GAUNTLET: An Explainable Model for Detecting AI-Generated Images

Ethics in Artificial Intelligence

Presented by Angelo Galavotti, Lorenzo Galfano

Università di Bologna, December 2024

The negative implications of AI

Generative AI had a rapid improvement over the last few years, which also lead to **AI-powered fraud**:

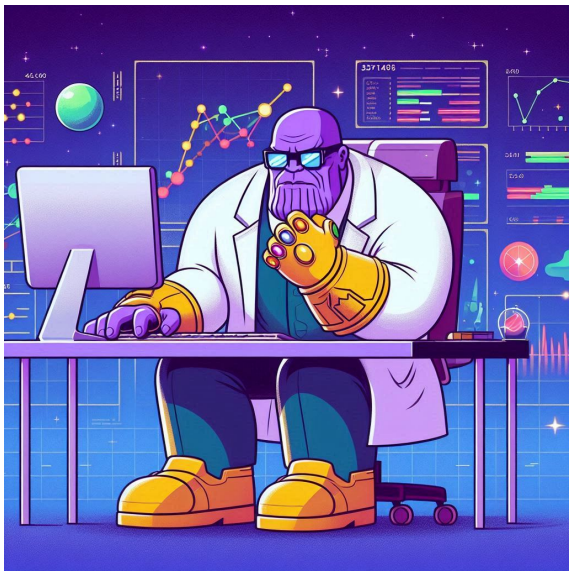
- Misinformation
- Identity fraud
- Fraudulent investment schemes

It's a significant concern since AI is now **more accessible** than ever. Nearly 22%[1] of younger individuals **fall for AI-enhanced fraud**, despite 73% of respondents expressing confidence in their ability to identify such scams.

[1] https://sift.com/index-reports-ai-fraud-q2-2024?aliId=

GAUNTLET

GAUNTLET aims at **detecting AI-generated images** while emphasizing **explainability**.
Our goal is not just to provide a tool for AI-image detection: we also want to **educate** the users and equip them with the knowledge to identify such images without necessarily relying on the model.



Model

The target model used was a **ResNet50**, which provided:

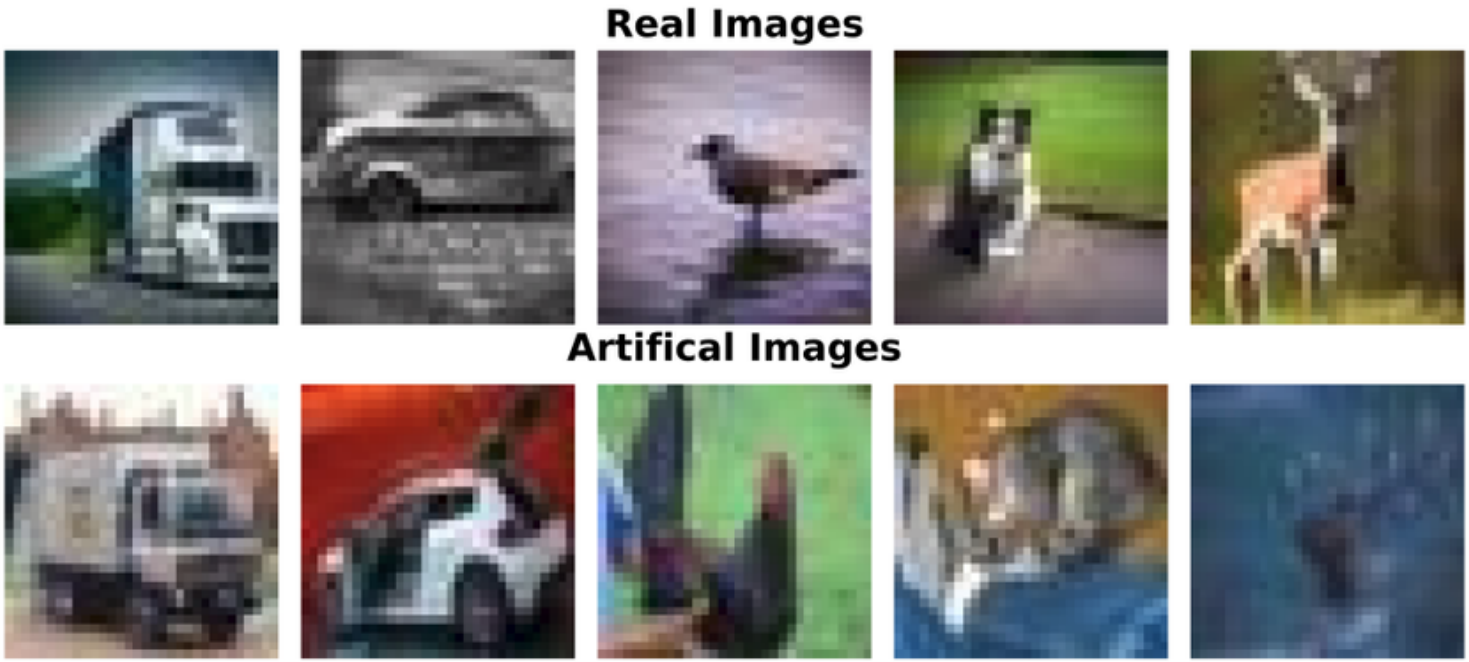
- Strong feature extraction capabilities.
- Excellent starting point for transfer learning.

As a baseline, we used a CNN comprised of:

- 3 convolutional layers with increasing channel sizes (32, 64, 128)...
- ...each followed by batch normalization, ReLU activation and max pooling.

CIFAKE Dataset

- Composed of 120k samples.
- Equally divided, half are real (from **CIFAR10**) and half are generated via **Stable Diffusion 1.4**.
- 32x32 resolution.



RVAA Dataset

- Composed of 975 samples.
- **Extremely heterogeneous**, generated via Midjourney, Stable Diffusion and even GANs.
- Significant variability in resolution, composition and model used to generate images

(a) Real image in RVAA

(b) Fake image in RVAA

MIXED Dataset

- **Mixture** of both CIFAKE and RVAA.
- We sampled about 1% of CIFAKE dataset (1200 samples).
- CIFAKE images were upscaled to match RVAA resolution.
- Designed to evaluate whether the model could learn and **generalize features** from **both dataset**.

Explainability techniques

The explainability techniques used in GAUTLET aim at **outcome explanation**.
Specifically, the technologies we adopted are:

- ScoreCAM
- AblationCAM
- LIME

Explainability techniques - GradCAM variants

ScoreCAM and AblationCAM are variants of **Grad-CAM**, which developed many variants throughout the years, creating something similar to a "family" of explainability techniques.

- By leveraging the gradients of a target class w.r.t final conv layer, it generates **heatmaps** to highlight regions of interest.

(a) Original Image

(b) Heatmap Overlay

Explainability techniques - GradCAM variants

Unlike GradCAM, which operates by working with gradients:

- **Score-CAM** utilizes the model's **score** as a measure of importance for each activation map. It **perturbs the input** and observes how each activation is contributing to the model's confidence in its prediction.
- **Ablation-CAM** operates by systematically ablating (i.e. **removing** or disabling) **specific neurons** in the feature maps of the convolutional layer and observing the impact on the model's output.

For these reasons, they produce less noisy outputs and are easier to integrate with GAUNTLET's evaluation pipeline.

- Cons: more costly!

Explainability techniques - LIME

On the other hand, LIME works as follows:

- **Perturbs input data**, by turning on and off some of the **super-pixels** (i.e. group of neighbouring pixels with similar characteristics) of the image.
- Observes changes in the model's output to determine which parts of the input contribute most to the prediction.
- Uses said super-pixels to then **segment** the image and evaluate the importance of each region by modifying them.

It is important to note that LIME is model agnostic.

Training setup

The ResNet50 model was trained using the "Linear Probing followed by Fine-Tuning" technique. It consists in a 2-step fine-tuning method:

- First we train a **new classification head**, while keeping the feature extractor **frozen**.
- Then we **unfreeze** the feature extractor and train the network again.

Training setup - Hyperparameters

Some specifications on the hyperparameters used:

- ADAM optimizer for RVAA and MIXED with lr of $3 * 10^{-4}$, ADAMW for CIFAKE.
- Different batch sizes for the datasets, 128 for CIFAKE, 32 for RVAA and 64 for MIXED.
- These hyperparameters were chosen after applying Grid Search.
- This configuration was used for both ResNet50 and the baseline.

Training setup - Data augmentation

We experimented with **data augmentation** on the RVAA dataset, and the RVAA portion of the MIXED dataset. To augment the data, we performed the following transformations:

- We applied a random rotation of 90 degrees at most.
- We randomly flipped the image.

Results in numbers - CIFAKE

For each experiment we evaluated the model based on accuracy, precision, recall and F1-score.

Models	Accuracy	Precision	Recall	F1-Score
CIFAKE_baseline	95.5	0.94	**0.97**	0.95
CIFAKE_ResNet-50_step1	82.7	0.83	0.83	0.83
CIFAKE_ResNet-50_step2	**98.0**	**0.97**	**0.97**	**0.97**

Table: Results on the CIFAKE dataset. The 2-step fine-tuning strategy achieves a very high accuracy score.

Results in numbers - RVAA

For each experiment we evaluated the model based on accuracy, precision, recall and F1-score.

Models	Accuracy	Precision	Recall	F1-Score
RVAA_baseline	76.7	0.73	0.73	0.73
RVAA_ResNet-50_step1	78.8	0.76	0.75	0.75
RVAA_ResNet-50_step2	81.5	**0.87**	0.74	0.80
RVAA_AUG_ResNet-50_step1	80.1	0.79	0.75	0.77
RVAA_AUG_ResNet-50_step2	**86.3**	0.85	**0.83**	**0.84**

Table: Results on the RVAA dataset. Data augmentation is beneficial to the network, and allow us to leverage the full potential of the dataset.

Results in numbers - MIXED

For each experiment we evaluated the model based on accuracy, precision, recall and F1-score.

Models	Accuracy	Precision	Recall	F1-Score
MIXED_baseline	83.1	0.92	0.86	0.89
MIXED_ResNet-50_step1	85.5	0.91	0.89	0.90
MIXED_ResNet-50_step2	90.3	**0.97**	0.90	0.93
MIXED_AUG_ResNet-50_step1	82.3	0.84	0.93	0.88
MIXED_AUG_ResNet-50_step2	**91.4**	0.93	**0.95**	**0.94**

Table: Results on the MIXED dataset. Data augmentation appears to slightly improve performance.

Qualitative results

Moving into qualitative results, the results showed:

- ScoreCAM and AblationCAM generally gave similar outputs, however **ScoreCAM** provided heatmaps that appeared **more aligned** with the model's decision-making process.
- The **resolution** impacted greatly the result: non-upscaled CIFAKE images generated broken or low-quality explanations in some cases.
- Based on the layers targeted by ScoreCAM and AblationCAM, the heatmaps may be **more defined**, but **less informative**.
- LIME explanations were significantly less intuitive, probably because it's best suited for a multi-category scenario.

Qualitative results

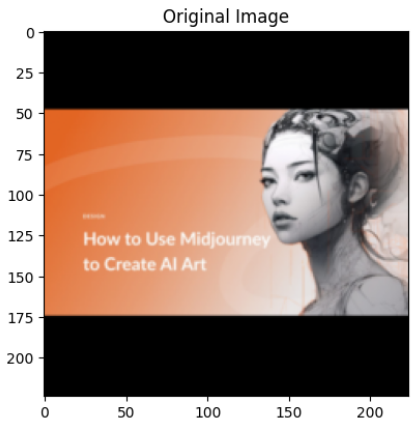


Figure: Original input

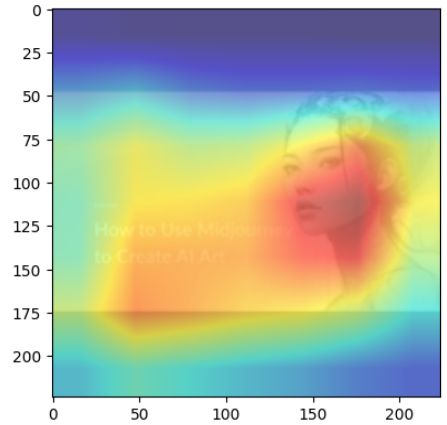


Figure: ScoreCAM on layer 4

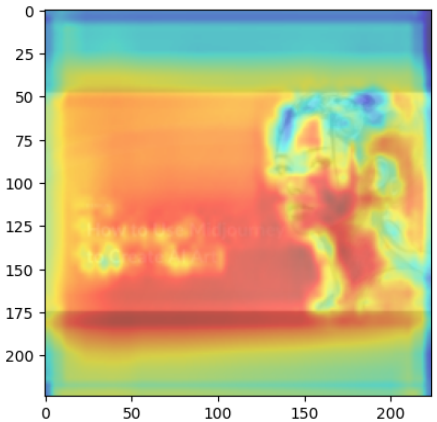


Figure: ScoreCAM on layers 2, 3, and 4

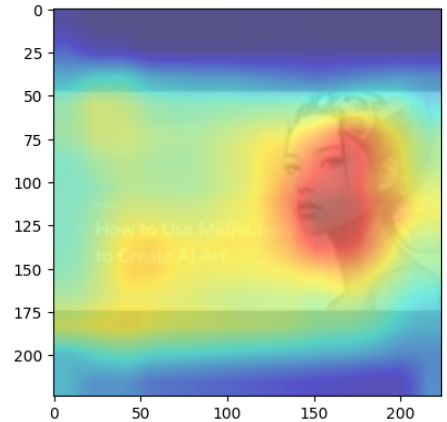


Figure: AblationCAM on layer 4

Web application

We developed a Web App[2] to showcase a use case of GAUNTLET:

- Built using React and Flask technologies.
- Provides a set of explanation methods that the users can choose (i.e. ScoreCAM, AblationCAM, LIME).
- Processes the input through the selected model and returns a **comprehensive visualization** using the selected explainability technique, along with **confidence scores** for the REAL and FAKE classes.

[2] A video demo of the web application can be found at the url `https://www.youtube.com/watch?v=THj-Gn8MYkw`

In conclusion

To wrap up:

- The **finetuned ResNet** allowed us to achieve **high scores** in terms of both accuracy and f1, and leverage the full potential of the datasets.
  - The RVAA dataset provides lowest performance in terms of accuracy.
- Visualization methods performed poorly on CIFAKE, probably due to the small resolution of images.
  - Upscaled CIFAKE showed more interpretable results.
- After testing, ScoreCAM on layer 4 seems to produce the **best quality explanations**.
- In the context of this project, AblationCAM and ScoreCAM provided much more interpretable visualizations than LIME.

In conclusion - Future Works

Future works could be concentrated on:

- Testing different kinds of network architectures $\rightarrow$ **CLIP** (Moskowitz et al., 2024).
- Using a better quality dataset $\rightarrow$ **WildFake**.
  - 3M images, from a wide range of generative models.

These directions, however, would require **significantly higher computational resources** and longer training times, given the increased complexity of the models and datasets involved.

Thank you!

Appendix: experiments - ModResNet and Dropout

During the development process, we experimented with a variation in the ResNet architecture and different techniques. However, we were not satisfied with the results obtained:

Models	Accuracy	Precision	Recall	F1-Score
ModResNet_step1	85.6	0.86	0.85	0.86
ModResNet_step2	97.2	0.97	0.97	0.97
ModResNet_Dropout_step1	85.3	0.88	0.83	0.86
ModResNet_Dropout_step2	96.9	0.97	0.97	0.97

Table: Additional experiments done with the CIFAKE dataset.

Appendix: experiments - CIFAKE on RVAA

Models	Accuracy	Precision	Recall	F1-Score
CIFAKE_UPSCALED_step2	97.8	0.98	0.98	0.98
CIFAKE_UPSCALED_step2 on RVAA	55.4	0.02	0.53	0.04

Table: Results of the model trained on the upscaled CIFAKE on the RVAA dataset.

Appendix: LIME results

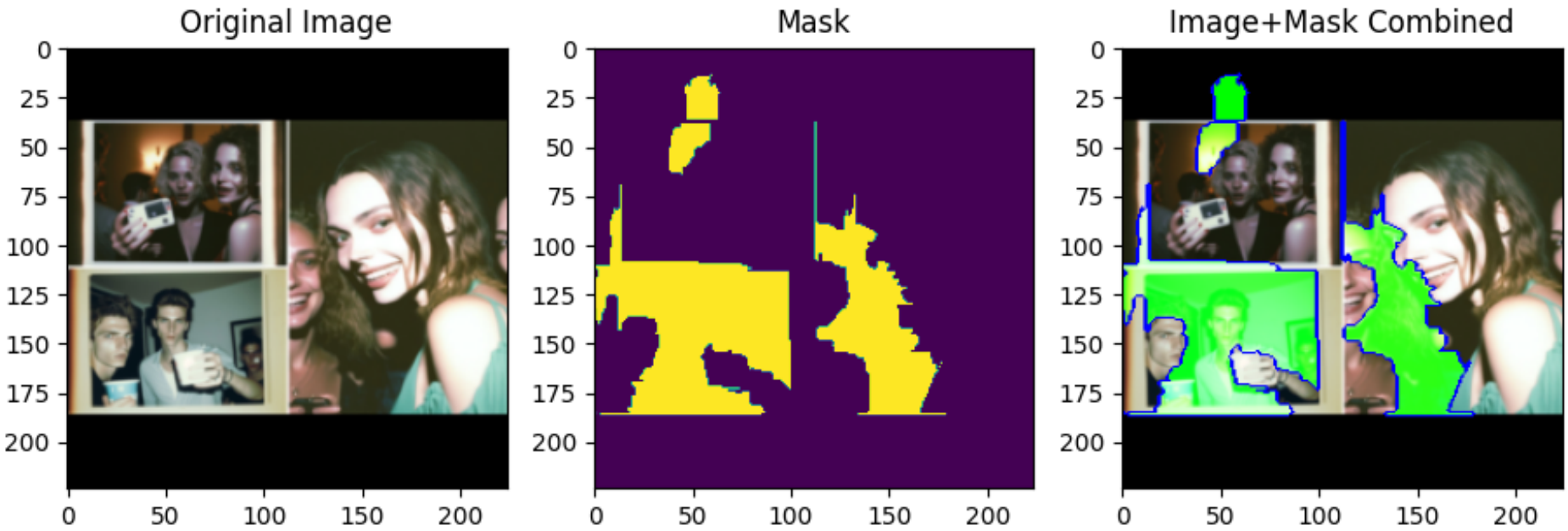


Figure: Example of LIME results