ADVERSARIAL FACE DE-IDENTIFICATION

Efstathios Chatzikyriakidis Christos Papaioannidis Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

contact@efxa.org, {chpapaioann, pitas}@aiia.csd.auth.gr

ABSTRACT

Recently, much research has been done on how to secure personal data, notably facial images. Face de-identification is one example of privacy protection that protects person identity by fooling intelligent face recognition systems, while typically allowing face recognition by human observers. While many face de-identification methods exist, the generated de-identified facial images do not resemble the original ones. This paper proposes the usage of adversarial examples for face de-identification that introduces minimal facial image distortion, while fooling automatic face recognition systems. Specifically, it introduces P-FGVM, a novel adversarial attack method, which operates on the image spatial domain and generates adversarial de-identified facial images that resemble the original ones. A comparison between P-FGVM and other adversarial attack methods shows that P-FGVM both protects privacy and preserves visual facial image quality more efficiently.

Index Terms— Privacy Protection, Face De-identification, Adversarial Examples, Deep Learning, Computer Vision

1. INTRODUCTION

In recent years, state-of-the-art deep learning and deep neural network methods have been applied for face recognition. At the same time, several efforts have been made for face de-identification, for person identity protection. In the past, several ad-hoc methods (e.g., masking, pixelization and blurring) [1-3] were used for face de-identification that are capable to fool various face classifiers by strongly altering the input facial image. However, many current state-of-the-art deep neural face recognizers are robust to such ad-hoc attacks. Furthermore, the aforementioned methods strongly alter face appearance in the de-identified image, thus making them useless in several applications (e.g., social networks).

Subsequently, various face de-identification techniques began to appear, which are based on the k-anonymity framework (e.g., k-Same [4] family of methods). They exploit statistical information from a set of facial images to produce more realistic de-identified facial images. Nevertheless, the result is often unsatisfactory, as the de-identified facial images eventually deviate significantly from the original input images. Furthermore, the depicted faces tend to resemble each other and thus, lose their unique characteristics related, e.g., to race, gender, age, expression or pose. Other interesting face de-identification techniques use a batch of facial images selected from a database, based on extracted well-defined facial features in order to de-identify an input facial image. The batch facial images are used as donors of facial characteristics, in order to alter the input facial image [5]. Alternatively, preexisting k-anonymity methods [6] are used to alter the input facial image, using the batch facial images. However, both techniques do not preserve the unique characteristics of the input facial images.

Recently, sophisticated techniques have been developed with the sole purpose of producing realistic de-identified facial images. Specifically, with the rise of Generative Adversarial Networks (GANs) [7] and, more generally, of generative models [8] various methods have been proposed [9-14] that replace the input face with a new, realistic and synthetic facial image. However, as we want both face de-identification and retaining the original face appearance, such methods fail to live to our expectations.

In this work, we propose the usage of adversarial examples [15, 16] to achieve face de-identification, so that the generated de-identified facial images are as realistic as possible and visually very similar to the original ones. Furthermore, we introduce the Penalized Fast Gradient Value Method (P-FGVM), a novel adversarial attack, which operates on the image spatial domain and generates adversarial examples for face de-identification that resemble the original facial images.

2. ADVERSARIAL EXAMPLES

Adversarial examples are inputs to machine learning classification models, which are carefully constructed and usually imperceptibly different from pre-existing original images that result to incorrect image classification. Specifically, let $(\boldsymbol{x}_i, y_i)$ be a dataset facial image entry which comprises of a feature vector $\boldsymbol{x}_i \in X \subseteq R^n$ and the corresponding ground truth label $y_i \in Y$. Suppose a deep neural network classifier has learned the mapping $f: X \to Y$, using a training dataset. Given an instance $\boldsymbol{x}$ with ground truth label y, such that $f(\boldsymbol{x}) = y$, it is possible to generate two types of adversarial examples—targeted and non-targeted ones. In both cases, the adversarial example $\widehat{\boldsymbol{x}}$ is crafted by adding a small adversarial perturbation to $\boldsymbol{x}$, so that $\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|_p \leq \varepsilon$ where ε is a small value to control the magnitude of the adversarial perturbation. For the non-targeted adversarial example we aim at $f(\widehat{\boldsymbol{x}}) \neq y$. For the targeted adversarial example, we aim at $f(\widehat{\boldsymbol{x}}) = \hat{y}$, where $\hat{y}$ is a specified target label, different than y.

The fast gradient-based adversarial example generation methods [17] use the gradient $\nabla_x \ell_f$ of the loss function ℓ_f (e.g., cross-entropy error) of the classifier f w.r.t. to an input $\boldsymbol{x}$, in order to transform $\boldsymbol{x}$ to an adversarial example $\widehat{\boldsymbol{x}}$. Iterative Fast Gradient Sign Method (I-FGSM) [16, 18] and Iterative Fast Gradient Value Method (I-FGVM) [18, 19] follow this methodology. They differ in the way they use the $\nabla_x \ell_f$ gradient. Specifically, the I-FGVM method changes the input $\boldsymbol{x}$ in the direction of the gradient, while the I-FGSM method uses only the sign gradient. The gradient descent update equations for the methods are the following ones:

I-FGVM

$$\widehat{\boldsymbol{x}}_0 = \boldsymbol{x},$$
$$\widehat{\boldsymbol{x}}_{i+1} = clip_{[0,1]}(clip_{[\boldsymbol{x}-\varepsilon, \boldsymbol{x}+\varepsilon]}(\widehat{\boldsymbol{x}}_i - \alpha \cdot \nabla_x \ell_f(\widehat{\boldsymbol{x}}_i, \hat{y})))$$

I-FGSM

$$\widehat{\boldsymbol{x}}_0 = \boldsymbol{x},$$
$$\widehat{\boldsymbol{x}}_{i+1} = clip_{[0,1]}(clip_{[\boldsymbol{x}-\varepsilon, \boldsymbol{x}+\varepsilon]}(\widehat{\boldsymbol{x}}_i - \alpha \cdot sign(\nabla_x \ell_f(\widehat{\boldsymbol{x}}_i, \hat{y})))$$

where α is the step size, $\boldsymbol{x}$ is the original image, $\nabla_x \ell_f(\widehat{\boldsymbol{x}}_i, \hat{y})$ is the first-order gradient term of the adversarial loss and $clip_{[a,b]}$ is a value constraint so that pixel values cannot go beyond the $[a, b]$ range. Thus, $clip_{[0,1]}$ constrains the pixel values to ensure data validity and $clip_{[\boldsymbol{x}-\varepsilon, \boldsymbol{x}+\varepsilon]}$ enforces the L^∞ norm of the adversarial perturbation to be within the limits defined by ε.

3. PENALIZED FAST GRADIENT VALUE METHOD

The proposed novel adversarial attack method Penalized Fast Gradient Value Method (P-FGVM) is inspired by the baseline adversarial attack method I-FGVM. P-FGVM combines an adversarial loss and a 'realism' loss term. It is capable of generating a targeted adversarial example $\widehat{\boldsymbol{x}}$ by using the following gradient descent update equations:

$$\widehat{\boldsymbol{x}}_0 = \boldsymbol{x},$$
$$\widehat{\boldsymbol{x}}_{i+1} = clip_{[0,1]}(\widehat{\boldsymbol{x}}_i - \alpha \cdot (\nabla_x \ell_f(\widehat{\boldsymbol{x}}_i, \hat{y}) + \lambda \cdot (\widehat{\boldsymbol{x}}_i - \boldsymbol{x})))$$

where α is the step size, $\boldsymbol{x}$ is the original image, λ is a weight coefficient, $clip_{[0,1]}$ constrains the pixel values to ensure data validity, $\nabla_x \ell_f(\widehat{\boldsymbol{x}}_i, \hat{y})$ is the first-order gradient term of the adversarial loss and $\widehat{\boldsymbol{x}}_i - \boldsymbol{x}$ is the 'realism' loss term.

4. EXPERIMENTAL RESULTS

We performed an experimental evaluation of the proposed P-FGVM method and compared it with the baseline I-FGVM and I-FGSM methods for face de-identification. We used as target models two deep convolutional neural networks, shown in Table 1. Both target models were trained (see Table 4 for training information and performance accuracies on the original data) on NVIDIA GeForce GTX 1080 GPU for face recognition with a subset of the CelebA dataset [20]. The model A has a simple architecture and the model B was fine-tuned with transfer learning based on the pre-trained state-of-the-art VGG-Face CNN descriptor [21], using the VGG-16 architecture [22]. Our CelebA subset contains 900 random, aligned, cropped and colored 178x218 pixel facial images, corresponding to 30 persons with 30 facial images each in order to have balanced labels.

First, we applied the P-FGVM method aiming to generate realistic de-identified facial images (as targeted adversarial examples) with high misclassification rate and having as input either Gaussian random noise or existing input facial images. Next, we applied the baseline I-FGVM and I-FGSM methods with the same objective, having as input only existing input facial images and requiring the L^∞ norm of the adversarial perturbation to be within the limits defined by ε. In all experiments we calculated the MSSIM similarity index between the de-identified and original facial images as well as the L^2 norm $\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|_2$ of the adversarial perturbation as the metrics for measuring the visual quality of the results.

The parameter values used in our experiments are shown in Table 2. The L^2 norm, the MSSIM similarity index, the misclassification rate as well as the percentage improvement in these metrics by the proposed P-FGVM method comparatively to the competing methods, are shown in Table 3. It is clearly seen that the proposed method produces de-identified images that are much closer to the original ones, while having better misclassification error than the competing methods. Examples of de-identified facial images are shown in Figure 1. Furthermore, the evolution of an example de-identified facial image, having as input Gaussian random noise is shown in Figure 2.

Table 1: The architecture of the target CNN models.

Model A
Conv(32, Kernel(5, 5), Padding(Same), L2Regularizer(0.001)) BatchNormalization+Relu MaxPooling(PoolSize(2, 2), Strides(2, 2))

Conv(64, Kernel(5, 5), Padding(Same), L2Regularizer(0.001)) BatchNormalization+Relu MaxPooling(PoolSize(2, 2), Strides(2, 2)) FC(512, L2Regularizer(0.001)) BatchNormalization+Relu Dropout(0.9) FC(30)+Softmax
Model B
VGG-Face CNN descriptor (VGG-16) FC(256, L2Regularizer(0.001)) BatchNormalization+Relu FC(30)+Softmax

Table 2: Parameter values (α: step size, N: iterations, ε: clipping threshold, λ: weight coefficient of 'realism' loss term) of the adversarial attack methods P-FGVM, I-FGVM and I-FGSM.

	Model A				**Model B**			
Method	**α**	**N**	**ε**	**λ**	**α**	**N**	**ε**	**λ**
P-FGVM	1.0	50	n/a	0.22	0.55	58	n/a	0.28
I-FGVM	1.0	50	0.022	n/a	0.1	40	0.022	n/a
I-FGSM	ε÷N	20	0.026	n/a	ε÷N	20	0.026	n/a

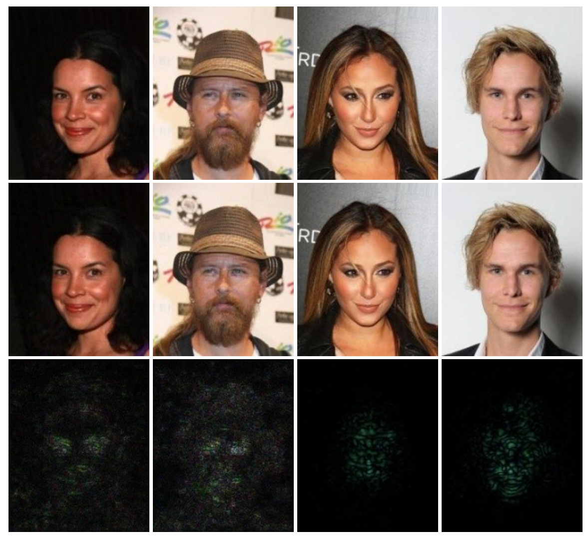
Table 3: The experimental results and the percentage improvement in metrics from the comparison between the proposed P-FGVM method and the baseline I-FGVM, I-FGSM methods. L2: Average L^2 norm of adversarial perturbation between the original and the de-identified images. SI: Average MSSIM similarity index between the original and the de-identified images. MR: Misclassification rate of the de-identified images.

Model A			**Model B**		
L2	**SI**	**MR**	**L2**	**SI**	**MR**
Experimental Results					
P-FGVM					
3.38	0.986	99.6%	2.11	0.995	96.0%
I-FGVM					
5.31	0.963	99.4%	2.67	0.993	93.2%
I-FGSM					
5.68	0.962	98.9%	5.74	0.968	94.4%
Percentage Improvement					
I-FGVM					
36.3%	2.3%	0.2%	20.9%	0.2%	3.0%
I-FGSM					
40.4%	2.4%	0.7%	63.2%	2.7%	1.7%

Table 4: The training information of the target CNN models.

	Model A	**Model B**
Dataset	CelebA	CelebA
Subset Classes	30	30
Subset Images	900	900
Image Size	178 x 218	178 x 218
Training Set Size	70%	70%
Testing Set Size	15%	15%
Validation Set Size	15%	15%
Normalization	MinMax	MinMax
Learning Rate	0.0001	0.0001
Optimization	Backprop+Adam	Backprop+Adam
Loss Function	Cross Entropy	Cross Entropy
Batch Size	16	16
Training Epochs	147	144
Validation Accuracy	80%	97.8%
Testing Accuracy	80.7%	95.4%

Model A **Model B**



(a)

(b)

(c)

Figure 1: Examples of de-identified facial images (two for each target model) generated by the adversarial attack method P-FGVM: a) clean input facial images, b) de-identified facial images, c) adversarial perturbation absolute value amplified by 10x.

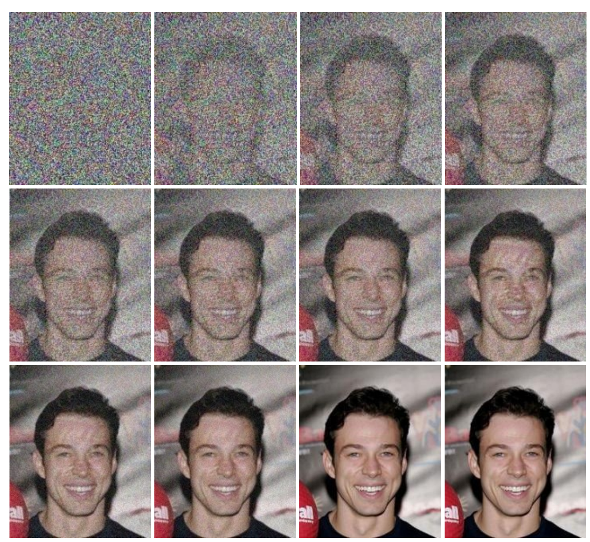


Figure 2: Evolution of an example de-identified facial image generated by the adversarial attack method P-FGVM using as input Gaussian random noise.

5. CONCLUSION

The existing adversarial face de-identification methods fail to preserve the face appearance of the original image. Therefore, we proposed the novel P-FGVM adversarial attack method for generating realistic de-identified facial images (as targeted adversarial examples) with high misclassification rate. By evaluating the proposed P-FGVM and baseline I-FGVM, I-FGSM methods on various deep convolutional neural network face classifiers trained on a subset of the CelebA dataset, we show that the P-FGVM method both protects privacy and preserves visual facial image quality more efficiently than its competitors.

6. ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE). This publication reflects only the authors' views. The European Commission is not responsible for any use that may be made of the information it contains.

7. REFERENCES

[1] Crowley, J. L., J. Coutaz, and F. Berard. "Things that see: Machine perception for human computer interaction." Communications of the ACM 43, no. 3 (2000): 54-64.

[2] Neustaedter, Carman Gerard, and Saul Greenberg. "Balancing privacy and awareness in home media spaces." University of Calgary, Department of Computer Science, 2003.

[3] Boyle, Michael, Christopher Edwards, and Saul Greenberg. "The effects of filtered video on awareness and privacy." In Proceedings of the 2000 ACM conference on Computer supported cooperative work, pp. 1-10. ACM, 2000.

[4] Newton, Elaine M., Latanya Sweeney, and Bradley Malin. "Preserving privacy by de-identifying face images." IEEE transactions on Knowledge and Data Engineering 17, no. 2 (2005): 232-243.

[5] Mosaddegh, Saleh, Loic Simon, and Frédéric Jurie. "Photorealistic face de-identification by aggregating donors' face components." In Asian Conference on Computer Vision, pp. 159-174. Springer, Cham, 2014.

[6] Du, Liang, Meng Yi, Erik Blasch, and Haibin Ling. "GARP-face: Balancing privacy protection and utility preservation in face de-identification." In IEEE International Joint Conference on Biometrics, pp. 1-8. IEEE, 2014.

[7] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In Advances in neural information processing systems, pp. 2672-2680. 2014.

[8] Oussidi, Achraf, and Azeddine Elhassouny. "Deep generative models: Survey." In 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), pp. 1-8. IEEE, 2018.

[9] Meden, Blaž, Žiga Emeršič, Vitomir Štruc, and Peter Peer. "k-Same-Net: k-Anonymity with Generative Deep Neural Networks for Face Deidentification." Entropy 20, no. 1 (2018): 60.

[10] Brkic, Karla, Tomislav Hrkać, Zoran Kalafatić, and Ivan Sikirić. "Face, hairstyle and clothing colour de-identification in video sequences." IET Signal Processing 11, no. 9 (2017): 1062-1068.

[11] Brkic, Karla, Ivan Sikiric, Tomislav Hrkac, and Zoran Kalafatic. "I know that person: Generative full body and face de-identification of people in images." In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1319-1328. IEEE, 2017.

[12] Meden, Blaž, Refik Can Mallı, Sebastjan Fabijan, Hazım Kemal Ekenel, Vitomir Štruc, and Peter Peer. "Face deidentification with generative deep neural networks." IET Signal Processing 11, no. 9 (2017): 1046-1054.

[13] Wu, Yifan, Fan Yang, Yong Xu, and Haibin Ling. "Privacy-Protective-GAN for Privacy Preserving Face De-Identification." Journal of Computer Science and Technology 34, no. 1 (2019): 47-60.

[14] Sun, Qianru, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. "Natural and effective obfuscation by head inpainting." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5050-5059. 2018.

[15] Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).

[16] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

[17] Yuan, Xiaoyong, Pan He, Qile Zhu, and Xiaolin Li. "Adversarial examples: Attacks and defenses for deep learning." IEEE transactions on neural networks and learning systems (2019).

[18] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." arXiv preprint arXiv:1607.02533 (2016).

[19] Rozsa, Andras, Ethan M. Rudd, and Terrance E. Boult. "Adversarial diversity and hard positive generation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 25-32. 2016.

[20] Liu, Ziwei, Ping Luo, Xiaogang Wang, and Xiaoou Tang. "Deep learning face attributes in the wild." In Proceedings of the IEEE international conference on computer vision, pp. 3730-3738. 2015.

[21] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition." In bmvc, vol. 1, no. 3, p. 6. 2015.

[22] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).