Improved Facial Mask-Based Adversarial Attack for Deep Face Recognition Models

Haoran Wang

The School of Natural and Computing Sciences, University of Aberdeen, King's College, Aberdeen, AB24 3FX, United Kingdom
u13hw21@abdn.ac.uk

Abstract. This paper explores the enhancement of security and robustness in the field of facial recognition by investigating adversarial example attacks. The author not only introduces an advanced adversarial example generation technique by utilizing key facial landmarks, but also investigates universal mask-based adversarial example generation strategy. These research efforts increase the precision and efficiency of attacks and extend the scope, affecting a broader range of users. Through extensive experimental setups with the Residual Network (ResNet)-50 model and the Chinese Academy of Sciences (CASIA) Face Image Database Version 5.0 (CASIA-FaceV5), this paper assesses the effectiveness of the proposed methods under different attack scenarios and various evaluation criteria, such as L0, L1 norms, and the Structural Similarity Index. These results demonstrate that mask-based attacks and universal perturbations significantly reduce recognition accuracy while maintaining the concealment of the examples. This study emphasizes the security aspect of current facial recognition technology, which has profound implications for the safety of digital life.

Keywords: Adversarial Attack, Face Recognition, Deep Learning

1 Introduction

Facial recognition technology, as a biometric technology based on artificial intelligence, has been widely applied in every aspect of people's daily life [1]. Not only in mobile phone unlocking, access control system, etc., but also in finance, justice, security monitoring and other fields. In 2022, the global market for facial recognition technology was evaluated to be worth $5.15 billion. It was expected to expand at a compound annual growth rate (CAGR) of 14.9% from 2023 to 2030. However, as the facial recognition technology becomes more widely used, its security issues have become increasingly important, especially in the context of adversarial attacks [2,3]. Additionally, more and more research show that face recognition systems can be easily fooled by simple printed masks or specially crafted glasses. Such attack cases remind the shortcomings of traditional facial recognition technology in defending against adversarial attacks.

In today's facial recognition technology area, adversarial example attacks have emphasized the significance of security. And here are five main types of adversarial

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_73

example attacks. Goodfellow et al. first introduced the Fast Gradient Sign Method (FGSM) in 2014 to produce adversarial examples through a gradient update, which is the one of the earliest and most widely known adversarial example generation methods [4]. Because of its fast generation speed and simple implementation, it has become a benchmark method in adversarial example area. Kurakin et al. introduced the Basic Iterative Method (BIM) which enhanced the FGSM attack through multiple iterations [5]. Projected Gradient Descent (PGD) is an advanced method based on BIM (and FGSM) proposed by Madry et al. In this method after each perturbation step, the adversarial example is re-projected onto the specific function [6]. DeepFool proposed by Moosavi-Dezfooli et al. seeks the minimal perturbation to push images to decision boundaries [7]. Carlini et al.'s Carlini and Wagner Attack (C&W Attack) finely tunes perturbations to minimize detectability and ensure high success rates [8]. These methods not only deepen the understanding on model existing vulnerabilities, but also show the importance of improving the security of facial recognition technology.

Therefore, this paper aims to enhance the security and robustness of facial recognition technology by deeply exploring adversarial example attacks in facial recognition technology. The author proposed an improved technique for generating adversarial examples, which includes a method based on key facial region masks to enhance the precision and efficiency of attacks. Besides, the author adopted a universal mask-based strategy to extend the scope of attacks to a broader range of users. These research not only improve the development of adversarial example generation technology, but also contribute to enhancing the security and robustness of facial recognition systems.

2 Method

2.1 Mask-Based Adversarial Example Generation

Mask Convex Hull Extraction. In this study, the author employes the Dlib library's face detector to identify faces within images, which is important in recognizing impressionable regions to adversarial attacks. The shape_predictor_68_face_landmarks model is utilized to detect 68 facial landmarks in each image, including critical areas such as the eyes, nose, mouth, and jawline. These landmarks are then used to create convex hulls with OpenCV's cv2.convexHull, defining regions for adversarial attacks. These hulls indicate the most influential facial features for recognition algorithms, ensuring that the attack is targeted. Each extraction process can be visualized by using OpenCV to verify accuracy and adjust the detection parameters accordingly.

Gradient Attack Execution. In this study, the author mainly employes gradient-based attack algorithms including FGSM and I-FGSM to operate pixels within the convex hull [9]. By calculating the gradient of the loss function in relation to the input image, these attack algorithms are able to determine optimal perturbation directions. The result of the function can be utilized to find out the disturbance in image that can misdirect classifier into mistakenly detecting the face. FGSM calculates these perturbations by first determining the loss function's gradient due to the image. The aim of it is assessing

how each change would affect the output. It then applies a sign function to these gradients, creating a vector that points to the direction of maximum increase in loss. This vector is controlled by epsilon which decides the magnitude of the perturbation. The epsilon not only ensures that the modifications are subtle enough to remain undetected by human, but also enough to mislead the classifier. The difference between FGSM and others is that the FGSM applies only single step update using the gradient sign, while I-FGSM introduces iterative process which can gradually approach the decision boundary. In order to ensure the naturalness of the area without convex hull, perturbations are only incrementally applied within the convex hull in the process of I-FGSM. Parameters such as step size, iteration number, and perturbation magnitude (epsilon) are iteratively tuned to balance the attack efficacy against concealment. The attack's effectiveness is assessed by its capacity to reduce recognition accuracy while maintaining the concealment of the adversarial examples.

2.2 Universal Mask-Based Adversarial Example Generation

This approach introduces a universal adversarial perturbation that attacks multiple users in a dataset by applying a single perturbation vector across various images. This process begins with initializing a zero tensor for the perturbation as the initial state for iterative updates. It will be iteratively updated using an optimization loop that processes batches of images from the data loader. During each iteration, the facial mask generated via Dlib's detectors focuses the perturbation on facial landmarks. The perturbation is refined through a series of gradient descent steps managed by the Adam optimizer. The Adam optimizer, an extensively employed optimization algorithm, is famous for its ability to effectively manage sparse gradients and its adaptive learning rate capabilities. Adam optimizer adapts the learning rates based on the averages of recent gradients for the weights. This approach allows for smaller optimization steps when gradients are large to prevent overshooting, and larger steps when gradients are small to accelerate convergence.

Additionally, the author uses a learning rate scheduler to continuously modify the learning rate across training iterations. The scheduler reduces the learning rate according to a predetermined schedule to strike a balance between the adversarial example's concealment and attack effect. In this paper, the step attenuation strategy is adopted. Every 10 steps, the learning rate will be reduced by a fixed proportion, so as to refine the disturbance and ensure the optimal balance between the attack effect and the concealment of the adversarial example.

The entire process is iteratively refined, ensuring the universal perturbation develops robust capabilities to generalize across the dataset.

2.3 Evaluation Indexes

In this experiment, the author mainly focuses on two primary metrics to assess adversarial attack’s efficacy: the success rate of the attack and the quality of adversarial examples.

Success Rate. In this experiment, the "Success Rate" metric is used to

assess the potency of the adversarial attacks. The success rate in the experiment is described as the probability that model will correctly classify the adversarial examples. This metric is critical in assessing the effectiveness of the adversarial methods applied, as a lower success rate indicates a stronger adversarial capability.

L0, L1 Norms. This experiment calculates the L0 norm to count the number of pixels that have been changed, which gives a recognition of the sparsity of the perturbation. The L1 norm, summing up the absolute differences across all pixels, provides insight into the total magnitude of change.

Structural Similarity Index Measure (SSIM). This experiment also utilizes Structural Similarity Index as one of the assessment criteria of concealment. This metric measures the visual similarity between the original and adversarial examples. SSIM accounts for luminance, contrast, and structure of the image. It will output a value between -1 and 1, where 1 indicates perfect similarity. A higher SSIM score indicates that the adversarial image is harder to detect both visually and algorithmically.

3 Results

3.1 Experimental Setups

The experiment uses the ResNet50 network architecture to train the model due to its robustness and widely used in facial recognition tasks. It utilizes the CASIA Face Image Database Version 5.0 (CASIA-FaceV5) as the dataset, which includes a diverse range of images representative of real-world variations [10]. The experiment is implemented using many libraries such as OpenCV and Dlib for image processing and facial landmarks extraction. Aimed at avoiding the problem of over fitting, several data augmentation techniques are used in the training process such as resizing, random rotation, and color jitter, simulating real-world variations. Fig. 1 shows an example of the Mask Convex Hull Extraction.

Fig. 1. Example of the Mask Convex Hull Extraction .

3.2 Face Image Selection

From the dataset, the author randomly selects an individual face image as target for the attack. Each selected image is subjected to a no-target attack which the goal is to misdirect the model to mistakenly classify the image without specifying a particular targeted class.

3.3 Different Attack Modes

For each image, the "shape_predictor_68_face_landmarks" from Dlib library will first extract the key facial landmark, which is assigned as the region of attack. Attack tasks including usual attack, iteration attack, and universal mask attack are selected respectively for testing.

Usual Attack. In the usual attack, the experiment sets two attack methods: FGSM applied to the whole image and FGSM targeted only at facial landmarks, under the same condition to see the different effect.

Iteration Attack. In each iteration of the 10 iterations, the loss's gradient is calculated. The original image is modified by applying the sign of the gradient, which is multiplied by alpha(the step size for each iteration, calculated as epsilon divided by the number of iterations), to introduce perturbations. After each modification, the image is clipped to ensure the perturbations stay within the valid pixel range [0,1].

Universal Mask Attack. The experiment initializes a zero tensor for the perturbation, which is optimized by the Adam optimizer through several iterations in each batch of images. The optimizer focuses on maximizing misclassification through perturbations on detected facial landmarks. During the process, the optimizer's learning rate is modified by the learning rate scheduler at specified intervals (every 10 steps), reducing it by a factor of 0.99. The aim of it is fine-tuning the optimization through the process. Fig. 2 shows a representative example of the perturbation.

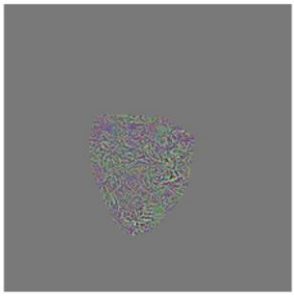

Fig. 2. Visualization of universal perturbation .

3.4 Result Analysis

In the experiment, the model's accuracy when exposed to adversarial examples from the CASIA-FaceV5 dataset is evaluated, and the correct classification rates are reported. Results are reported for epsilon in different conditions. Results are listed in Table 1. These consequences indicate a significant decline in model's accuracy with the increase in epsilon in all methods. The accuracy even drops significantly from 97.45% to a mere 7.32% as epsilon increases from 0.01 to 0.3 in whole page FGSM attack. These drastic reductions suggest that these attack strategies are capable of misdirecting the model into incorrect classification. Compared to FGSM attacks that focus on facial landmarks, global FGSM attacks typically achieve stronger impacts as they modify the entire image, potentially exploiting more vulnerabilities in the facial recognition system.

Compared to the FGSM, I-FGSM (both focusing on facial landmarks) is more effective because it applies FGSM multiple times, making small adjustments each time. This allows for finer manipulation to find the smallest perturbations, resulting in better effectiveness.

Table 1. Model sensitivity to adversarial attacks

	Raw Accuracy	Epsilon 0.01	Epsilon 0.05	Epsilon 0.1	Epsilon 0.3
FGSM (whole page)	97.45%	94.04%	51.12%	20.01%	7.32%
FGSM (facial landmarks)	97.45%	95.41%	91.33%	56.07%	20.43%
I-FGSM	97.45%	91.33%	75.15%	44.90%	18.30%
Universal Attack	97.45%	96.34%	90.47%	58.22%	46.29%

The universal attack performs not as well as other three methods in different epsilon since its broader applicability across different images. Selecting a group of pictures to experiment, the average model identification accuracy can drop to 55.82% while maintaining a great concealment.

The concealment of the adversarial examples is assessed with the L0, L1 norms, and SSIM. Results are shown in Table 2. While the attack proves effective in decreasing the performance of the model, the concealment of the adversarial images, as the SSIM index in this experiment, remains high (0.99 for both FGSM with facial landmarks and I-FGSM, 0.98 for Universal Attack). But the FGSM on whole page only gets 0.36 on same condition. This suggests that despite the successful deception of the model by attacks on facial landmarks, the visual quality of the images remains high. The visualization comparison is demonstrated in Fig. 3.

The L0 and L1 norms further prove these findings. The number of pixels changed (L0) and the total magnitude of changes (L1) are relatively low for the facial landmarks targeted FGSM and I-FGSM attacks compared with the FGSM applied to whole images.

Table 2. Performance measured by L0 norm, L1 norm, and SSIM.

	L0 norms	L1 norms	SSIM
FGSM(whole page)	50175	1883223	0.36
FGSM(facial landmarks)	20982	106789	0.99

I-FGSM	20901	98981	0.99
Universal Attack	40147	252830	0.98

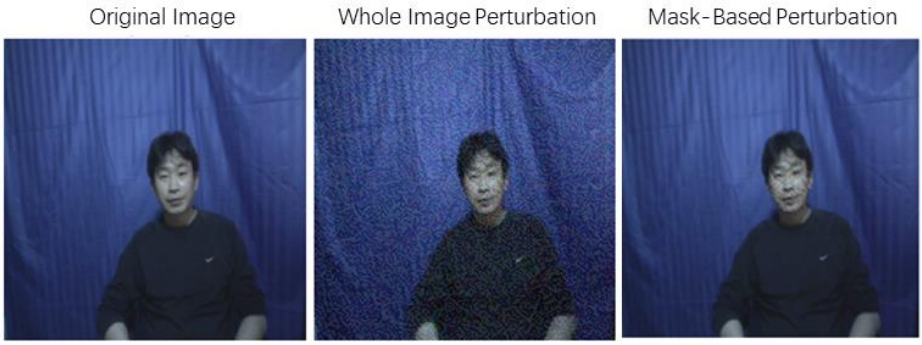

Fig. 3. Comparison of whole image and the proposed mask-based perturbation, with epsilon equals 0.05.

4 Discussion

In the experiment, the author tested the model's accuracy against adversarial examples across different attack methods. The result shows that accuracy significantly decreases as the epsilon increases. For example, under the FGSM attack on facial landmarks, accuracy decreases from 97.45% to 20.43% as epsilon increased from 0.01 to 0.3. This steep decline indicates that these attack methods effectively misdirect the model into incorrect classifications.

Compared to the standard FGSM attack, the FGSM and I-FGSM on facial landmarks effectively reduce the model's recognition accuracy while maintaining high stealthiness. According to the SSIM index, attacking methods based on facial landmarks demonstrate high stealthiness (SSIM value of 0.99), compared with the traditional FGSM, which shows much lower stealthiness (SSIM value of 0.36). This suggests that attacks focus on facial features are less detectable to the naked eye yet still effectively reduce recognition accuracy.

While the attack methods demonstrated in this study are proven effective, they still have their limitations. For example, the universe attack perturbation strategy though quite adaptable across multiple targets in test, does not perform as well under different epsilon values compared to other methods. It still needs further research to explore how to optimize this strategy to enhance its effectiveness in broader applications.

Moreover, the current attack models focus primarily on the effectiveness against a single model. Future endeavors could consider multi-model or cross-model attack strategies to enhance the generalizability of adversarial examples. By employing ensemble learning techniques, effective attacks against various facial recognition models could be achieved, significantly improving the universality of adversarial examples.

5 Conclusion

This paper utilized a generation strategy based on facial landmarks and universal mask. The author found that, compared to the normal strategy, these strategies based on facial landmarks can keep a great concealment while maintaining an effective misdirection. By iteratively optimizing the mask across the entire dataset, this paper found out a universally applicable perturbation capable of attacking on multiple targets. This universe attack strategy makes the attack no longer limited to a single target user, but can widely affect multiple users which greatly expand the scope of the attack. It means the perturbation vector can be added to different input data and misdirect a well-trained model about these perturbed data.

In the future, the author will try to combine the mask-based adversarial examples and the universal mask with the concept of ensemble learning. A multi-model attack strategy will be employed to promote adversarial example's transferability. By attacking multiple different face recognition models, the adversarial example can achieve efficient attacks which significantly improve the universality of adversarial examples.

References

1. Kaur, P., Krishan, K., Sharma, S. K., & Kanchan, T.: Facial-recognition algorithms: A literature review. Medicine, Science and the Law, **60**(2), 131-139 (2020).
2. Sati, V., Garg, D., Choudhury, T., & Aggarwal, A.: Facial recognition-application and future: A review. In 2018 International Conference on System Modeling & Advancement in Research Trends, 231-235 (2018).
3. Lai, X., & Rau, P. L. P.: Has facial recognition technology been misused? A public perception model of facial recognition scenarios. Computers in Human Behavior, **124**, 106894 (2021).
4. Goodfellow, I. J., Shlens, J., & Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014).
5. Kurakin, A., Goodfellow, I. J., & Bengio, S.: Adversarial examples in the physical world. In Artificial intelligence safety and security, 99-112 (2018).
6. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A.: Towards Deep Learning Models Resistant to Adversarial Attacks. In International Conference on Learning Representations, 1-28 (2018).
7. Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2574-2582 (2016).
8. Carlini, N., & Wagner, D.: Towards evaluating the robustness of neural networks. In 2017 ieee symposium on security and privacy, 39-57 (2017).
9. Kurakin, A., Goodfellow, I., & Bengio, S.: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236 (2016).
10. CASIA Face Image Database Version 5.0. URL: http://english.ia.cas.cn/db/201610/t20161026_169405.html. Last Accessed 2024/04/26.