Unfair CNN? Debias.

Group of Horribly Optimistic STatisticians

Enhancing Fairness in Neural Networks with Debiasing Techniques

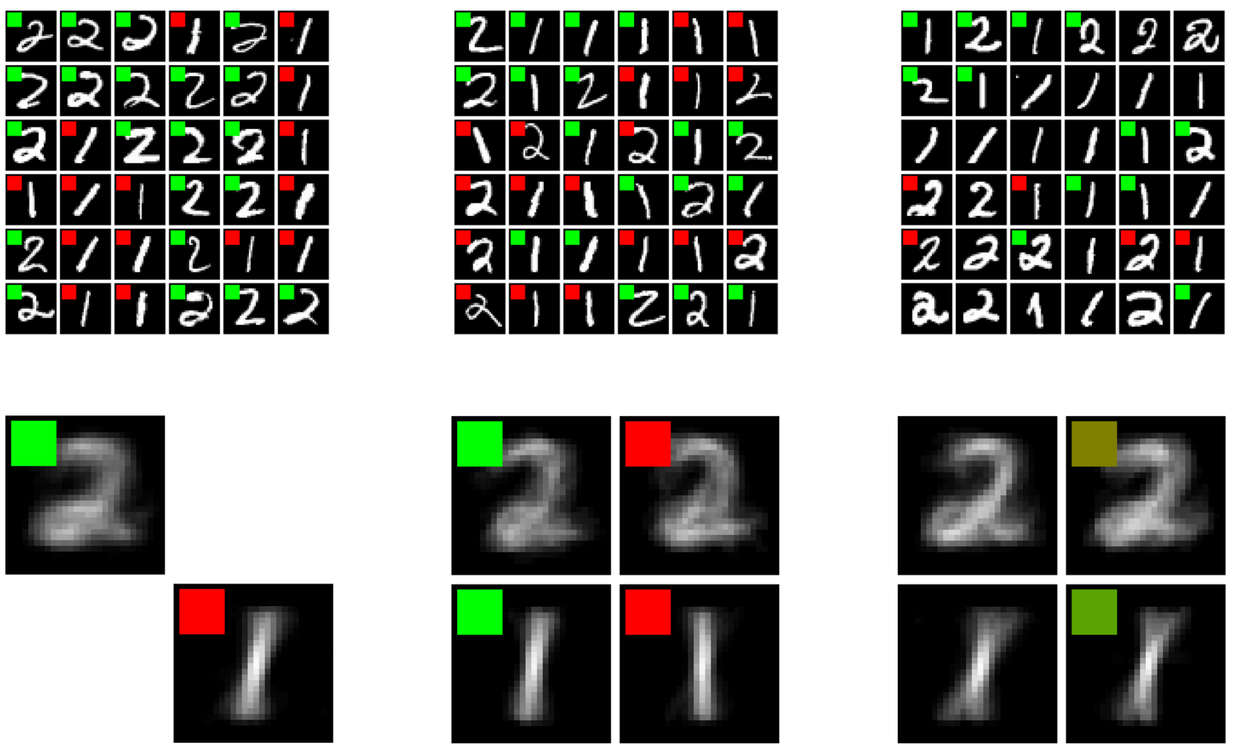
Łukasz Sztukiewicz, Ignacy Stępka, Michał Wiliński, Jerzy Stefanowski
Poznan University of Technology, Poznań, Poland

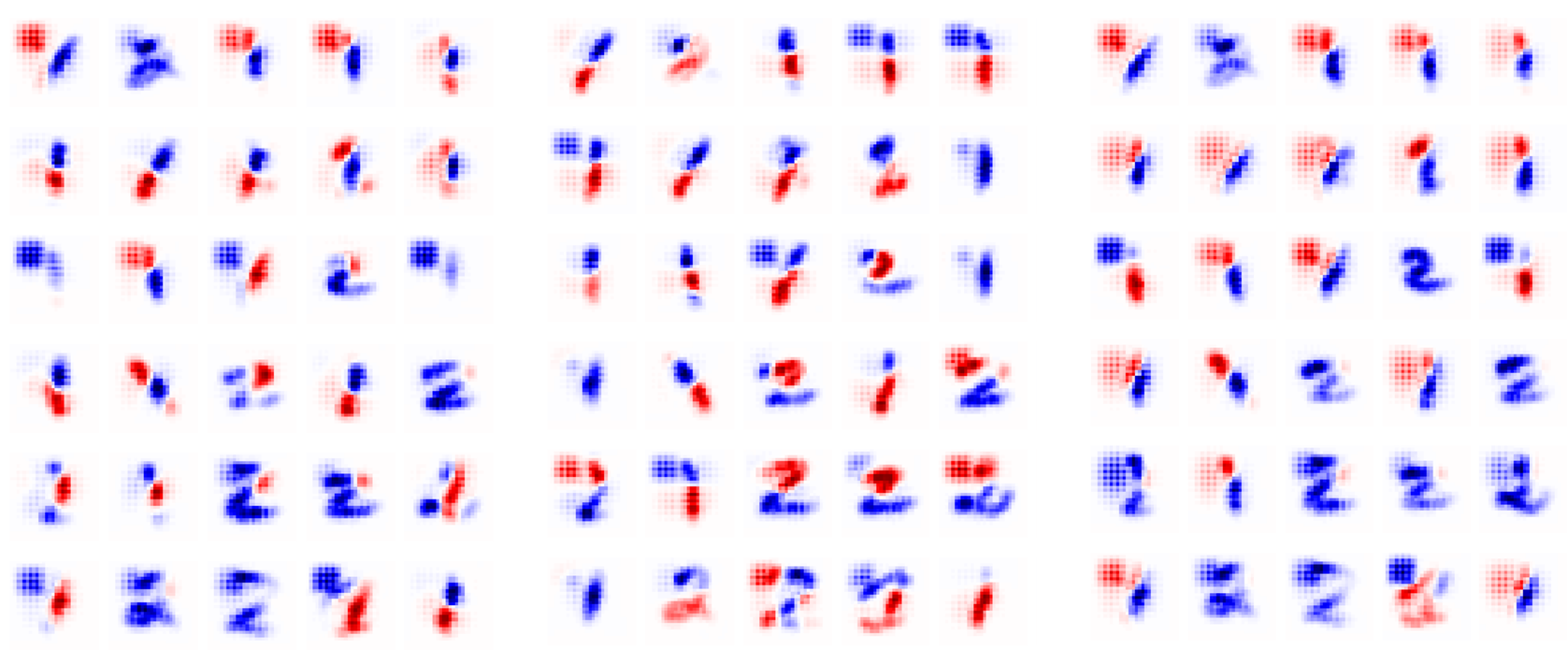
Motivation

- Convolutional neural networks (CNNs) often learn harmful biases (concepts), leading to potential unfair treatment of protected groups
- We aim to determine whether unlearning harmful concepts can enhance model fairness
- Bias mitigation is critical for development of both trustworthy and ethically responsible AI systems

Methods

- $h_{ClArC}(a) = (I - vv^T)a + vv^T \mu_{\mathbf{A}_{non-artifact}}$
- $h_{\text{mass-mean}}(a) = a - (\mu_{\mathbf{A}_{artifact}} - \mu_{\mathbf{A}_{non-artifact}})$





Results discussion

- The threshold optimizer targeting Equalized Odds collapses, while targeting Demographic Parity shows slight improvement in the metric
- Layer-Wise Relevance Propagation (LRP) visualizations reveal bias and the impact of debiasing
- Debiasing techniques enhance fairness by targeting prediction construction issues
- P-ClArC significantly boosts both performance and fairness metrics
- Mass-mean probing, despite its simplicity, yields promising results

	Macro Accuracy ( )	Demographic Parity ( )	Equalized Odds ( )	Equality of Opportunity ( )
Base CNN	51.11	0.99	0.98	0.99
Threshold Optimizer (DP)	51.02 (-0.09)	0.97 (-0.02)	0.98 (+0.00)	0.99 (+0.00)
Threshold Optimizer (EO)	50.00* (-1.11)	0.00* (-0.99)	0.00* (-0.99)	0.00* (-0.99)
P-ClArC	96.51 (+45.40)	0.07 (-0.92)	0.06 (-0.92)	0.05 (-0.94)
Mass-mean	68.25 (+17.14)	0.98 (-0.01)	0.54 (-0.44)	0.98 (-0.01)

References

[1] Bach, Sebastian, et al. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation." PloS one 10.7 (2015)
[2] Anders, Christopher J., et al. "Finding and Removing Clever Hans: Using Explanation Methods to Debug and Improve Deep Models." Information Fusion 77 (2022): 261-295.
[3] Marks, Samuel, and Max Tegmark. "The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets." arXiv preprint arXiv:2310.06824 (2023).
[4] Weerts, Hilde, et al. "Fairlearn: Assessing and Improving Fairness of AI systems." Journal of Machine Learning Research 24.257 (2023): 1-8.