Information and Networking Event
Horizon Europe 2023-2024 Calls
Co-Funded by the Government of India (DST)

HORIZON-CL4-2024-HUMAN-03-02: Explainable and Robust AI

24 May 2024

RedTeaming AI models for Trustworthy AI

prof. Przemyslaw *[p-sh-eh-m-eh-k]* Biecek
Przemyslaw.Biecek@pw.edu.pl
https://pbiecek.github.io/

Warsaw University of Technology, Poland
https://mi2.ai/

Why do we XAI?

ICML 2024

Computer Science > Artificial Intelligence

[Submitted on 21 Feb 2024]

Explain to Question not to Justify

Przemyslaw Biecek, Wojciech Samek

Explainable Artificial Intelligence (XAI) is a young but very promising field of research. Unfortunately, the progress in this field is currently slowed down by divergent and incompatible goals. In this paper, we separate various threads tangled within the area of XAI into two complementary cultures of human/value-oriented explanations (BLUE XAI) and model/validation-oriented explanations (RED XAI). We also argue that the area of RED XAI is currently under-explored and hides great opportunities and potential for important research necessary to ensure the safety of AI systems. We conclude this paper by presenting promising challenges in this area.

Fallacies behind the XAI crisis

models are either interpretable or not
(single) XAI silver bullet exists
true explanations exists
user study is the ultimate validation
only users need explanations

solved after sorting out

Two XAI Cultures

validation-oriented	*value-oriented*
Research	responsi**B**le
Explore	**L**egal
Debug	tr**U**st
	Ethics

new research perspectives

So what? New Challenges

supplementary and complementary explanations
multiple models (Rashomon perspective)
explorer mindset for data and models
benchmarks, tools, standards
XAI for Science

	Model-validation oriented RED XAI	Human-values oriented BLUE XAI
Why explanations are produced?	Research on data, Explore models, Debug models	responsiBle models, Legal issues, trUst in predictions, Ethical issues
When explanations are read and used?	Empower model developer, mostly during training	Empower user, mostly during model inference
Who is the direct audience of the explanations?	Power user, Model developers, AI researchers	Lay user, Customer, Patient
What are desired characteristics of explanations	Faithful to model and data, Actionable	Simple and easy to understand

Red Teaming with GenAI

CVPR 2024

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 2 Apr 2024]

Red-Teaming Segment Anything Model

Krzysztof Jankowski, Bartlomiej Sobieski, Mateusz Kwiatkowski, Jakub Szulc, Michal Janik, Hubert Baniecki, Przemyslaw Biecek

Foundation models have emerged as pivotal tools, tackling many complex tasks through pre-training on vast datasets and subsequent fine-tuning for specific applications. The Segment Anything Model is one of the first and most well-known foundation models for computer vision segmentation tasks. This work presents a multi-faceted red-teaming analysis that tests the Segment Anything Model against challenging tasks: (1) We analyze the impact of style transfer on segmentation masks, demonstrating that applying adverse weather conditions and raindrops to dashboard images of city roads significantly distorts generated masks. (2) We focus on assessing whether the model can be used for attacks on privacy, such as recognizing celebrities' faces, and show that

Bonus 1: Adversarial attacks for SAM

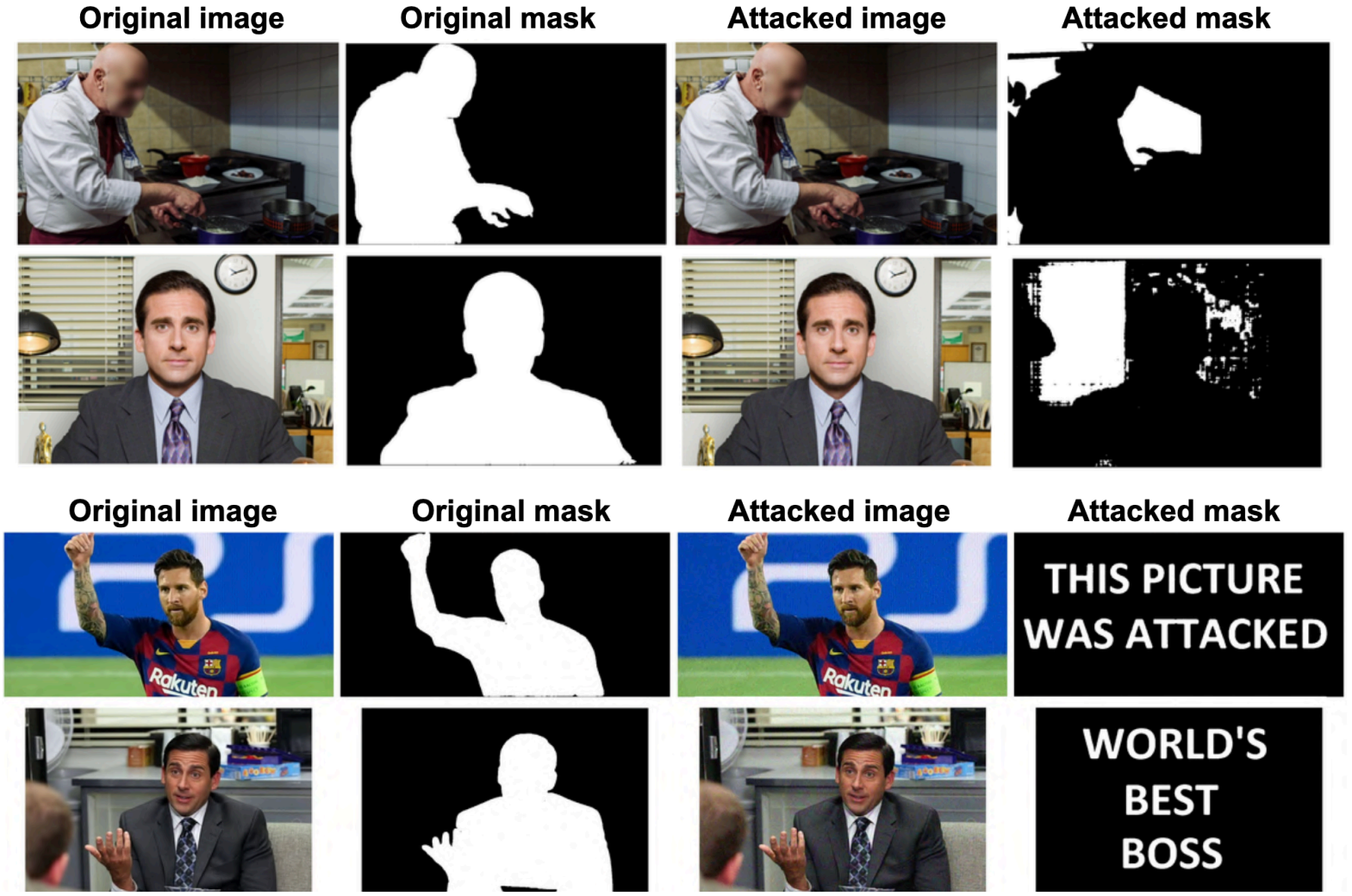


Figure 6. Examples of targeted attacks, generated using an FGSM-based approach. Masks can be changed to an arbitrary text.

Red Teaming Vision Models

ICLR 2024

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 12 Mar 2024 (v1), last revised 14 Mar 2024 (this version, v2)]

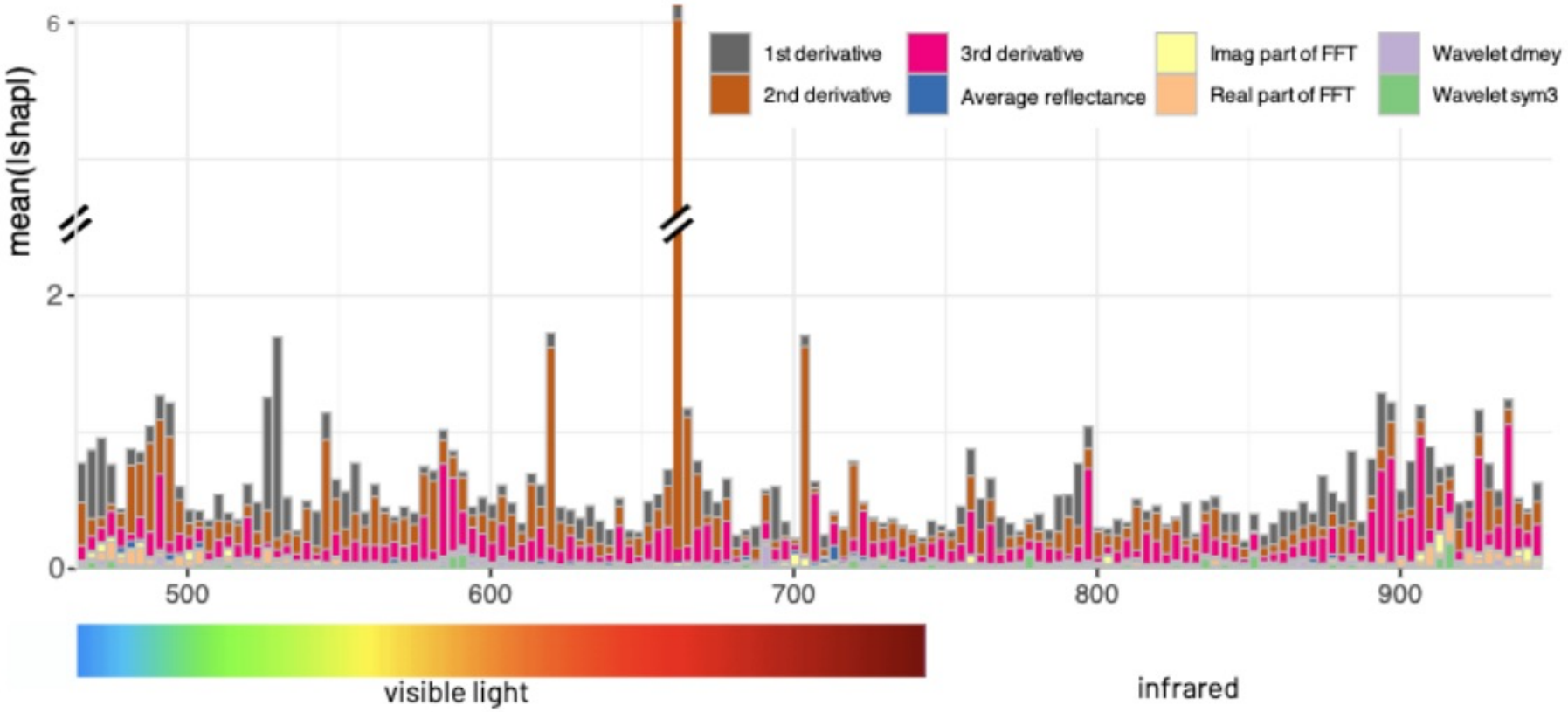
Red Teaming Models for Hyperspectral Image Analysis Using Explainable AI

Vladimir Zaigrajew, Hubert Baniecki, Lukasz Tulczyjew, Agata M. Wijata, Jakub Nalepa, Nicolas Longépé, Przemyslaw Biecek

Remote sensing (RS) applications in the space domain demand machine learning (ML) models that are reliable, robust, and quality-assured, making red teaming a vital approach for identifying and exposing potential flaws and biases. Since both fields advance independently, there is a notable gap in integrating red teaming strategies into RS. This

Aggregation Analysis

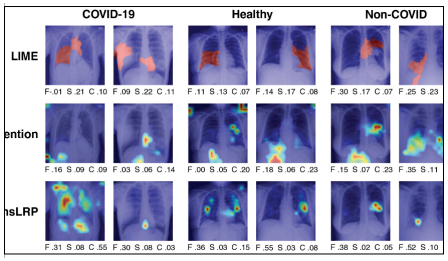
The ability to aggregate Shapley values (<n_samples, features, class>) by hyperspectral bands and data transformation groups enabled a **richer exploration**, as presented in Figure 4. This plot shows that key features are distributed across various bands rather than concentrated in specific areas.



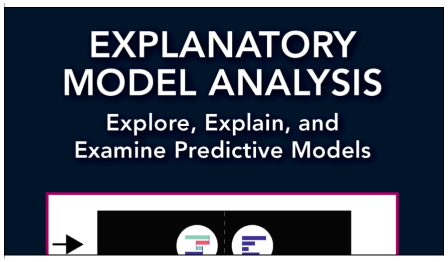
MI2.AI is here to fix AI

MI REDTEAM

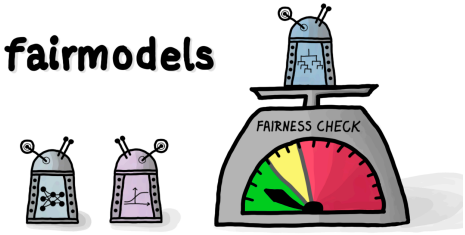
Papers



Books



Software



Teaching