Efficient Image Poisoning as Defense: Disrupting Profile Matching on OSNs and Preserving Human Comprehension

Course: CS577 Data Privacy
Course Instructor: Asst. Prof. Sinem Sav

Group Number 03:
Aqsa Shabbir
Kousar Saleem
Ecem İlgün
Noor Muhammad
Mehmet Kadri Gofralılar

Date: 7th May, 2024

Facial Recognition:

- Facial recognition is extensively utilized in everyday devices.
- Expedite identification processes and enhance overall safety measures.
- Retailers utilize facial recognition technology to analyze customer demographics.
- Social media platforms integrate facial recognition algorithms for automatic tagging of friends in photos.

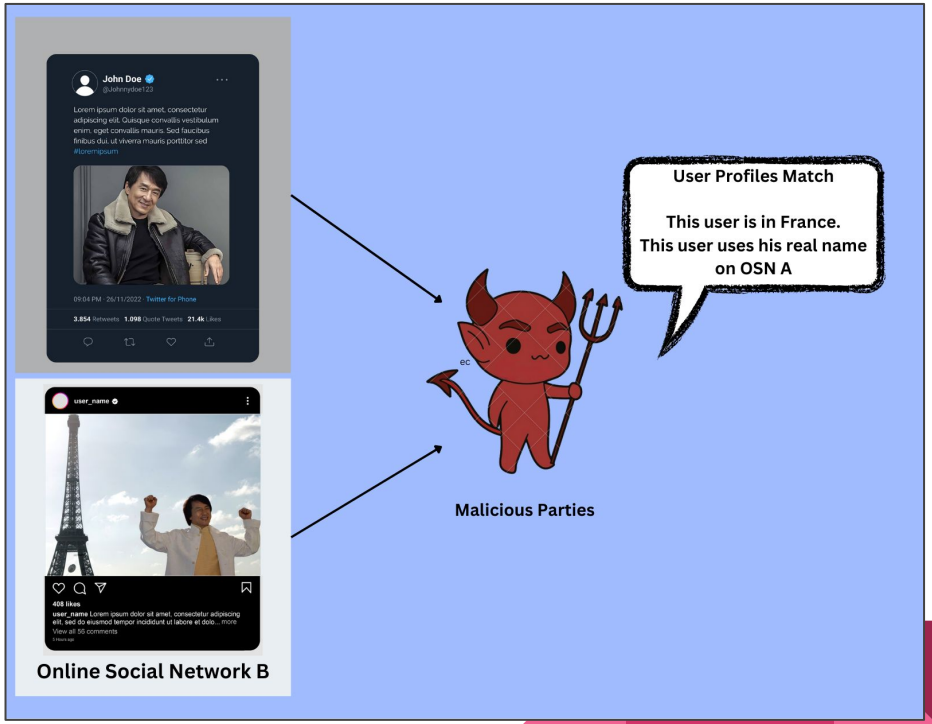
What About ***Privacy*?**

Invasive Data Collection

Surveillance and Tracking

Lack of Consent

Potential for Misuse



Objective: Obfuscate faces in images

Enhanced Privacy Protection

Prevents Biometric Data Collection

Mitigating the risks of unauthorized facial recognition

Protecting individuals' identities

Existing Approaches:

- Blurring/pixelating[1]
- Adversarial Perturbations[2]
- Physical Obfuscation[3]

Panda Perturbation Gibbon

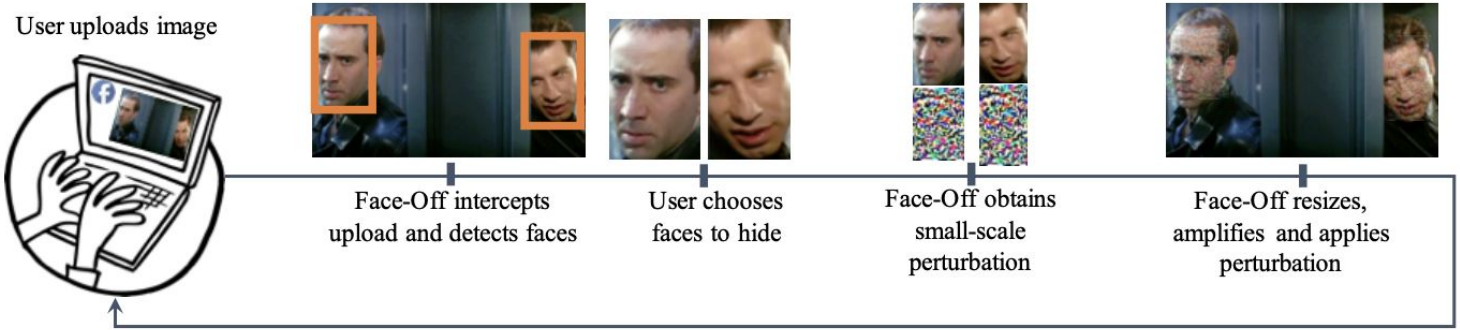
1) Richard, McPherson, Shokri Reza, and Shmatikov Vitaly. "Defeating image obfuscation with deep learning." *arXiv preprint arXiv:1609.00408* (2016)
2)Sharif, Mahmood, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." *Proceedings of the 2016 acm sigsac conference on computer and communications security*. 2016
3) Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014)

Overview Of Face-off [1]

A privacy-preserving framework that introduces strategic perturbations to the user's face to prevent it from being correctly recognized.

Challenges:

- Finding the right perturbation patterns that can successfully mislead recognition systems.
- Black box nature of facial recognition models of social media platforms.



1) Chandrasekaran, V., Gao, C., Tang, B., Fawaz, K., Jha, S., & Banerjee, S. (2020). Face-off: Adversarial face obfuscation. arXiv preprint arXiv:2003.08861.

Related Work

AdvFaces: Adversarial Face Synthesis[2]

Relevance	Introduction	Methodology	Dataset	Strengths	Weakness
Making changes to facial pictures to specifically trick facial recognition systems to make mistake.	Creating adversarial images that both have high perceptual quality and can be generated quickly.	Utilizes GANs to generate minimal perturbations in salient facial regions.	LFW	High success rates in evading detection by black-box face recognition systems.	Utility

2) Deb, D., Zhang, J., & Jain, A. K. (2020, September). Advfaces: Adversarial face synthesis. In 2020 IEEE International Joint Conference on Biometrics (IJCB) (pp. 1-10). IEEE.

Related Work

A Study of Face Obfuscation in ImageNet[3]

Relevance	Introduction	Methodology	Dataset	Strengths	Weakness
face obfuscation techniques on the accuracy and reliability.	comprehensive analysis of face obfuscation on ImageNet.	Obfuscation techniques applied to faces within ImageNet images, such as blurring, pixelation.	ImageNet [4]	Effectiveness of obfuscation techniques in preserving individual privacy.	Negatively impacting the utility of images.

3) Kaiyu Yang, Jacqueline H Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet. In International Conference on Machine Learning, pages 25313–25330. PMLR, 2022.
4) Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Conference on Computer Vision and Pattern Recognition, 2009.

Related Work

Analysis of Defeating Image Obfuscation with Deep Learning[5]

Relevance	Introduction	Methodology	Dataset	Strengths	Weakness
Maintaining privacy in the age of advanced machine learning.	Evaluating the effectiveness of privacy-preserving technologies against neural network-based attacks.	Ability of networks to recover unobfuscated content from altered images.	CIFAR-10 [6] MNIST[7] AT&T[8]	Multiple datasets ensures robustness of findings.	Resource Intensive.

5) McPherson, Richard, Reza Shokri, and Vitaly Shmatikov. "Defeating image obfuscation with deep learning." *arXiv preprint arXiv:1609.00408* (2016).
6) Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009. Available on: https://www.cs.toronto.edu/~kriz/cifar.html
7) Xiao, Han, Kashif Rasul, and Roland Vollgraf. "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms." *arXiv preprint arXiv:1708.07747* (2017)
8) http://www.uk.research.att.com/facedatabase.html

Adversarial Model

Adversary

Social Media Service Provider (semi-honest party)

Vulnerability

Individuals, places and objects can be easily tagged in the images and used for additional inference.

Threat

Perform additional inferences using the user's images beyond user's expectations.

Harm

Unwanted surveillance and/or information leak.

Face-Off: Replication Challenges

- Despite our best efforts, we were unable to replicate the results of the Face-Off paper.

https://github.com/theyusko/face-off

- After solving various issues:
  CW attack fails to assign delta with certain Margin-Target combinations

perturbation

$$\min \ ||\delta||_p$$
$$s.t. \ ||\delta||_p \leq \epsilon;$$
$$G(\mathbf{x}+\delta, t) \leq 0$$

Hinge Loss (includes margin)

Original image

target

Face-Off: Replication Challenges

Possible Reasons:

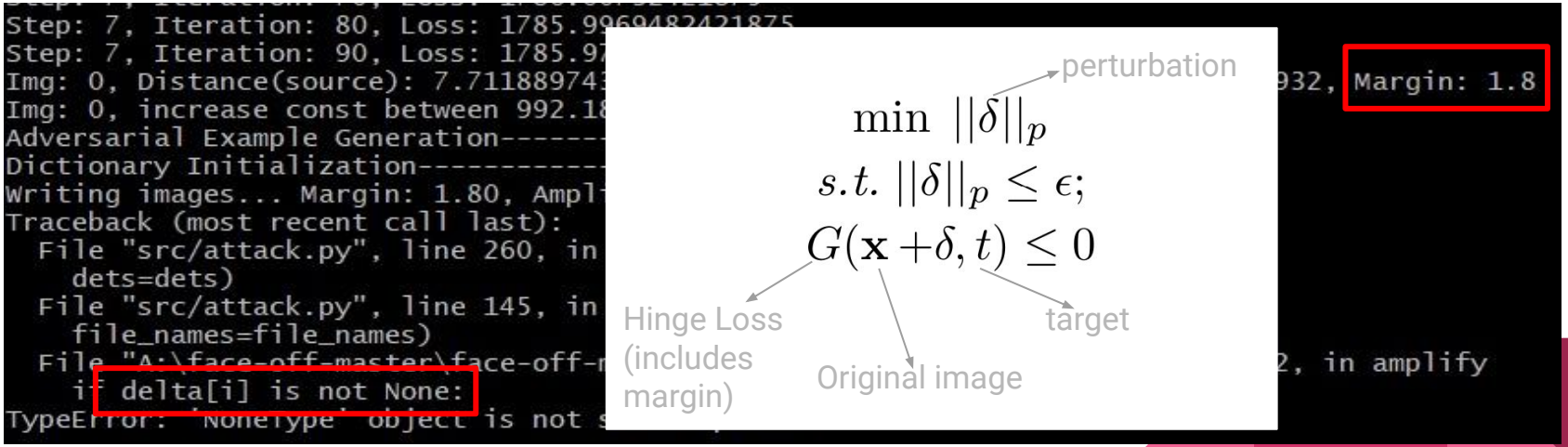
1. Authors did not publish their final code
2. Authors used libraries not supported as of 2024

```
Step: 7, Iteration: 80, Loss: 1785.9969482421875
Step: 7, Iteration: 90, Loss: 1785.9757080078125
Img: 0, Distance(source): 7.711889743804932, Distance(target): 7.711889743804932, Margin: 1.8
Img: 0, increase const between 992.18984375 and 996.094921875
Adversarial Example Generation----------------------719.340692100006
Dictionary Initialization---------------------------0.015797400003066286
Writing images... Margin: 1.80, Amplification: 1.000
Traceback (most recent call last):
  File "src/attack.py", line 260, in <module>
    dets=dets)
  File "src/attack.py", line 145, in outer_attack
    file_names=file_names)
  File "A:\face-off-master\face-off-master\src\utils\attack_utils.py", line 72, in amplify
    if delta[i] is not None:
TypeError: 'NoneType' object is not subscriptable
```

Face-Off: Replication Challenges

Possible Reasons:

1. Authors did not publish their final code
2. Authors used libraries not supported as of 2024



Novelty in our approach

- Boundary box (decision based adversarial attack)
    - Gaussian Blur

Novelty in our approach

- Boundary box (decision based adversarial attack)
    - Gaussian Blur

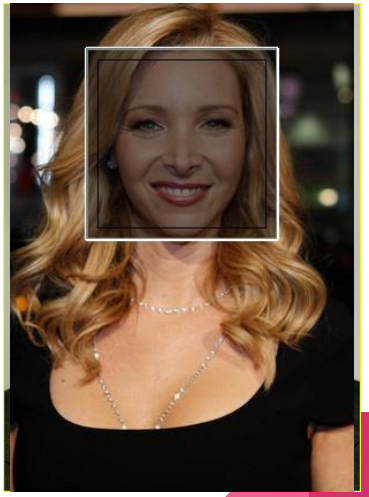
```
Total images proccessed: 608
Original Images correctly predicted: 443
Blurred Images correctly predicted: 411
Original Accuracy: 72.86184210526315
Blurred Accuracy: 67.59868421052632
```

Novelty in our approach

- Eyes Off: Make the obfuscation even more selective:
    - Only on some salient features such as eyes.

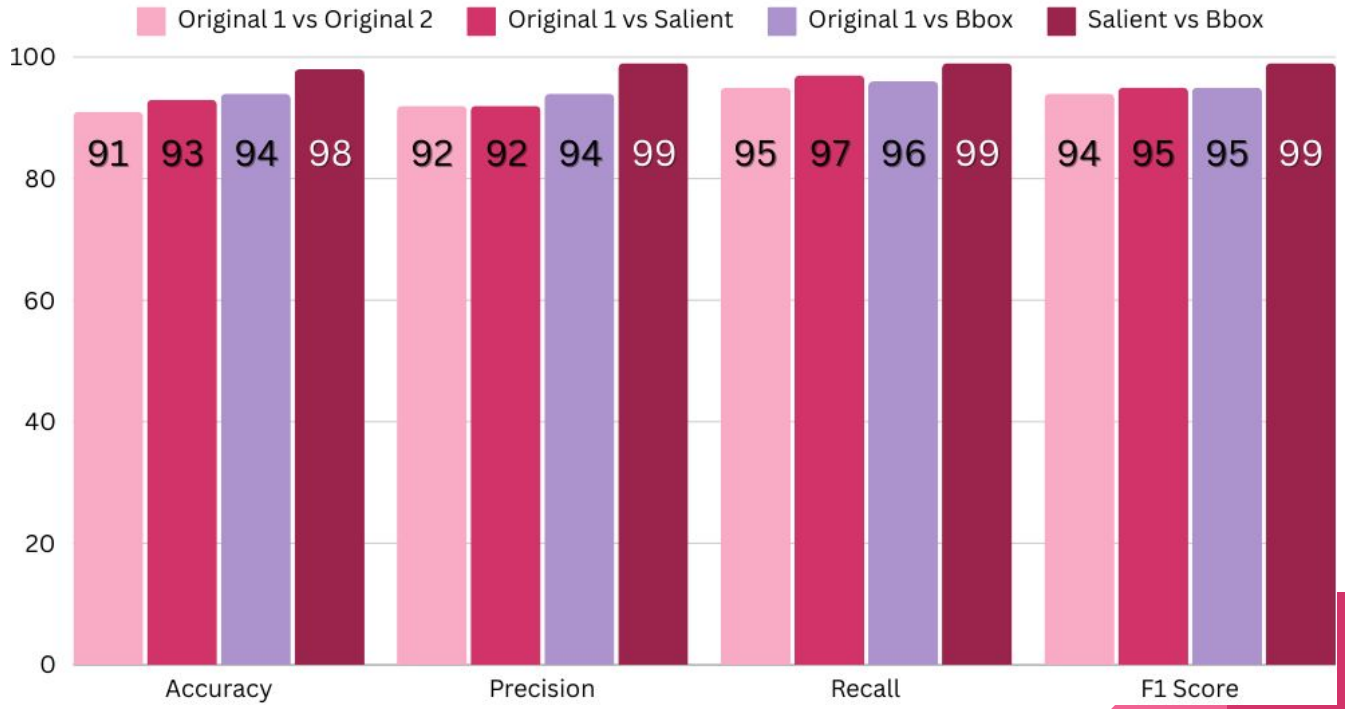
Original Image

Salient Features Obfuscated Image



Whole Face Obfuscated Image (Black Box)

Results


Original 1 vs Original 2
Original 1 vs Salient
Original 1 vs Bbox
Salient vs Bbox
100
80
60
40
20
0
91
93
94
98
92
92
94
99
95
97
96
99
94
95
95
99
Accuracy
Precision
Recall
F1 Score

Future Direction & Open Issues: Dataset Access

Dataset Name	Identities / Samples	Access
LFW(Deep funneled) [a]	5,749 / 13,233	Accessible
FaceScrub [b]	530 / 106,863	Accessible
Celeb [c]	100,000 / 10,000,000	Stopped
VGGFace2[d]	9,131 / 3,310,000	Stopped
PubFig [e]	200 / 58,797	Accessible

a) G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," 2008.
b) Garofalo, G., Rimmer, V., Preuveneers, D., & Joosen, W. (2018). Fishy faces: Crafting adversarial images to poison face authentication. In 12th USENIX Workshop on Offensive Technologies (WOOT 18).
c) Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in European conference on computer vision. Springer, 2016, pp. 87–102.
d) Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018, pp. 67–74.
e) Kumar, N., Berg, A. C., Belhumeur, P. N., & Nayar, S. K. (2009, September). Attribute and simile classifiers for face verification. In 2009 IEEE 12th international conference on computer vision (pp. 365-372). IEEE.

Future Direction & Open Issues: What will we do?

- Deciding on the dataset
- Replicating the work done in Face-Off [7]
    - Comparing it with the other papers (e.g. Poison Frogs) using the same robust metric
- Improving our new obfuscation approaches and/or an additional surrogate model
- Analyzing and documenting the results

Thank you for listening!