Protecting Multimedia Privacy from Both Humans and AI

Bo Liu
Department of Engineering
La Trobe University
Melbourne, Australia
Email:b.liu2@latrobe.edu.au

Jian Xiong
Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, China
Email: xjarrow@sjtu.edu.cn

Yiyan Wu
Communication Research Center
Ottawa, Canada
Email: Yiyan.wu@ieee.org

Ming Ding
Data61, CSIRO
Sydney, Australia
Email: ming.ding@data61.csiro.au

Cynthia M. Wu
Queens University
Kingston, Ontario, Canada.
Email: cindy.wu@sympatico.ca

***Abstract*—With the development of artificial intelligence (AI), multimedia privacy issues have become more challenging than ever. AI-assisted malicious entities can steal private information from multimedia data more easily than humans. Traditional multimedia privacy protection only considers the situation when humans are the adversaries, therefore they are ineffective against AI-assisted attackers. In this paper, we develop a new framework and new algorithms that can protect image privacy from both humans and AI. We combine the idea of adversarial image perturbation which is effective against AI and the obfuscation technique for human adversaries. Experiments show that our proposed methods work well for all types of attackers.**

***Index Terms*—multimedia, image, privacy, deep learning, face recognition**

I. Introduction

The Internet of Things (IoT) and advancements in Artificial Intelligence (AI) are significantly improving the quality of life for many people [1]. Security camera images can help search for missing people. Smart homes can turn on air conditioners or heaters while we are on our way home. AI-based face recognition systems can be used for access control. Intelligent vehicular network records can track our daily routines and where we have been. These modern conveniences depend on large amounts of IoT device data and AI derived learning models. These data and models contain a variety of personal information, such as faces, license plate numbers, routes, and behaviors. When these data are shared on social network platforms or social media, it poses severe privacy risks. The latest developments in deep learning technology have allowed malicious entities to use artificial intelligence when stealing private information from various multimedia data. Therefore, within the context of AI, the rules of multimedia privacy need to be redefined.

As a result, we need to consider two types of attacks: AI and human. Traditional obfuscation techniques such as blurring and pixelation work well when the adversary is human, however, they are not effective against AI-assisted attackers. There has been recent research on combating special AI-based attackers who are mostly using deep learning technologies. The fundamental idea is to generate a small but intentional worst-case disturbance to an original image, which misleads deep neural networks (DNN) without causing a significant difference perceptible to human eyes. The perturbed image is called an "adversarial example" [2] [3] and the speically generated noise is named adversarial image perturbations (AIP). A few papers have discussed the potential of AIP in privacy protection. For example, Liu et al. [3] investigated the AIP-based privacy protection scheme for the image classification problem. However, no previous work has investigated how to combine both these types of privacy protection into the same framework.

Due to the current situation, there is a need to develop a new framework and new algorithms that will protect multimedia privacy from both humans and AI. Image and video are the most common multimedia types. Video can be interpreted as consecutive images. This article will focus on image privacy. It follows that the results can provide insight to video privacy as well. Since the face is one of the most important human identifiers, an existing facial recognition application was used. DNN-based AI are used in this paper.

In summary, the contributions of this paper are as follows:

- Developing an image privacy protection framework against both human and AI.
- Proposing a probability-based image privacy metric with the context of human and AI as adversaries.
- Proposing a dual target image privacy protection scheme, which performs well against both human and DNN-based adversaries.

The rest of the paper is organized as follows. Section II discusses the system model and formulates the research problem. In Section III, the dual target image privacy protection scheme is presented. Section IV shows the experimental results. Finally, the results are concluded in Section V.

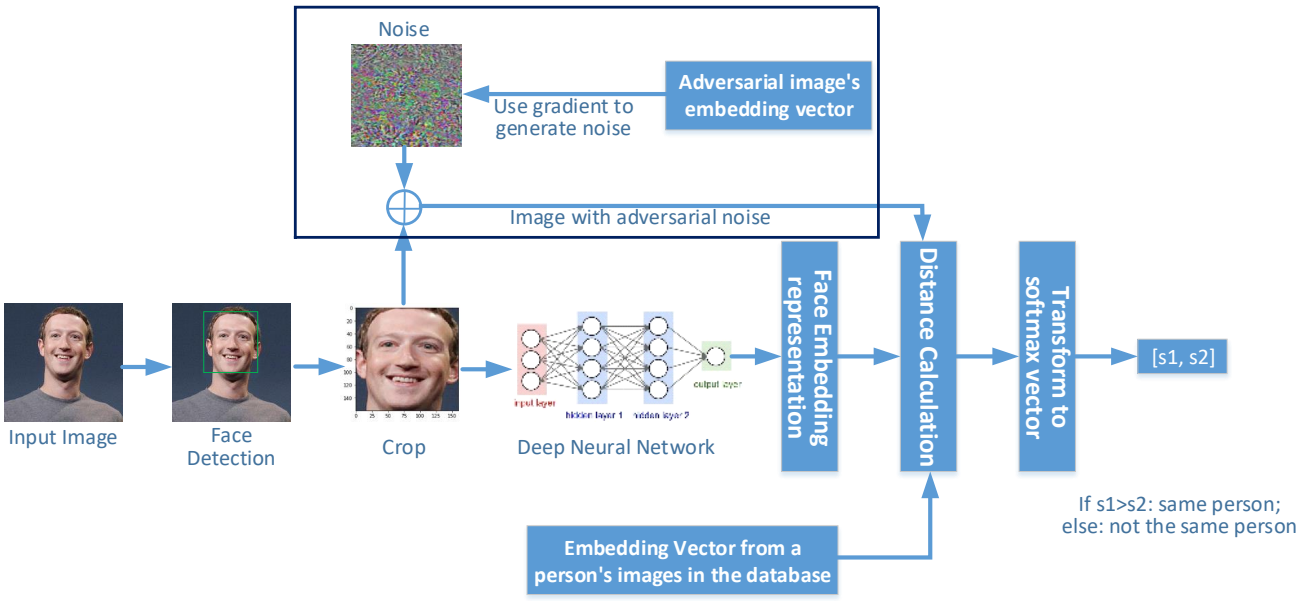


Fig. 1: Illustration of a typical face recognition system and the process of generating adversarial image perturbation.

II. System Model and Problem Formulation

A. AI based Face Recognition Systems

A face recognition system is a technology that is capable of recognizing or authenticating a person from an image or a video frame. With recent advanced deep learning neural networks, the accuracy of artificial intelligence-based face recognition systems have begun to surpass human accuracy in some benchmark tests. As a result, they are beginning to see a wider range of uses in many applications, such as access control and security monitoring.

Fig. 1 depicts a typical face recognition system. When an input image is received, it first detects the position of the face and crops the face to the size that is aligned with the system settings. The DNN is used to calculate a face embedding (a numerical vector representing the facial features) from the face image. Then the system can calculate the distance between the embedding of the input face and any given embedding from the system database. The distance is converted to a vector containing two soft values that indicate the face recognition result: if the first value is greater than the second one, then the two embeddings are from images of the same person. Otherwise, they are the images of two different persons.

In a sense, the face recognition system is similar to the person in performing the task of recognizing another person: the person compares the new image with their memory. If the image looks close to someone in their memory, they reckon it as the same person. The only difference is how DNN and humans measure the "distance" between images.

B. System Model

Based on the above principle of face recognition, the basic idea of the proposed image privacy protection is to interfere with the measurement of image distances by introducing noise. Fig. 2 shows the privacy protection framework. The original data is a person's photo. The face is considered to be private information. The original photo can be correctly recognized both by humans and pre-trained DNNs with high confidence. Now there are three different privacy protection scenarios. In Scenario 1 (machine-recognizable picture sharing), the person is happy to share their photo with an AI (learning algorithm) but does not want to be recognized by other people. The photo is processed by adding some human-sensitive noise (blur), so that it can barely be recognized by humans but can be identified by a learning algorithm. In Scenario 2 (human-recognizable picture sharing), the person would like to share their photo with friends but is not inclined to share their photo with a training centre. Crafted noise can be added to the photo to destroy the AI's learning features. Even if the AI could recognize some features, it could only do so with poor accuracy. In Scenario 3 (non-recognizable picture sharing), the person does not want to be recognized by either humans or by machines (e.g., in Google street view or video news), so the noises are combined.

C. Privacy Metric

The three scenarios in Fig. 2 demonstrate very typical requirements of image privacy. The challenge is to quantify these levels of recognition. The proposed solution is to define privacy metrics associated with the nature of human and learning recognition. As the output of DNN is soft values that

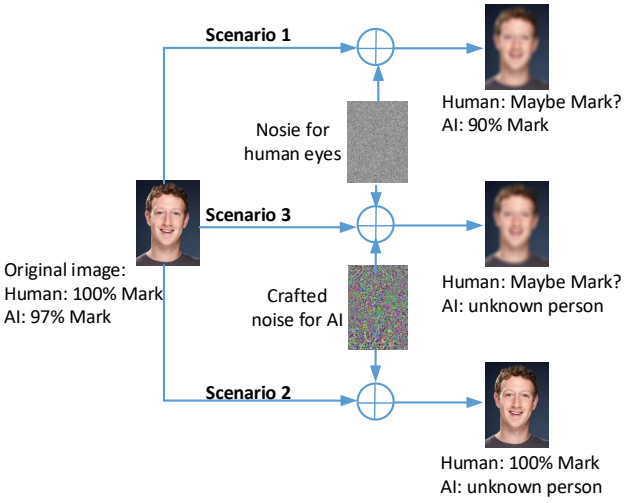


Fig. 2: Privacy protection framework that deals with three different scenarios.

represent probabilities, the privacy metrics can also be defined by probabilities.

Probability-based Metric: The probability-based metric is defined as: "Privacy protection success rate". It is calculated as the probability that the face recognition system cannot identify the person correctly, i.e.,

$$\Pr(ID_{\boldsymbol{x}'} \neq ID_{\boldsymbol{x}}), \tag{1}$$

where $ID_{\boldsymbol{x}}$ is the identity of the original image, and $ID_{\boldsymbol{x}'}$ is the identity of the image with perturbation.

D. Problem Formulation

Based on the privacy protection successful rate metric, the proposed image privacy protection problem can be formulated as:

$$\text{P: } \max \Pr(ID_{\boldsymbol{x}'} \neq ID_{\boldsymbol{x}}). \tag{2}$$

III. Dual Target Image Privacy Protection Scheme

As stated in Section II, the aim is to mislead both humans and the DNNs so that privacy in images can be preserved.

For the noise to defend from human adversaries, the classic Gaussian blurring scheme is used.

For the noise to defend from AI, the idea of AIP was used [2] [4] [5], as the DNN can be fooled by adding a small amount of "well designed" AIP to the original image. Fast gradient sign method (FGSM) [2] is the most classic and straightforward algorithm to generate AIP. A brief introduction of FGSM and its iterative version "projected gradient descent (PGD)" is as follows.

1) Fast Gradient Sign Method (FGSM): Let $\boldsymbol{x}$ be the original input image, $\boldsymbol{\theta}$ the model parameters, y the adversarial target (the goal is to mislead the DNN to think that $\boldsymbol{x}$ belongs to y), and $J(\boldsymbol{\theta}; \boldsymbol{x}; y)$ be the cost function that is used to train the neural network. Then an optimal max-norm constrained AIP can be obtained by calculating

$$\boldsymbol{\eta} = \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}; \boldsymbol{x}; y)), \tag{3}$$

where ϵ is a small scalar that adjusts the amount of noise, and $\nabla_{\boldsymbol{x}}$ is the gradient with regard to the input image $\boldsymbol{x}$, i.e,

$$\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}; \boldsymbol{x}; y) = \frac{\partial J}{\partial \boldsymbol{x}}. \tag{4}$$

The required gradient can be computed efficiently using the backpropagation scheme.

The release image with AIP is generated by

$$\boldsymbol{x'} = \boldsymbol{\eta} + \boldsymbol{x}. \tag{5}$$

2) Projected Gradient Descent (PGD): A more powerful adversary is the multi-step variant $FGSM^N$, which is essentially projected gradient descent (PGD) on the negative loss function [6] [7].

In PGD, the FGSM will be repeated for N times or until the absolute value of noise reaches a pre-defined upper bound, i.e.,

$$\begin{aligned} \boldsymbol{x}'_0 &= \boldsymbol{x} \\ \boldsymbol{x}'_n &= \boldsymbol{x}'_{n-1} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}; \boldsymbol{x}'_{n-1}; y)) \\ &= \boldsymbol{x}'_{n-1} + \boldsymbol{\eta}_{n-1}, 1 \leq n \leq N. \end{aligned} \tag{6}$$

An image privacy preservation scheme against AI can be based on the above introduced FGSM and PGD methods. The process of the scheme is shown in Alg. 1. First, a different person was specifically or randomly selected. Then the embedding vector of this adversarial person will be calculated and used as the value of y in Equ. (6). The image with adversarial perturbation is generated by the PGD algorithm and finally tested using the face recognition system.

Algorithm 1: AIP-based image privacy preservation scheme.

Parameters: Noise scalar ϵ.
Noise limit η_{max}.
Iteration number N.
Input: The original image $\boldsymbol{x}$.
Output: The released privacy preserving image $\boldsymbol{x}'_n$.
Initialization: Overall noise $\boldsymbol{\eta} = 0$, $\boldsymbol{x}'_0 = \boldsymbol{x}$.
Randomly select an adversarial person in the database.
y is the embedding generated from this person's images.
for $1 \leq n \leq N$ **do**
    $\boldsymbol{\eta}_{n-1} = \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}; \boldsymbol{x}'_{n-1}; y))$;
    $\boldsymbol{\eta} = \boldsymbol{\eta}_{n-1} + \boldsymbol{\eta}$;
    Clip the element in $\boldsymbol{\eta}_{n-1}$ if $\boldsymbol{\eta}$ exceeds η_{max};
    $\boldsymbol{x}'_n = \boldsymbol{\eta}_{n-1} + \boldsymbol{x}'_{n-1}$;
end

Based on the above algorithms, the dual target image privacy protection scheme can be implemented by selecting

different combinations according to the scenario. In scenario 1, Gaussian noise based blurring can be used. In scenario 2, the proposed AIP-based image privacy preservation scheme can be adopted. Finally in scenario 3, the two schemes can be combined to achieve the dual target protection goal.

IV. Experiments and Discussions

A. Experiment Setup

In the experiments, the open source face recognition system FaceNet [8] was used. It has the system structure shown in Fig. 1 and uses Inception ResNet v1 architecture [9] for the DNN part. The models of FaceNet are trained on several large-scale datasets, including MS-Celeb-1M [10] (faces from 100K celebrities), VGGFace2 [11] (3.3M faces from 9000 persons), and CASIA-WebFace (453,453 images over 10,575 identities) datasets. The newest model achieves 0.9965 accuracy when validating on the Labeled Faces in the Wild Home (LFW) dataset [12].

B. Experiment Results

First, the method was tested using an image of Mark Zuckerburg: a color image with 3 channels (RGB) where the maximum value for each pixel was 255. The adversarial noise was generated by PGD using Bill Gates as the adversarial target. The parameters were set as $\epsilon = 0.3$ with iteration number 100. Then the perturbed image was inputted into the FaceNet system. The output vector was $[s1, s2] = [0.39029777, 0.6097022]$. Since $s1 < s2$, it means that the system thinks it is not Mark Zuckerburg. Fig. 3 shows the original image, the perturbed image, and the adversarial noise. As the noise value was small (the mean square root of the adversarial noise power was 5.1361), the noise was amplified by normalization. It can be seen that it is still quite easy for a human to recognize the perturbed image as Mark Zuckerburg, while the FaceNet system has lost the accuracy.

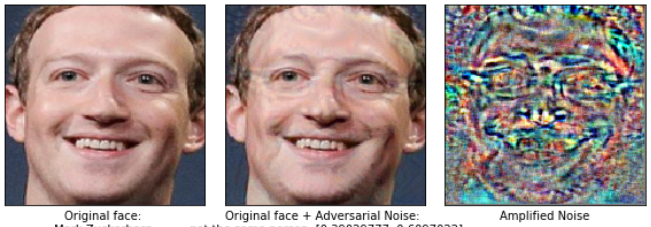


Fig. 3: Performance of AIP-based image privacy preservation scheme when $\epsilon = 0.3$ and iteration number = 100.

Fig. 4 shows the effect of the traditional image privacy preservation method of simple blurring. Gaussian noise with similar level of power ($\delta = 5$) was added, and then the blurred image was inputted into FaceNet. The system outputted the vector $[s1, s2] = [0.75867796, 0.24132206]$. This meant that it still had high confidence in identifying this blurred image as Mark Zuckerburg, but it was difficult for a human to recognize the identity.

Fig. 4: Performance of traditional blurring method with Gaussian noise ($\delta = 5$).

Then the AIP-based scheme and blurring methods were combined together. As shown in Fig. 5, the FaceNet system could not identify the person correctly, and it was also challenging for humans.

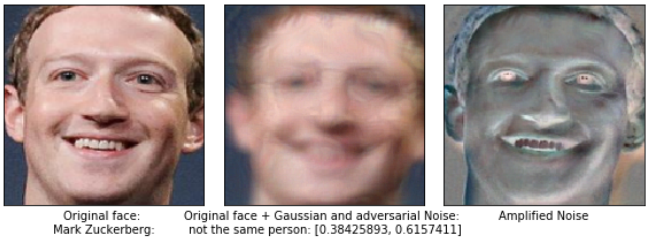


Fig. 5: Performance of dual target image privacy protection scheme.

We now further investigate the amount of noise needed to protect privacy from AI. As shown in Fig. 6, we can fail the FaceNet system with a small amount of noise ($\epsilon = 0.1$). The "false" confidence of the system increases with the increase of the noise (from $\epsilon = 0.1$ to $\epsilon = 0.3$), but then becomes stable later (from $\epsilon = 0.3$ to $\epsilon = 1$). This tells us that we do not need much AIP to fail the current face recognition systems.

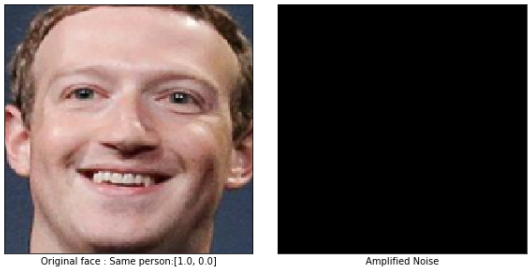
Finally, large-scale experiments were run on the LFW dataset. 2000 images of different persons were used as the input image and the adversarial target was randomly selected. It can be seen in Fig. 7 that only an imperceptible noise ($\epsilon = 0.01$) can mislead the FaceNet system with over 60% probability. The probability increases to over 90% when $\epsilon = 0.3$.

C. Discussions

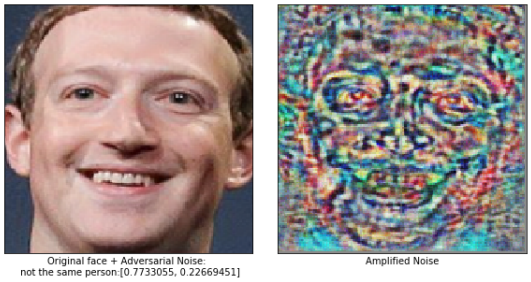
From the above results, it can be seen that privacy protection from AI and humans are completely different tasks, at least at the current stage when DNN is used as the mainstream technique for AI. The DNN-based face recognition system can be misled by a small amount of elaborately crafted adversarial noise, while at the same time can maintain resilience against random noises.

The transitional image privacy protection schemes such as blurring are not effective in the context of AI. This is not only because of the power of the DNN, but also due to the ability to reverse the blurring process easily [13].

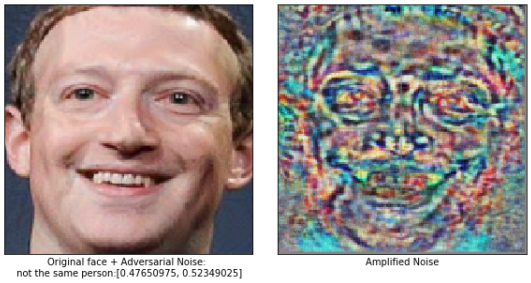
Furthermore, the transferability of adversarial perturbations means adversarial noises crafted for one DNN model, can



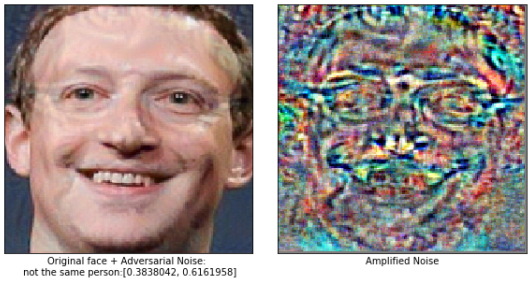
(a) No adversarial noise



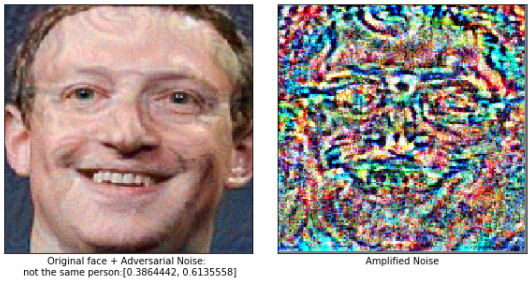
(b) $\epsilon = 0.05$



(c) $\epsilon = 0.1$



(d) $\epsilon = 0.3$



(e) $\epsilon = 1$

Fig. 6: Performance of AIP-based image privacy preservation scheme with different values of ϵ

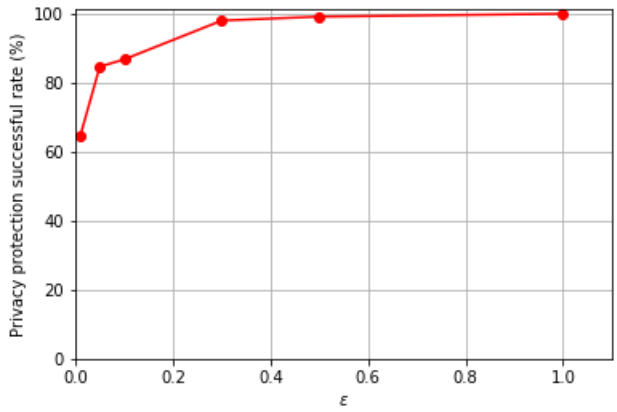


Fig. 7: Performance of PGD on large-scale dateset.

work on other models as well. To test the transferability of the proposed privacy protection method, the online face recognition API: Clarifai Celebrity Recognition [14] can be used. As we had no access to the model of the Clarifai Celebrity Recognition system, the FaceNet model was used to generate the perturbed image. As shown in Fig. 8(a), the Clarifai Celebrity Recognition identified the original image as Mark Zuckerburg with high confidence (0.97 out of 1). However, it failed to recognize the image with adversarial perturbation, instead giving the information "No celebrity detected", as shown in Fig. 8(b).

V. Conclusions

The introduction and development of artificial intelligence and deep learning has greatly reformed the context of privacy protection. It has become more challenging than ever before to protect sensitive multimedia information shared on social network platforms. In order to solve this problem, we proposed a scheme to protect image privacy, especially the facial identity, from both humans and AI. The contributions of this paper are two-fold. First, we introduced the image privacy preservation scheme against AI, based on the adversarial image perturbation theory. Second, we designed the dual target image privacy protection scheme by combining our proposed scheme with traditional obfuscation method. The results show that good face identity protection can be achieved against state-of-art face recognition systems based on deep neural networks by adding a small amount of noise. And by adding some Gaussian noise, we can protect the privacy from humans as well.

References

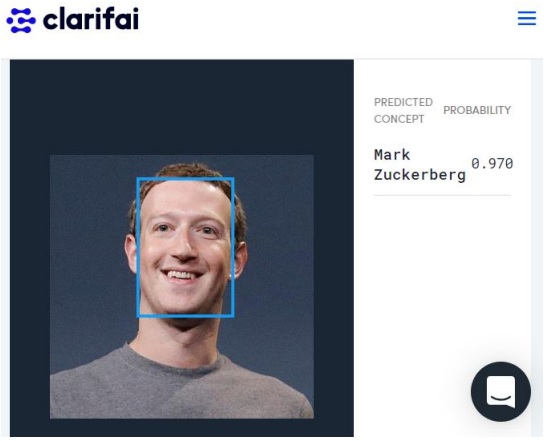
[1] R. Crist, "How ai is spreading everywhere with the rise of smart machines," https://www.techrepublic.com/, Nov. 2018.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

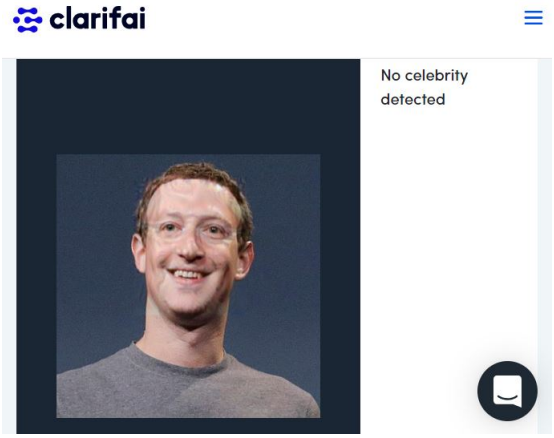
[3] B. Liu, M. Ding, T. Zhu, Y. Xiang, and W. Zhou, "Adversaries or allies? privacy and deep learning in big data era," *Concurrency and Computation: Practice and Experience*, p. e5102.

[4] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," *arXiv preprint*, 2017.

[5] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, "Generative adversarial perturbations," 2017.



(a) No adversarial noise



(b) $\epsilon = 0.3$

Fig. 8: Test results using Clarifai Celebrity Recognition API.

[6] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[8] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[9] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[10] 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.

[11] 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.

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[13] R. McPherson, R. Shokri, and V. Shmatikov, "Defeating image obfuscation with deep learning," *arXiv preprint arXiv:1609.00408*, 2016.

[14] "Clarifai celebrity recognition [online]," https://clarifai.com/models/celebrity-image-recognition-model-e466caa0619f444ab97497640cefc4dc, accessed: 2018-04-13.