Unfair CNN? Debias.

Group of Horribly Optimistic STatisticians

Enhancing Fairness in Neural Networks with Debiasing Techniques

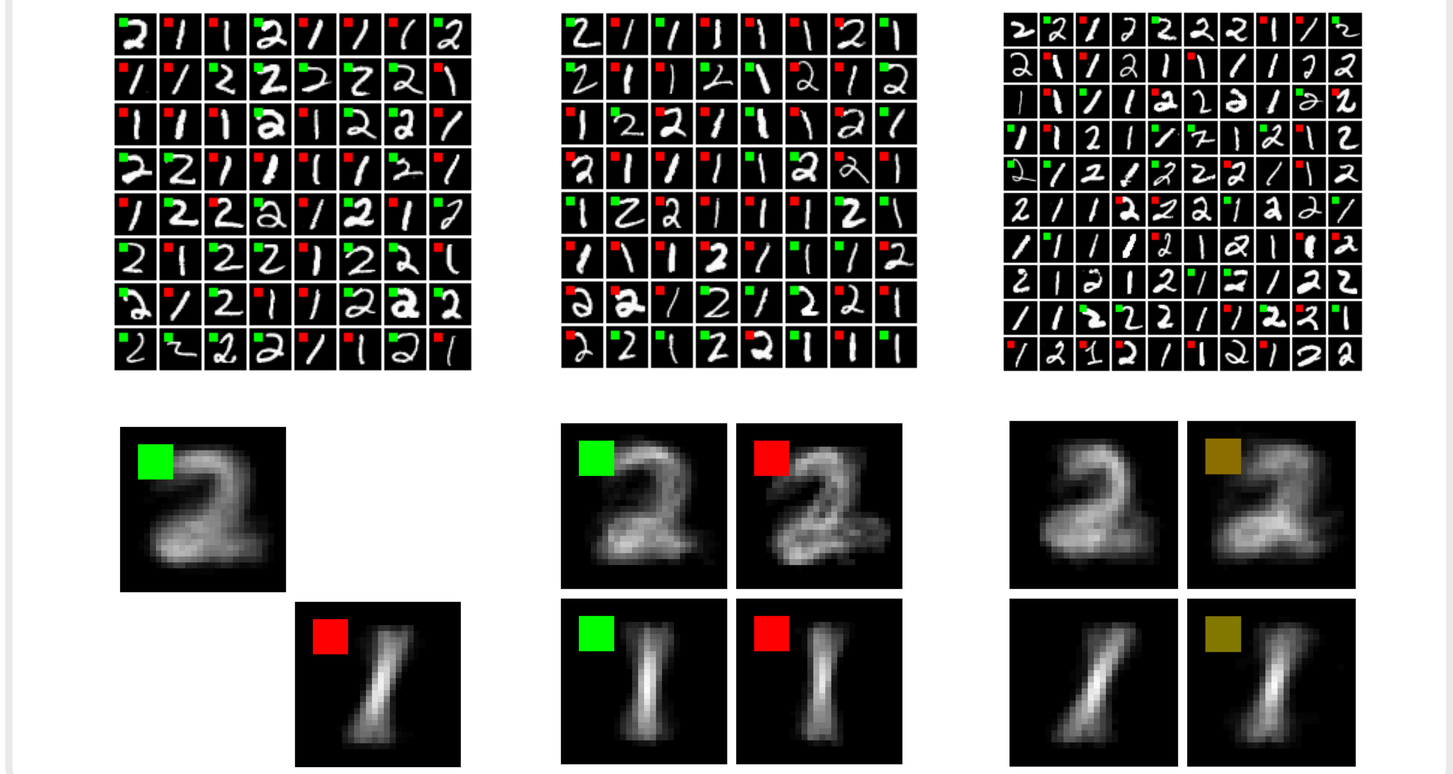
Łukasz Sztukiewicz, Ignacy Stępka, Michał Wiliński, Jerzy Stefanowski
Poznan University of Technology, Poznań, Poland

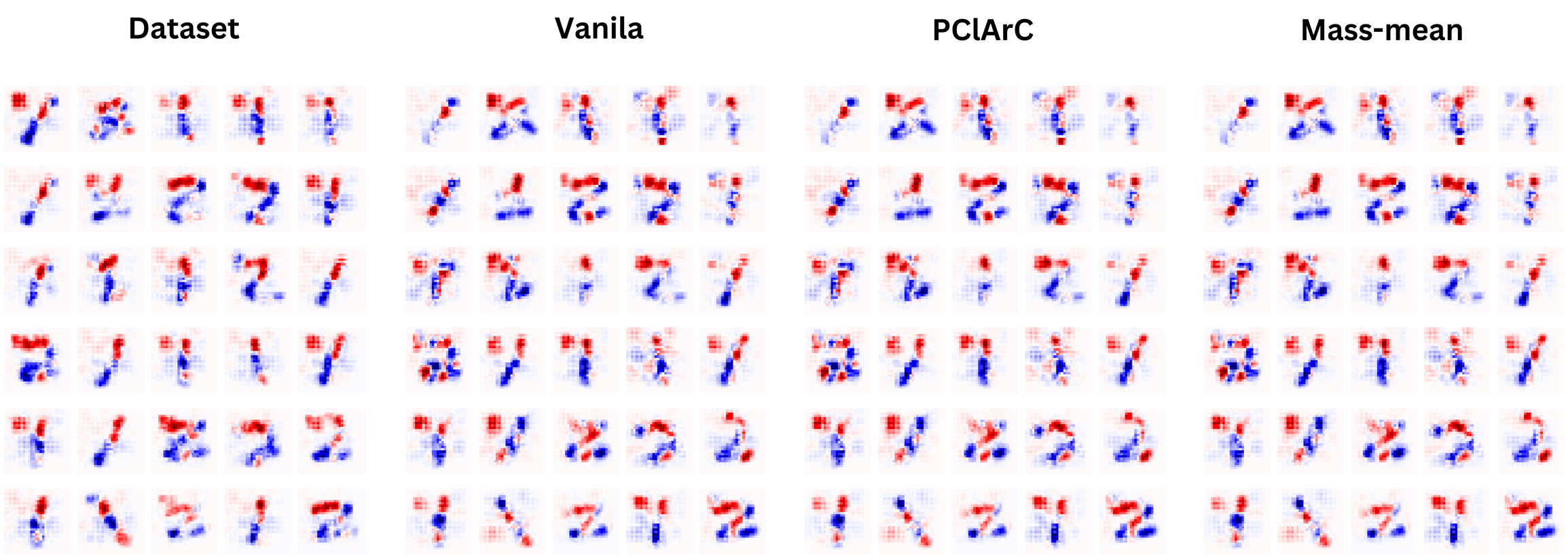
Motivation

- CNNs often learn harmful biases, leading to unfair treatment of protected groups
- We aim to check whether unlearning harmful concepts can improve model fairness
- Addressing bias is crucial for building 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

Co zaobserwowaliśmy

Tabelka z wynikami

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.