Towards Targeted Obfuscation of Adversarial Unsafe Images using Reconstruction and Counterfactual Super Region Attribution Explainability

Authors: Mazal Bethany¹, Andrew Seong¹, Samuel Henrique Silva¹, Nicole Beebe², Nishant Vishwamitra², and Paul Rad¹

¹Secure AI and Autonomy Lab, ²Department of Information Systems and Cyber Security
University of Texas at San Antonio

USENIX Security 2023
Disclaimer

- This presentation contains discussions on harmful image content, such as sexually explicit, cyberbullying, and self-harm images that are highly offensive and might disturb the readers.
Adversarial Unsafe Images

- **Adversarial Images**: Deceptive digital images that fool AI-based image recognition systems, causing misclassification, while appearing unchanged to human viewers.

- **Unsafe Images**: Potentially harmful or offensive content requiring effective detection and moderation to protect viewers.
Detection of Adversarial Unsafe Images

• Small perturbations can fool AI based detectors while preserving visual semantic content

Detected as cyberbullying
Detection of Adversarial Unsafe Images

- Small perturbations can fool AI based detectors while preserving visual semantic content

Detected as cyberbullying

$\epsilon^*$
Detection of Adversarial Unsafe Images

- Small perturbations can fool AI-based detectors while preserving visual semantic content.
Adversarial Unsafe Images

- Adversarial perturbations compound the issue of unsafe images
- Frequent exposure to unsafe images can cause harm to image reviewers
- Moderator lawsuits for mental damages
How Do Existing Methods Perform?

• With adversarial attacks the detection performance drops almost 40% on average across state-of-the-art API
How Do Existing Methods Perform?

• With adversarial attacks, the detection performance drops almost 40% on average across state-of-the-art API

Problem 1: Existing methods are insufficient against adversarial unsafe images

Need a new method that can remove perturbations
Explainability Based Image Obfuscation

- Image obfuscation for protecting reviewers of sensitive images
- Grad-CAM, LIME, Integrated Gradients

Problem 2: Existing explanation methods are unsuitable for image obfuscation

Need new obfuscation methods that are suitable for obfuscation
Motivation Overview

Problem 1: Existing methods are insufficient against adversarial unsafe images

Problem 2: Existing explanation methods are unsuitable for image obfuscation

Reconstruction to remove adversarial perturbations

Counterfactual super region attribution explainability to obfuscate
Datasets

- Sexually Explicit [1]
- Cyberbullying [2]
- Self-Harm
  - Self-harm (self-cutting, self-bruising, eating disorder, depicted or promoted self-harm) (2,100 images)
  - Non-self-harm (neutral social media images) (4,200 images)

System Design Intuition

• Reconstruction of Adversarially Perturbed Image with Robust Classifier
  • Image Reconstruction to remove the perturbations as an input transformation defense
  • Robust classifier with adversarial training to detect unsafe content

• Obfuscating Unsafe Content with Counterfactual Explainability
  • Explainability to detect the unsafe parts of the image and obfuscate them
Approach Overview

• uGuard (unsafe image Guard)

  • Image reconstruction module: Adaptive Clustering of Robust Semantic Representations (ACSR)

  • Explainability-based image obfuscation module: Counterfactual Super Region Attributions (CSRA)
uGuard Robust model

• Training a robust image detection model
  • Adversarial attacks pushes image to tail of training distribution
  • Standard adversarial training:
uGuard Image Reconstruction

• **Removing Adversarial Perturbation**
  • Reconstruct high-frequency component of image
  • Decompose images into high and low frequency components using the Tikhonov filter
  • Convolutional Dictionary Learning to learn a dictionary from clean (unattacked) images to reconstruct the high frequency component of an image from the low frequency component

\[
x_{rec} = x_{low} + x_{high}^{rec}
\]

\[
\arg \min_{x_{low}} \frac{1}{2} \| x_{low} - x \|^2_2 + \frac{\lambda}{2} \sum_j \| G_j x_{low} \|^2_2
\]

\[
x_{high}^{rec} \approx Dr = d_1 r_1 + \cdots + d_M r_M
\]
uGuard Explainability Based Image Obfuscation

- **Targeted Image Obfuscation**
  - Counterfactual examples
  - $2^K$ different combination of regions to potentially mask
  - Attribution maps point us to likely regions to sample from

![Original Image](image1.png)
![Attribution Map](image2.png)
![Superpixel Segmentation](image3.png)
![Average Region Scoring](image4.png)
![Ideal Obfuscation Determined by CSRA](image5.png)
uGuard Explainability Based Image Obfuscation

- Split an image into regions
- Generate attribution map
- Average attribution scores within each segment
- Perform counterfactual analysis of top K scored segments to determine a combination of segments to obfuscate
uGuard System Architecture

Reconstruction Subsystem
- Tikhonov Filter
  - $X_{high}^{low}$
  - High Frequency Features
  - Low Frequency Features

Obfuscation Subsystem
Using Counterfactual Super Region Attribution (CSRA) Explainability
- Grad-CAM Attribution Map
- SLIC Segmentation
Identify and Block Unsafe Regions using Counterfactual Super Region Attribution

Train Robust Classifiers to Generate Semantic Features
- NSFW Images
- Cyberbullying Images
- Self-Harm Images

Reconstruct High Frequency Features using CDL
- $X_{rec}^{high}$
Evaluation: Public API vs uGuard

- Public API are unable to perform targeted obfuscation, and perform worse on adversarially perturbed unsafe images.

<table>
<thead>
<tr>
<th></th>
<th>Public API</th>
<th>uGUARD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adversarially Perturbed Accuracy %</td>
<td>45.60</td>
<td>88.07</td>
</tr>
<tr>
<td>% Adversarially Perturbed Images Obf. to be Safer</td>
<td>96.67</td>
<td>99.50</td>
</tr>
<tr>
<td>Obfuscation %</td>
<td>27.00</td>
<td>13.37</td>
</tr>
</tbody>
</table>

| Class                |  |  |  |
|----------------------|------------------|------------------|
| Sexually Explicit    | 45.60            | 88.07            |
| Cyberbullying        | N/A              | 95.36            |
| Self-Harm            | N/A              | 90.07            |
|                      | 96.67            | 94.67            |
|                      | 27.00            | 14.00            |
Additional Evaluations

• Adversarial robustness
  • Robustness to seen attacks and some unseen attacks

• Explainability-based obfuscation
  • More images made safer, with less obfuscation overall
  • Preserves more important context than other techniques

• In-the-wild Experiment
  • Human evaluations on sexually-explicit and self-harm images
  • Over 90% of unsafe images made safer
Future Work

• Other unsafe image categories
• Investigating using targeted obfuscation methods in conjunction with Vision Language Models to assist in protecting social media image moderators
Conclusions

• We investigated adversarial unsafe image detection systems and explainability based obfuscation of unsafe images

• State-of-the-art systems that detect unsafe image content are vulnerable to adversarially attacked images

• We presented uGuard to detect and perform targeted obfuscation of adversarial unsafe images across three datasets

• Our evaluations showed that uGuard was able to sufficiently detect and obfuscate adversarially unsafe images
Q&A