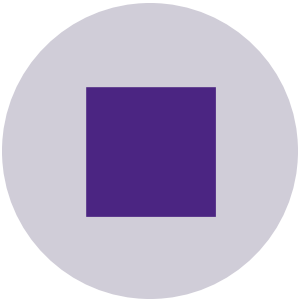
Object-detection for content moderation with xAI

David Pichler
Timo Rodinger

Background and Rationale

DIGITAL SERVICES ACT

SOCIAL MEDIA PLATFORMS



BLACK BOX AND LACK OF TRANSPERANCY

Research question

- Can the implementation of Explainable AI techniques effectively improve content moderation with object detection systems by providing transparent insights into decision-making processes and enhancing moderators‘ ability to assess detection outcomes?
  - *Yes, but…*

    - RQ1.1: Can state of the art ML / DL models effectively be used to help with violent content moderation in social media? - *Yes*
    - RQ1.2: Can these SOTA Models be used together with xAI and Captum? - *Generally yes, but…*
    - RQ1.3: How do different SOTA ML image detection models differ in their explainabilty from each other? Which models are better explainable than others? - *Future Research*

Results

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

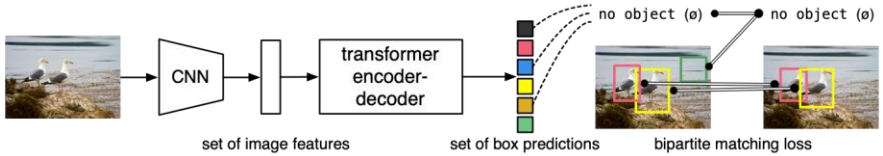
EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

Models & Libraries

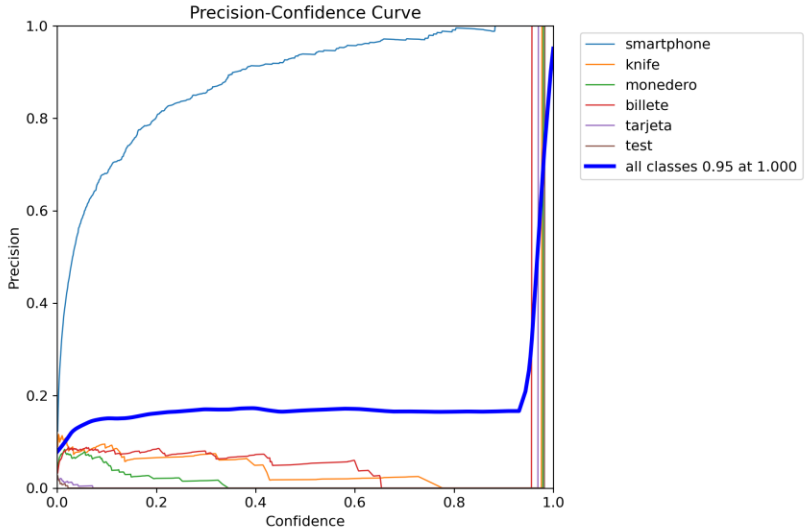
ultralytics
YOLOv8



Captum

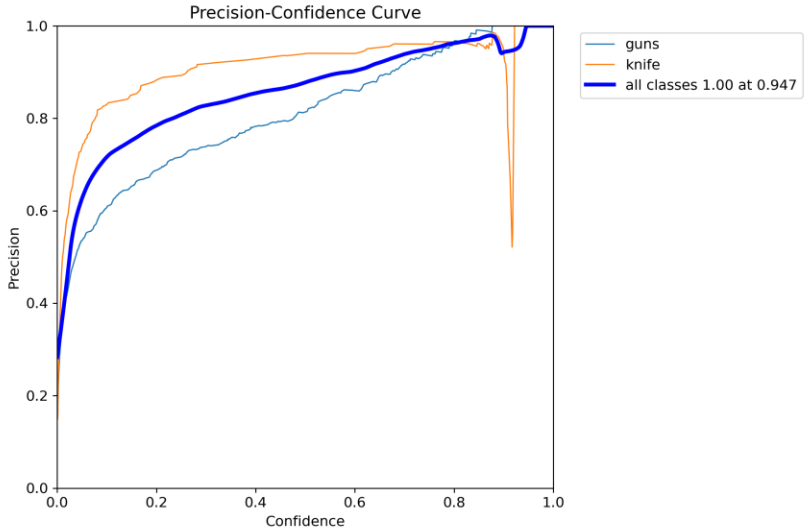
Self-Trained Yolov8

- Dataset based on weapons and similar data
- Good at identifying smartphones
- Bad at every other category
- Overall Precision of 0.2



Pre-Trained model

- Only 2 classes
- Guns and knifes
- Way better performance
- More fitting for our use case



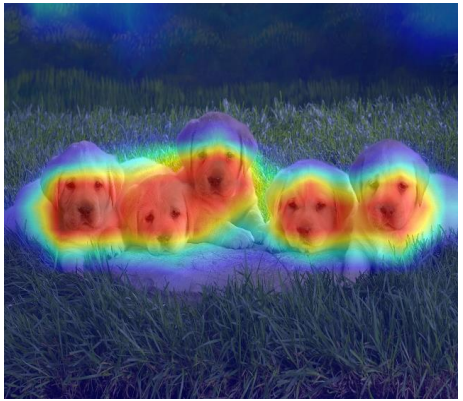
Dataset & Model Classes

- **Sohas_weapon-Detection** *(Our model)*
  - 5000 training images
  - 6 classes
    - Pistol
    - Knife
    - Smartphone
    - Bill
    - Purse
    - Card
- **Weapon 2 Computer Vision Project** *(Pre-Trained Model)*
  - 4098 training images
  - 2 classes
    - Knife
    - Gun

Why does the pre-trained model work better?
Assumption: 6 vs 2 Classes

xAI with Eigen Grad-CAM

- Highlight the most critical features -> PCA
- Visualize the important regions of an input image that contribute to the CNN's output
- Uses PCA to reduce the dimensionality of the gradients, potentially highlighting the most significant features more effectively



Our final result

Standard Model

Our self-trained model

Our final result

Pre-Trained Weapon Model

Eigen Grad-CAM

What would be our next steps…

- Improve object / weapon detection model
  - Test Different Frameworks (e.g. DETR, Yolov10)
  - Better Data
- Compare different explainability techniques
  - Grad-CAM ++
  - Augmented Grad-CAM
- Try it on social media
  - Real-time detection
- Usage in Security Context
  - Real-time security footage monitoring

More Examples

WU
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

EFMD EQUIS ACCREDITED
AACSB ACCREDITED
ASSOCIATION AMBA ACCREDITED

Hurl

Standard Model

Our model

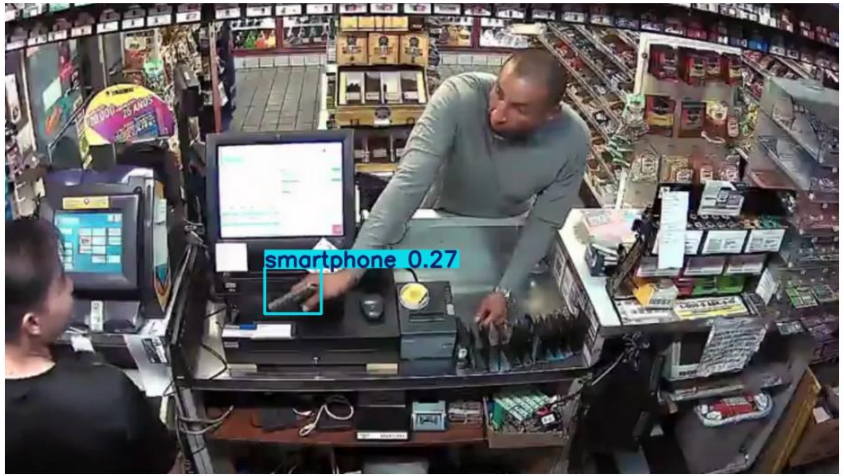
Hurl

Weapon Model

Grad-CAM

Robbery

Standard Model



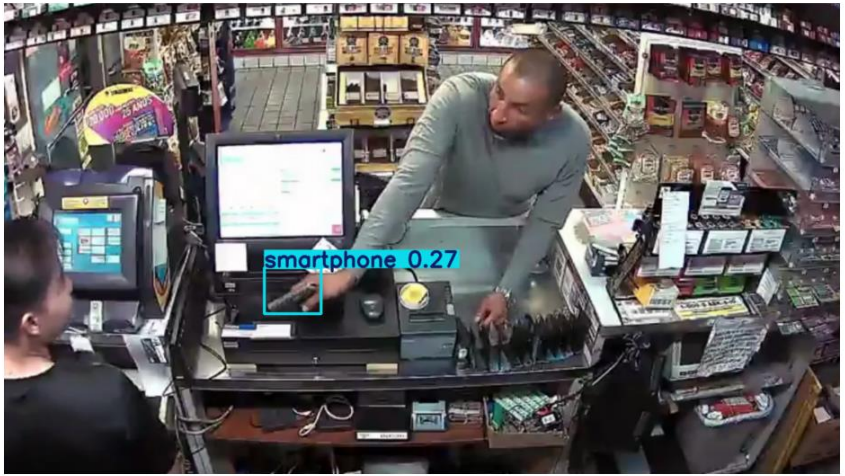
Our model

Robbery

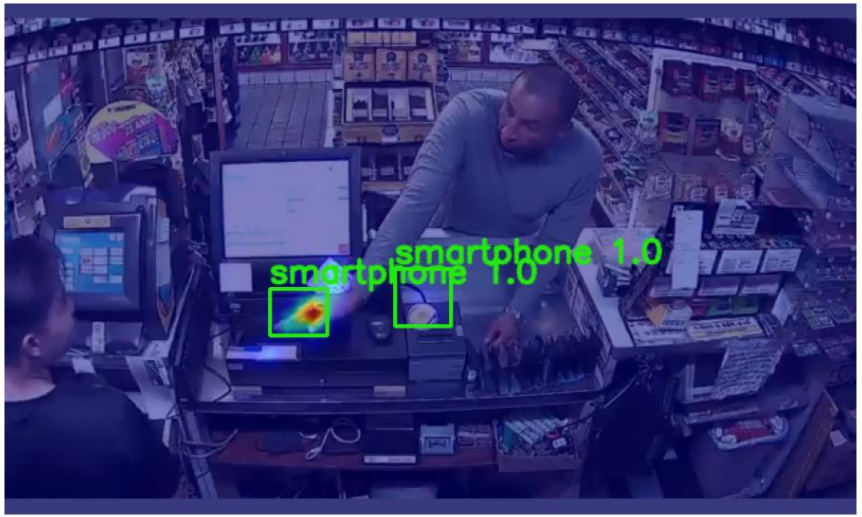
Weapon Model

Grad-CAM

Robbery



Our Model



Our model GradCam

Demo

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED