DetoxAI: a Python Toolkit for Debiasing Deep Learning Models in Computer Vision

Ignacy Stępka[1] ⋆[0009−0004−4575−0689], Lukasz Sztukiewicz[1] ⋆[0009−0008−7077−0226], Michał Wiliński[1] ⋆[0009−0004−4818−8417], and Jerzy Stefanowski[1] (✉)[0000−0002−4949−8271]

Institute of Computing Science, Poznan University of Technology
jerzy.stefanowski@cs.put.poznan.pl

Abstract. While machine learning fairness has made significant progress in recent years, most existing solutions focus on tabular data and are poorly suited for vision-based classification tasks, which rely heavily on deep learning. To bridge this gap, we introduce DetoxAI, an open-source Python library for improving fairness in deep learning vision classifiers through post-hoc debiasing. DetoxAI implements state-of-the-art debiasing algorithms, fairness metrics, and visualization tools. It supports debiasing via interventions in internal representations and includes attribution-based visualization tools and quantitative algorithmic fairness metrics to show how bias is mitigated. This paper presents the motivation, design, and use cases of DetoxAI, demonstrating its tangible value to engineers and researchers.

Keywords: Fairness · Deep Learning · Computer Vision · Debiasing

1 Introduction

Ensuring fairness in machine learning models has become critical, particularly in high-stakes fields [5]. While several libraries address fairness, most focus on tabular data and are ill-suited for unstructured, high-dimensional tasks like computer vision. We identify two major gaps in the current landscape.

The first is technical: existing tools such as AIF360 [2] and Fairlearn [4] are built around the scikit-learn API, expecting datasets to fit in memory as Pandas DataFrames or NumPy arrays. This design is incompatible with deep learning workflows, where data must be processed in small batches, forcing practitioners to abandon popular toolkits and manually reimplement fairness methods.

The second issue is methodological: current tools primarily offer quantitative metrics or basic visualization capabilities, which are insufficient for analyzing biases in computer vision models. Furthermore, the post-hoc debiasing techniques they provide typically operate only on the model's outputs - such as by adjusting classification thresholds - without addressing or removing bias in the model's underlying reasoning process.

⋆ Equal contribution

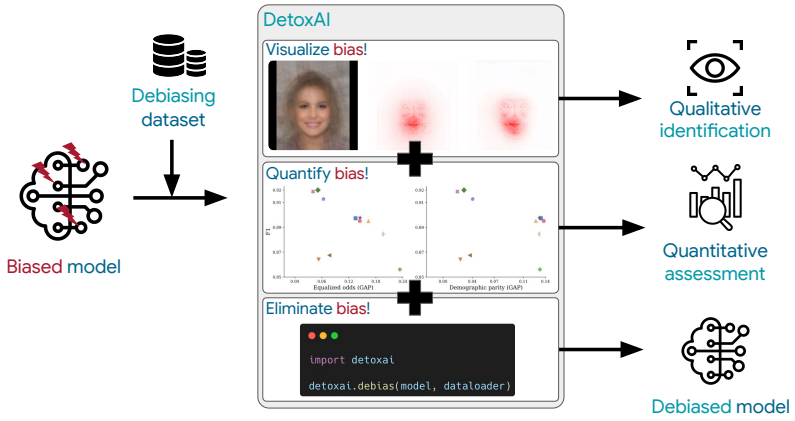

Fig. 1. DetoxAI incorporates multiple tools to mitigate biases in vision models. Biased model with debiasing dataset is passed to our module, where biases can be inspected, measured and finally eliminated.

To address these gaps, we introduce DetoxAI, a post-hoc debiasing toolkit for image classification that operates at the representation level. DetoxAI enables desensitization of neural networks to protected attributes (e.g., gender, race) without requiring full retraining. Designed for deep learning and seamlessly integrated with PyTorch, DetoxAI equips AI practitioners and researchers with a practical tool for empirical evaluation, comparative studies, and the deployment of fairness interventions in real-world vision models.

2 Use Cases

Engineering Use Case: Debiasing a facial expression recognition system Consider a facial expression recognition system deployed to detect whether customers are smiling. In practice, it is observed that the system consistently misclassifies individuals wearing neckties as not smiling, revealing unintended bias. Using DetoxAI, engineers can address this by treating the necktie as a protected attribute. DetoxAI allows them to load the existing model (e.g., an already trained and deployed ResNet50) and apply targeted debiasing techniques - all without full model retraining. The updated model can then be seamlessly redeployed, mitigating the bias while maintaining overall performance. Engineers can evaluate fairness metrics, visualize attribution shifts, and select the optimized model, enhancing both fairness and system reliability.

Research Use Case: Fairness studies and comparative benchmarking Researchers studying fairness can leverage DetoxAI as a standardized platform for systematic benchmarking. DetoxAI enables easy comparison of post-hoc debiasing methods under consistent pipelines, evaluation on common fairness metrics, and qualitative analysis of shifts in feature importance. Its modular design

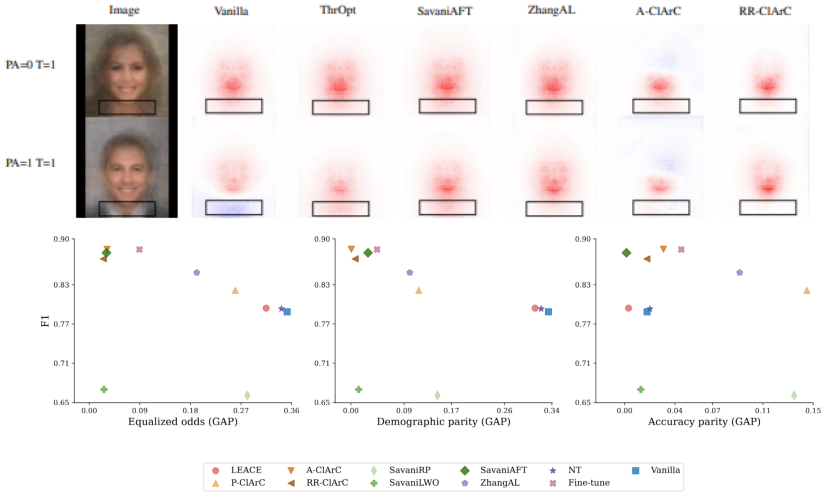

Fig. 2. In the upper image, we visualize saliency maps before (Vanilla) and after bias mitigation using a selection of implemented methods. The image on the bottom is a quantitative evaluation of the debiasing techniques on the predictive performance (F1 Score) - fairness trade-off. Fairness metrics are EqualizedOdds, DemographicParity and AccuracyParity, calculated in a difference between protected groups variation [5].

and clear abstractions simplify the integration of novel debiasing methods and benchmarking against established techniques. This flexibility can help accelerate fairness research in image classification and supports rigorous, extensible and reproducible experimentation.

3 System Overview

DetoxAI implements several post-hoc debiasing methods, including Savani and Zhang based methods [6], LEACE [3], and post-hoc Threshold Optimization [5] It also includes ClArC variants [1] originally used as artifact-removal methods but repurposed for fairness [7]. All methods, except Threshold Optimization, modify internal representations instead of merely calibrating outputs, enabling deeper mitigation of learned biases.

Debiasing techniques are accessed through a unified `detoxai.debias(...)` interface (Fig. 1). The API supports debiasing, performance evaluation (e.g., F1, GMean, Balanced Accuracy), and fairness metric computation (e.g., Equalized Odds, Demographic Parity, Accuracy Parity), offering seamless integration into existing PyTorch workflows.

DetoxAI is model-agnostic and focuses on binary classification tasks with binary protected attributes. Even though certain debiasing methods require internal model interventions (e.g., hooks) and fine-tuning, DetoxAI automatically adapts to a variety of PyTorch models without requiring additional user in-

put. All debiasing methods come with default configurations, empirically tuned for robustness across various model sizes and architectures. Advanced users can override these defaults by passing custom configurations to the API to better meet specific needs, such as computational budget or optimized fairness metrics. All components follow an object-oriented design, making it straightforward to extend DetoxAI with new methods, metrics, or visualization tools (see Sec. 4 for examples and documentation).

4 Conclusions

DetoxAI offers a unified platform for implementing fairness interventions in deep learning systems for image classification tasks. Designed to be production-ready, yet highly extensible, the toolkit supports a wide range of practical applications across both industry and research. Its simple and consistent API lowers the barrier to applying bias mitigation techniques, making fairness interventions accessible to engineers and researchers alike.

Additional Information

Webpage and video: `https://detoxai.github.io`, GitHub: `https://github.com/DetoxAI/detoxai`, Documentation: `https://detoxai.readthedocs.io`.

References

1. Anders, C.J., Weber, L., Neumann, D., Samek, W., Müller, K.R., Lapuschkin, S.: Finding and removing clever hans: Using explanation methods to debug and improve deep models. Information Fusion **77**, 261–295 (2022)
2. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al.: Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development **63**(4/5), 4–1 (2019)
3. Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., Biderman, S.: Leace: perfect linear concept erasure in closed form. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp. 66044–66063 (2023)
4. Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K.: Fairlearn: A toolkit for assessing and improving fairness in ai. Microsoft, Tech. Rep. MSR-TR-2020-32 (2020)
5. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys **54**(6) (Jul 2021)
6. Savani, Y., White, C., Govindarajulu, N.S.: Intra-processing methods for debiasing neural networks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 2798–2810. Curran Associates, Inc. (2020)
7. Sztukiewicz, L., Stępka, I., Wiliński, M., Stefanowski, J.: Investigating the relationship between debiasing and artifact removal using saliency maps (2025), `https://arxiv.org/abs/2503.00234`