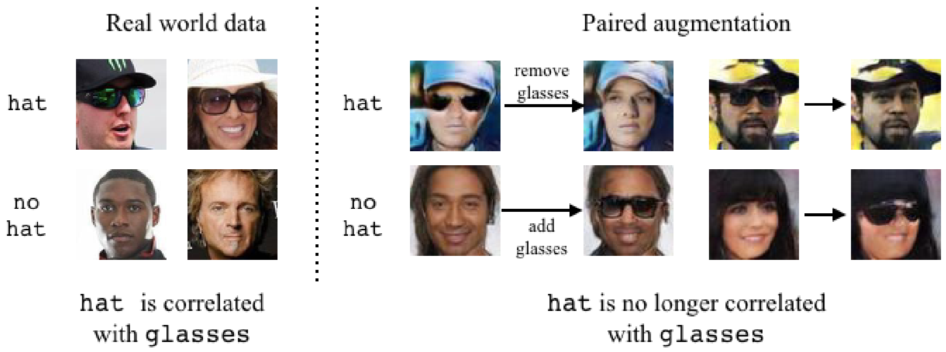
(https://eccv2020.eu/)

Princeton Visual AI Lab (https://visualai.princeton.edu)

Fair Attribute Classification through Latent Space De-biasing

Vikram V. Ramaswamy (https://www.cs.princeton.edu/~vr23/) Sunnie S. Y. Kim (https://www.cs.princeton.edu/~suhk/) Olga Russakovsky (https://www.cs.princeton.edu/~olgarus/)
Princeton University
{vr23, suhk, olgarus}@cs.princeton.edu



Training a visual classifier for an attribute (e.g., wearing hat) can be complicated by correlations in the training data. For example, the presence of hats can be correlated with the presence of glasses. We propose a dataset augmentation strategy using Generative Adversarial Networks (GANs) that successfully removes this correlation by adding or removing glasses from existing images, creating a balanced dataset.

(https://arxiv.org/abs(https://github.com/p(https://youtu.be/j5L(https://youtu.be/1eb(https://colab.research.google
debiasing) deeplearning-ai/GANs-Public/blob/master/C2W2_G

Paper | Code | 2min Talk | 10min Talk | Colab Notebook

Abstract

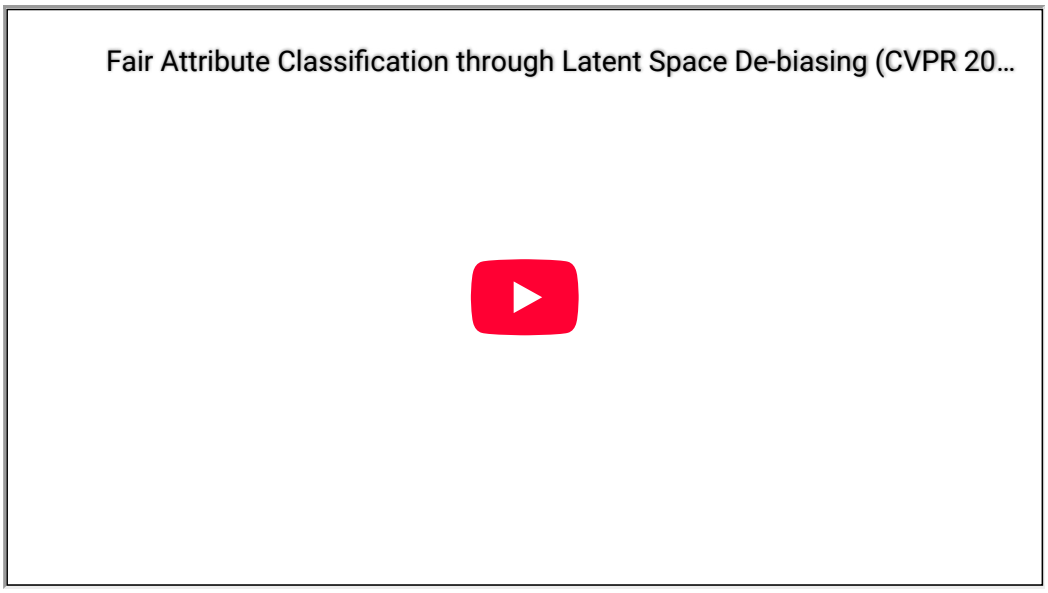
Fairness in visual recognition is becoming a prominent and critical topic of discussion as recognition systems are deployed at scale in the real world. Models trained from data in which target labels are correlated with protected attributes (e.g., gender, race) are known to learn and exploit those correlations. In this work, we introduce a method for training accurate target classifiers while mitigating biases that stem from these correlations. We use GANs to generate realistic-looking images, and perturb these images in the underlying latent space to generate training data that is balanced for each protected attribute. We augment the original dataset with this perturbed generated data, and empirically demonstrate that target classifiers trained on the augmented dataset exhibit a number of both quantitative and qualitative benefits. We

conduct a thorough evaluation across multiple target labels and protected attributes in the CelebA dataset, and provide an in-depth analysis and comparison to existing literature in the space.

Citation

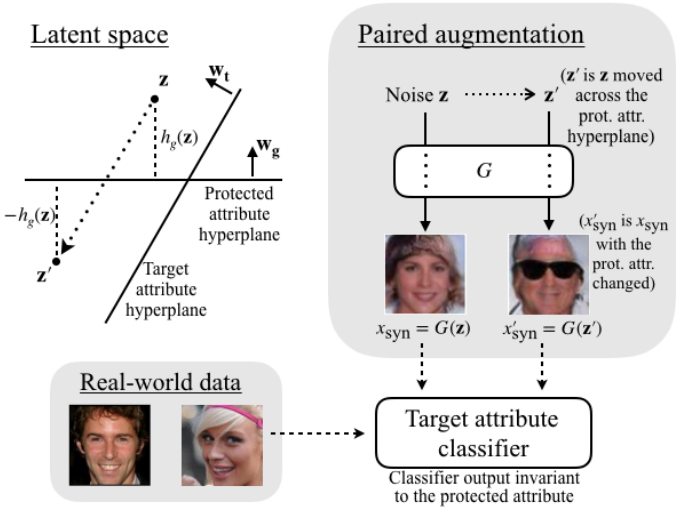
```
@inproceedings{ramaswamy2020gandebiasing,
    author = {Vikram V. Ramaswamy and Sunnie S. Y. Kim and Olga Russakovsky},
    title = {Fair Attribute Classification through Latent Space De-biasing},
    booktitle = {IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)},
    year = {2021}
}
```

5-Minute Talk



Creating a De-biased Dataset to Train Fairer Attribute Classifiers

We propose a method for perturbing latent vectors in the GAN latent space that successfully de-correlates target and protected attributes and allows for augmenting and de-biasing the real-world dataset.



(Top left) The trained GAN learns a distribution from which it samples z. For each z sampled, we compute z' such that its target attribute (e.g., wearing hat) score remains the same according to wa, while its protected attribute (e.g., wearing

glasses) score is negated according to wg.

(Top right) We add images G(z) and G(z') to our training set, and train a target classifier on both the real-world dataset and the balanced synthetic dataset generated through our paired augmentation method.

Results

In the below table, we compare our model (i.e. target classifier trained on both the real-world dataset and the balanced synthetic dataset) with a baseline model trained on the real-world dataset. We evaluate the models with four metrics: average precision (AP), difference in equality of opportunity (DEO), bias amplification (BA), and KL divergence between score histograms (KL). The results are averaged for each attribute category: inconsistently labeled, gender-dependent, and gender-independent.

Attr. type	AP ↑		DEO ↓	
	Baseline	Ours	Baseline	Ours
Incons.	**66.3 ± 1.8**	65.2 ± 1.9	21.5 ± 4.4	**16.5 ± 4.2**
G-dep	**78.6 ± 1.4**	77.8 ± 1.4	25.7 ± 3.5	**23.4 ± 3.6**
G-indep.	**83.9 ± 1.5**	83.0 ± 1.6	16.7 ± 5.0	**13.9 ± 5.2**
Attr. type	BA ↓		KL ↓	
	Baseline	Ours	Baseline	Ours
Incons.	2.1 ± 0.6	**0.5 ± 0.6**	1.7 ± 0.3	**1.3 ± 0.4**
G-dep	2.3 ± 0.5	**1.6 ± 0.5**	1.3 ± 0.2	**1.2 ± 0.2**
G-indep.	0.3 ± 0.6	**0.0 ± 0.5**	1.1 ± 0.5	**0.9 ± 0.6**

Our model performs better on all three fairness metrics, DEO, BA and KL, while maintaining comparable AP.

We provide an in-depth analysis of our method and comparison to existing literature in the paper. Our findings show the promise of augmenting data in the GAN latent space in a variety of settings.

Related Work

Below are some papers related to our work. We discuss them in more detail in the related work section of our paper.

Image Counterfactual Sensitivity Analysis for Detecting Unintended Bias. (https://arxiv.org/abs/1906.06439) Emily Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, Andrew Zaldivar. CVPR 2019 Workshop on Fairness Accountability Transparency and Ethics in Computer Vision.

Fairness GAN. (https://arxiv.org/abs/1805.09910) Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, Kush R. Varshney. IBM Journal of Research and Development 2019.

Contrastive Examples for Addressing the Tyranny of the Majority. (https://arxiv.org/abs/2004.06524) Viktoriia Sharmanska, Lisa Anne Hendricks, Trevor Darrell, Novi Quadrianto. arXiv 2020

Fair Generative Modeling via Weak Supervision. (https://arxiv.org/abs/1910.12008) Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, Stefano Ermon. ICML 2020

Towards Causal Benchmarking of Bias in Face Analysis Algorithms. (https://arxiv.org/abs/2007.06570) Guha Balakrishnan, Yuanjun Xiong, Wei Xia, Pietro Perona. ECCV 2020

Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation. (https://arxiv.org/abs/1911.11834) Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, Olga Russakovsky. CVPR 2020

Acknowledgements

This work is supported by the National Science Foundation under Grant No. 1763642 and the Princeton First Year Fellowship to SK. We also thank Arvind Narayanan, Deniz Oktay, Angelina Wang, Zeyu Wang, Felix Yu, Sharon Zhang, as well as the Bias in AI reading group for helpful comments and suggestions.

Contact

Vikram V. Ramaswamy (https://www.cs.princeton.edu/~vr23/) (vr23@cs.princeton.edu)