Enhancing Cross-Task Transferability with Dispersion Reduction

Yunhan Jia[1]*, Yantao Lu[1]*, Senem Velipasalar[2] , Zhenyu Zhong[1], Tao Wei[1]

[1] Baidu X-Lab , [2] Syracuse University

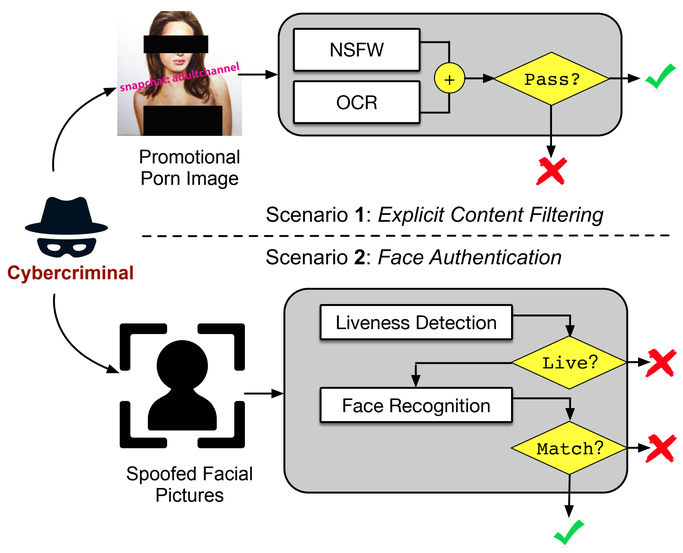
Introduction

- CV based detection systems have been extensively deployed in security-critical applications such as content censorship, authentication with face biometrics.

Real-world promotional porn image that contains explicit content for seduction, and urls/qr codes for promotional

- This incentivizes cybercriminals to generate adversarial examples that transfer across CV tasks to evade all detection (e.g., NSFW detections, and OCR) at once.

Figure1. Real-world vision based detection systems usually employ an ensemble of detection mechanisms



Attack Algorithm

- DR takes a multi-step approach that creates an adversarial example by iteratively reducing the dispersion at layer k with Adam optimizer.

- DR can use any layer of any feature extractor for generation. We compare the transferability of AEs generate on layers from shallow to deep of off-the-shelf feature extractors:
  - The result on 1000 randomly chosen samples from ImageNet shows that targeting on **middle** layer e.g., `conv3.3` of VGG-16 and `conv3.8.3` of Res-152 provides better transferability.

$$\min_x g(f(x', \theta))$$
$$s.t. \parallel x' - x \parallel_\infty \leqslant \epsilon$$

g( ) is the standard deviation of feature vectors.

Algorithm 1 Dispersion reduction attack

Input: A classifier f, original sample x, feature map at layer k; perturbation budget ϵ

Input: Attack iterations T, learning rate ℓ.

Output: An adversarial example x' with $\parallel x' - x \parallel_\infty \leqslant \epsilon$

procedure DESPERSION REDUCTION

$x'_0 \leftarrow x$

for $t = 0$ to $T - 1$ **do**

Forward x'_t and obtain feature map at layer k:

$$\mathcal{F}_k = f(x'_t)|_k \quad (3)$$

Compute standard deviation of $\mathcal{F}_k$: $g(\mathcal{F}_k)$

Compute its gradient $w.r.t$ the input: $\bigtriangledown_x g(\mathcal{F}_k)$

Update x'_t by applying Adam optimization:

$$x'_t = x'_t - Adam(\bigtriangledown_x g(\mathcal{F}_k), \ell) \quad (4)$$

Project x'_t to the vicinity of x:

$$x'_t = clip(x'_t, x - \epsilon, x + \epsilon) \quad (5)$$

return x'_t

Figure1. Dispersion reduction algorithm

Attack Effectiveness

- Evaluation on 4 Google Cloud Vision APIs for different tasks:
  - **`Labels`**, **`SafeSearch`**, **`Objects`**, **`Texts`**.
  - Datasets: 100 images from ImageNet, NSFW Data Scrapper [2], and COCO-Text[3]
  - Generation model: clean trained VGG-16 and Resnet-152, which are used by previous work MI-FGSM[3] and DIM[4].
- Adversarial examples crafted on VGG-16 model by DR brings down the performance of all GCV APIs by a large margin
- DR outperforms previous techniques in enhancing transferability when transfer across different CV tasks.

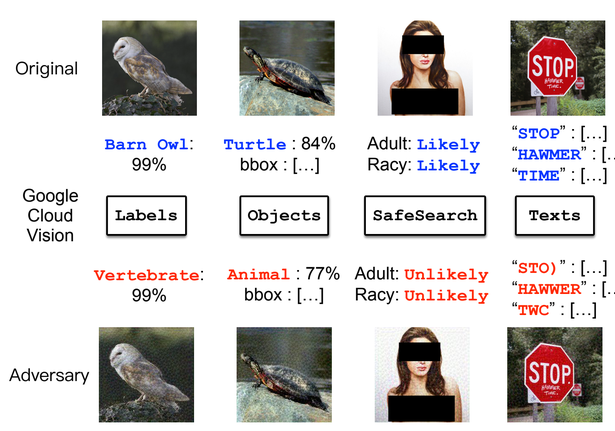


Figure3. AEs generated by DR with $l_\infty \leq 16$ fools GCV models

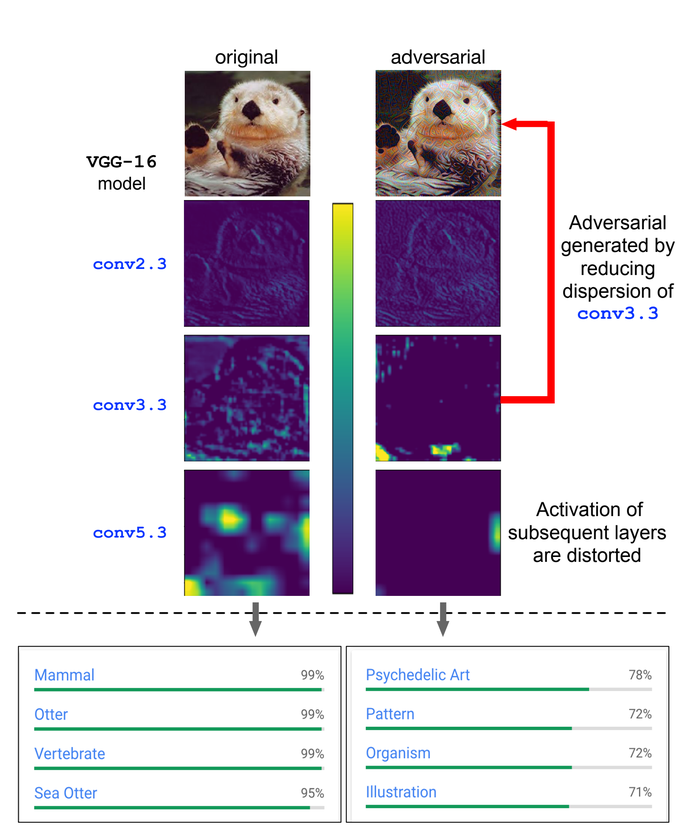
Model	Attack	`Labels` acc.	`Objects` mAP(IoU=0.5)	`SafeSearch` acc.	`Texts` AP(IoU=0.5)	`Texts` C.R.W[2]
baseline (SOTA)[1]		82.5%	73.2	100%	69.2	76.1%
VGG-16	MI-FGSM	41%	42.6	62%	38.2	15.9%
	DIM	39%	36.5	57%	29.9	16.1%
	DR (**Ours**)	**23%**	**32.9**	**35%**	**20.9**	**4.1%**
Resnet-152	MI-FGSM	37%	41.0	61%	40.4	17.4%
	DIM	49%	46.7	60%	34.2	15.1%
	DR (**Ours**)	**25%**	**33.3**	**31%**	**34.6**	**9.5%**

Figure4. The degraded performance of four GCV models where we generate AE from models in left column

Dispersion Reduction

- *Making an image "featureless"*: As lowering the *contrast* of an image would make the objects indistinguishable, we presume that reducing the "contrast" of internal feature map would also degrade the recognizability of the subjects in the image.
- Dispersion reduction (DR) attack minimizes the dispersion of feature map at internal layer
- Measure of dispersion:
  - **Standard deviation**, Gini, KL divergence, …
- We hypothesis that the low lever features extracted by early convolution layers share many similarity across CV models. Thus the distortions caused by DR in feature space, are ideally suited to fool any CV models, whether designed for classification, object detection, OCR, or the other vision tasks.

Figure2. DR attack targets on the dispersion of feature map at specific layer of feature extractors



Transfer to Google Cloud Vision **`image_label`** API

References

1. . K.Yuan et al. "Stealthy porn: understanding real-world adversarial images for illicit online promotion". S&P 18'
2. NSFW Data Scrapper on Github
3. Y. Dong et al. "Boosting adversarial attacks wih momentum". CVPR 18'
4. C.Xie et al. "Improving transferability of adversarial examples with input diversity" CVPR, 19'
5. Google Cloud Vision API

Baidu X-Lab

https://github.com/advboxes/perceptron-benchmark

Welcome to check out our project!

SYRACUSE UNIVERSITY