



CSE 509 Research Project Presentation

Team: Security Sheriffs

Members: Neelesh, Jagadeesh, Kamalnath



What is NSFW?

- Not Safe For Work
- How?
- Consists of Images in these categories:
 - Explicit Adult Content
 - Violent or Gore Content
 - Hate Speech and Harassment
 - Drug and Substance Abuse
 - Explicit Language
 - Graphic or Disturbing Images
 - Sensitive or Triggering Material

A large red diamond-shaped warning sign with the text "NSFW" inside. The sign is oriented with its vertices at the top, bottom, left, and right. The text "NSFW" is written in a bold, black, sans-serif font, centered within the diamond. The red border of the diamond is thick and prominent.

NSFW



Project Objective?

- Proliferation of Social Media platforms with huge, diverse user base
- These platforms use content moderation tools to filter the content that an end user can see
- They work!
- Our goal is target these moderation tools and disguise NSFW images as normal image!



What is our **Motivation**?

- The research aims to conduct attacks on these systems to reveal vulnerabilities, evaluate resilience, and propose enhancement strategies.
- There's a lack of research on NSFW-based adversarial attacks, motivating this study to understand and address NSFW detector limitations.
- Ultimately, the goal is to contribute to the development of more robust content moderation mechanisms, creating safer and respectful online environments.





Our Contributions

- An NSFW clean dataset
- A new white-box and black-box attack
- First to use Grad-CAM for adversarial attacks specifically



Data Collection

- Scraping Images from reddit forums
- Removing Deleted Images, Duplicates and corrupted images from dataset
- Using GCP's Detect explicit content (SafeSearch) API to classify images downloaded
- Ended up with 100K images and PTSD!





How to attack Content Monitoring tools?

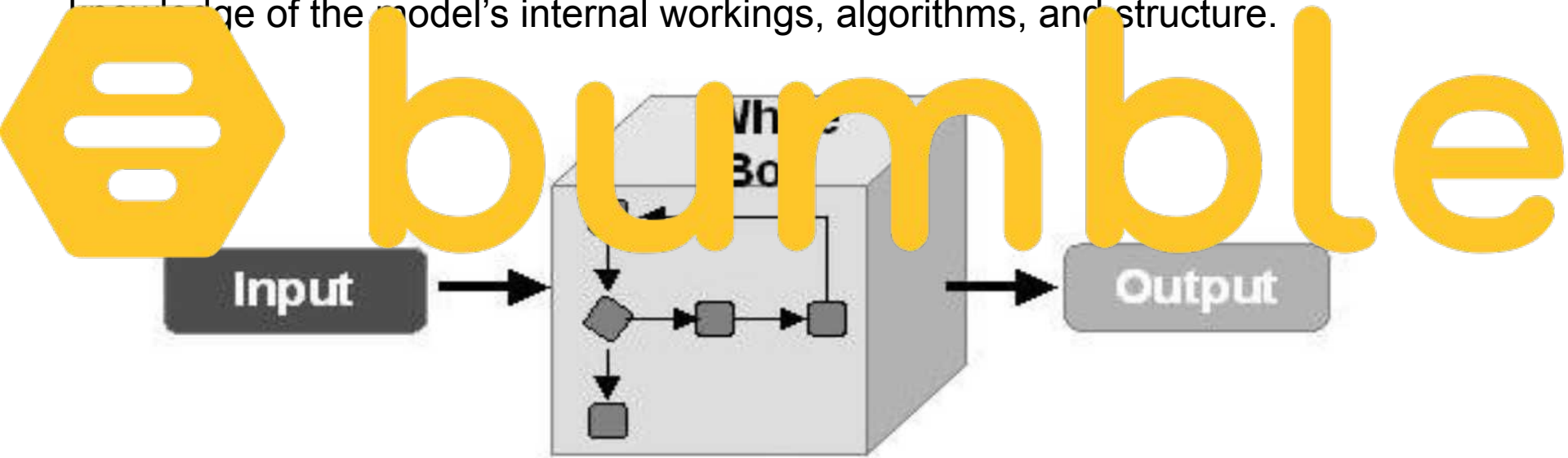
There are two types of attacks we can do here:

- WhiteBox Attack
- BlackBox Attack



What is WhiteBox Attack?

White-box attacks involve testing vulnerabilities by gaining full access and knowledge of the model's internal workings, algorithms, and structure.



As part of our project we have attacked...



Bumble Attack

Bumble has [released](#) an **Open-Source** version of their Private **Detector** AI Feature to help tech community **Combat Cyberflashing**

The developers have trained the model in tensorflow and provided the model weights, but we copied the architecture into PyTorch and trained on our dataset

We used the GradCAM tool to generate adversarial images for pytorch model

Stephen (Senior ML Scientist at Bumble) helped us at times, the findings will be shared with them to improve their systems.

Stephen O' Farrell
[Steeeeephen \(Ste\) \(github.com\)](#)





Bumble Attack Results

There are 2 different models used in Bumble platform, one for profile pictures and other for in chat media. We targeted in chat media model.

Out of 15 adversarial images, only one got tagged as NSFW and the rest were undetected.

Control Images:

3 NSFW images were added and all were detected

3 Normal Images (non NSFW/SFW)

The results are at the end of ppt



What is BlackBox Attack?

Black-box attacks involves testing the vulnerabilities of these detection systems without direct access or knowledge of their internal workings, relying solely on input-output observations to identify weaknesses.



We chose to target Reddit for this attack, as it is more forgiving than Facebook.





Our Methodology

- Grad-CAM is a “visual explanation” tool
- Tells us which pixels/regions a model is giving importance to
- Hypothesis
 - Different models for the same task look at similar image regions
 - Does not depend whether we use same dataset for training/testing of all models
 - If there are some “important region(s)” in an image, we can add noise gradually to the region(s)
 - There will be a time when this noisy image will start getting misclassified
 - Need to control the denoising as well



Selection of Models

- To cover different types of models that have been introduced, we selected 4 models
 - Resnets (4 types - 18, 34, 50, 101)
 - InceptionNet (v2, v3)
 - Vit (base model)
 - VGG16
- The models have been introduced at different times (VGG 2014 to Vit 2020)



Selection of Models Cont.

- Grad-CAM outputs for some of these models
- Number of regions produced can be more than one
- 3 of these 5 are pretty similar (how similar?)



Inception-v3



Resnet-34



Resnet-50



Resnet-101



ViT



Similarity metrics based on Earth Mover Distance

- We took a sample of 100 images (grad cam outputs) for 5 models and computed mean EMD distances

	Inception-v3	Resnet-34	Resnet-50	Resnet-101	ViT
Inception-v3	0				
Resnet-34	0.0023	0			
Resnet-50	0.0063	0.0014	0		
Resnet-101	0.0041	0.0035	0.0011	0	
ViT	0.0087	0.0063	0.0049	0.0061	0



White-Box Attack Cont.

- Algorithm
 - Train a model M on the dataset D (NSFW + SFW Images)
 - Take an image $I^{H \times W \times 3} \sim D$
 - Get the grad-cam output of this image $G = \text{gradcam}(I)$
 - Assume G has N number of regions (disconnected red regions)
 - Initialize N gaussians with mean $\mu_1, \mu_2 \dots$ and variance $\sigma_1, \sigma_2, \dots$
 - At these locations, apply these noises in a patch of height and width x
 - Apply these gaussians to the image to get the new image $I'(\mu, \sigma, x)$



White-Box Attack Cont.

- Algorithm

- Assuming that the original label is y , the loss function corresponding to this image is

$$L(M(I'(\mu, \sigma, x)), y)$$

- We want to maximize this loss such that the added noise remains small

- $\|I' - I\|_{\infty} < \epsilon$ (Allowable noise)

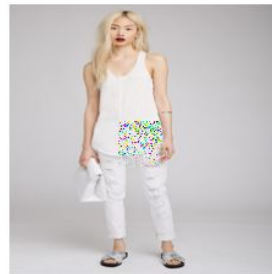
- Final equation becomes

$$\arg \max_{I'} \left(L(M(I'(\mu, \sigma, x)), y) \right) \text{ such that } \|I' - I\|_{\infty} < \epsilon$$



White-Box Attack Cont.

- Such optimization equations can be solved using PGD (implementations available and heavily used)
- For simplicity we used 2 Gaussians only



Progression of noise on a normal image for initial steps



White-Box Attack Accuracy

	Inception-v3	Resnet-34	Resnet-50	Resnet-101	ViT
Inception-v3	78%				
Resnet-34		87%			
Resnet-50			81%		
Resnet-101				88%	
ViT					81%

These are for 100 images



Black-Box Attack

- We now know that heat-maps of grad-cam are important
- Based on the similarity scores we saw earlier, we may say that for any other “classification model”, heat-maps will be similar
- So we can use these heat-maps for black-box attacks too



Black-Box Attack Accuracy

	Inception-v3	Resnet-34	Resnet-50	Resnet-101	ViT
Inception-v3		50%	41%	39%	44%
Resnet-34	17%		19%	12%	10%
Resnet-50	38%	53%		46%	50%
Resnet-101	56%	61%	56%		56%
ViT	41%	47%	51%	48%	

These are for 100 images



Black-Box Attack for Reddit

- Algorithm

- Train multiple models M_1, M_2, \dots
- For any image I , generate grad-cam heatmaps G_1, G_2, \dots
- Take the intersection of these heatmaps $G_1 \cap G_2 \cap \dots$
- Using the final heatmap, do the same procedure as in White-box attack

$$\arg \max \left(L \left(M \left(I'(\mu, \sigma, x) \right), y \right) \right) \text{ such that } \| I' - I \|_{\infty} < \epsilon$$

- M is **Resnet101** (ensemble can be used)



Results

- We have been able to post image on Reddit not tagged as NSFW
- Because of multiple queries for initial days, our accounts keep getting banned (either due to spamming or NSFW pics)
- Hence, not able to do an extensive amount of testing
- Out of the 10 images we prepared, 2 were passed



Possible reasons of some failures

- At the end of the day, Reddit might not be doing classification at all, it might be doing detection
- Classification attacks may not work on detection



FGSM attack working on classification but not on detection



Problems and Future Work

- Optimization equation
- Gaussian noise vs Random noise
- Better way to handle heatmaps ?
- Generalization of attacks for detection and classification
- A detection dataset of NSFW (Nudenet)

Failed Attempts/Attacks

- Segment&Complete implementation (WIP)
- Adversarial Patch method



Results


If Approved, the results will be posted here



QUESTIONS

Will be taken now



A man in a dark suit, white shirt, and patterned tie stands in an office. He is positioned in front of a glass door with horizontal blinds. To his left is a window with closed horizontal blinds and a black office chair. The text "THANK YOU" is overlaid in large white letters at the bottom of the image.

THANK YOU