Deceiving Face Recognition Neural Network with Samples Generated by Deepfool

Jingsong Xue, Yu Yang, and Dongsheng Jing
State Grid Jiangsu Electric Power Limited Company, Suzhou Power Supply Branch
No. 555 Laodong Road, Suzhou Jiangsu 215004, China
Email: 6802569@qq.com
Email: yyjady@qq.com
Email: jds19810119@163.com

Abstract. Image-based identity authentication systems have been extensively used recently. However, it holds the risk of illegal use of the images for authentication by others, which will cause a severe threat to personal privacy and security protection. The concept of adversarial samples, which is often utilized in the field of deep learning to deceive classification models, has become more and more popular, and has accumulated lots of significant efforts. Meanwhile, previous work mainly focused on interpreting images by features rather than by the structural characteristics of the human face, which made the deception of face recognition model ineffective. To solve the problem, face recognition neural network deceiving method that is based on Deepfool algorithm is proposed. For a specific face recognition neural network, white-box attack is used to generate adversarial samples, and Euclidean distance is utilized to optimize adversarial samples to obtain face photos with misleading attributes. The proposed approach can be used for privacy and security protection. Experiments on several face testing data sets verify effeteness of our proposed approach.

1. Introduction
With the maturity of face recognition technology, the era of "face scan" is coming. face recognition technology combines with all walks of life, bringing more and more application scenarios. Although face recognition technology will bring changes to life, its threat to citizens' privacy protection is also worth considering. In particular, personal photos on social networks are often the most intense areas of privacy leakage. In the era of information as value, the commercial value of personal privacy information has become increasingly prominent. Enterprises can excavate more value according to the large amount of information collected. Once this information is not properly kept and leaked, the user's personal privacy will not be guaranteed, and even personal safety will be threatened. Therefore, the protection of personal privacy is becoming ever more important.

At present, the commonly used deceiving face recognition model is a kind of deep learning technology called adversarial sample[1], which confronts two artificial intelligence algorithms and generates similar samples to confuse the face recognition model. Szegedy et al.[2] proved for the first time that misclassification can be made by adding a small amount of human-imperceptible noise to the image. Fast Gradient Sign Method (FGSM) algorithm proposed by Goodfellow et al.[1] can deceive neural networks by adding noise linearly to generate countermeasure samples in high-dimensional space. Kurakin et al.[3] proposed BIM algorithm, which increases the loss function of the classifier and disturbs the image by iterating along a certain direction. Papernot et al.[4] introduced Jacobian-based Saliency Map Attack (JSMA) algorithm to generate adversarial samples by restricting the $l_0$ norm of the disturbance and only changing several variables in the image, rather than disturbing the whole
image, to deceive the classifier. Su et al.\[5\] put forward One Pixel Attack algorithm, which uses differential evolution algorithm to change only one pixel value in the image to resist attacks.

Although the methods mentioned above can produce adversarial samples from and improve the quality of adversarial samples and the effect of confusing classifiers, the classical FGSM image perturbation algorithm is to disturb images at the classification layer\[6\], and it needs to be considered that the size of the disturbance factor is determined, resulting in the false classification effect of face recognition is not obvious. While Deepfool\[7\] image perturbation algorithm can generate approximate minimum interference to the image, and confuse the classifier. Therefore, we proposes face recognition neural network deceiving method which is based on Deepfool for FaceNet\[8\], confuses the FaceNet classifier, and uses Euclidean distance\[9\] to evaluate the accuracy of misclassification in FaceNet embedding\[10\] layer. Experiments demonstrate that the method based on Deepfool is effective.

2. Related Works

2.1. Triplet Loss

Triplet Loss\[8\] is a loss function in deep learning. It is used to train samples with less difference, such as faces. Feed data includes Anchor samples, Positive samples and Negative samples. By optimizing the distance between anchor samples and positive samples is less than that between anchor samples and negative samples, the similarity of samples is calculated. Principle of Triplet Loss is shown as Figure 1.

\[
L = \sum_{i} \left[ \|f(x_i^a) - f(x_i^p)\| - \|f(x_i^a) - f(x_i^n)\| + a \right]^{+}
\]

In the Equation 1, for each element in Triplet Loss triple\[11\], a parameter-sharing or non-sharing network is trained, and the feature expressions of three elements are obtained, which are recorded as \(f(x_i^a)\), \(f(x_i^p)\) and \(f(x_i^n)\) respectively. Through learning, the distance between feature expression of \(x_i^a\) and \(x_i^p\) is as small as possible, while the distance between feature expression of \(x_i^a\) and \(x_i^n\) is as large as possible, and the distance between \(x_i^a\) and \(x_i^p\) and between \(x_i^a\) and \(x_i^n\) should have a minimum distance \(a\). So the distance is calculated by Euclidean distance. The notation of + at the lower right corner is given in the maximum value. When the bracket internal value is positive, the value is loss. When the bracket internal value is negative, the value is 0, which is loss.

2.2. FaceNet

FaceNet proposed by Schroff et al.\[8\] is distinct from the current universal face recognition model. The existing face recognition model based on deep neural network\[12\] uses classification layer: the middle layer is the vector mapping of face image, and the classification layer is the output layer. The disadvantage of this method is that it is not direct and inefficient. Unlike current methods, FaceNet uses convolution neural network\[13\] learning to map images to Euclidean space, and uses Loss function of Triplets-based Largest Margin Nearest Neighbour (LMNN)\[14\] Classification to train the neural network.
network, which directly outputs 128-dimensional vector space. The selected triplets contain two matched face thumbnails and one unmatched face thumbnail. The goal of loss function is to distinguish positive and negative classes by distance boundary. The structure of FaceNet is shown as Figure 2.

![Figure 2. Structure of FaceNet](image)

As shown in the Figure 2, the first half of the model is a traditional convolutional neural network. DEEP ARCHITECTURE generally uses Inception network [15], normalizes it before calculating the two norms, establishes the embedding space, and uses triplets to classify the end of the model structure directly.

2.3. Deepfool
Since Szegedy et al. [2] first discovered that deep neural network image classification domain is vulnerable to the attack of adversarial samples, researchers have proposed several methods to generate countermeasure samples. These countermeasure samples have only a slight disturbance, so that the human visual system cannot detect the disturbance, but such an attack will lead to a complete change in the classification results of the images by the neural network.

The disturbance factor in FGSM is set artificially. When the activation function is a linear function, such as softmax [16] and rulu [17], the gradient is constant, so FGSM is always effective. When the activation function is not a linear function, FGSM cannot generate countermeasure samples. Therefore, to solve this problem, Moosavi-Dezfooli et al. [7] proposed Deepfool, which generates minimum adversarial disturbance by iteration method, and pushes the image located in the classification boundary step by step outside the boundary until the wrong classification occurs.

3. Methods

3.1. VGGface2
VGGface2 [18] data set is a data set that can be used to recognize faces of different postures and ages. The VGGface2 data set contains 9131 identities, with a total of 3.31 million photos. The data set is roughly gender-balanced, with 59.3% males and 362.6 photos per identity on average. The pose, age, light and background of each identity's photograph vary greatly, which ensures the comprehensiveness of the data set in the face recognition process.

The whole data set includes training set and test set, in which there are 8631 identities in training set and 500 identities in test set. The training set is not intersected with test set. The VGGface2 data set collects photos of each identity in different environments, and provides a high-precision image data set for model training.

3.2. Data Preprocessing
In the VGGface2 data set, each photo is a color photo, with three channels. The size of each photo is different, and the color of each photo is complex, which results in that the sample data are not in the same scale, and the training time is increased. At the same time, it may make the neural network unable to converge. Therefore, it is necessary to normalize the image before training.

Data normalization is to scale the data into a small specific interval to solve the comparability between data indicators. The common normalization methods are Min-Max Normalization and Z-score. The former is a linear transformation of the original data, and the latter is a data standardization...
based on the mean and standard deviation of the original data. The Min-Max Normalization normalization equation as follows:

\[ X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \]  \hspace{1cm} (2)

Because the pixel values of the image are in the [0,255], the extremum values are 0 and 255 respectively, and the relationship between the original data needs to be preserved, the Min-Max Normalization normalization method is used to eliminate the influence of the quantity rigidity and the range of the pixel values. The normalization formula and the formula for restoring the pixel value as follows:

\[ X_{\text{norm}} = \frac{X \times 2.0}{255.0} - 1.0 \]  \hspace{1cm} (3)
\[ X = \frac{(X_{\text{norm}} + 1.0) \times 255.0}{2.0} \]  \hspace{1cm} (4)

3.3. Model Structure

Current face recognition models based on deep learning generally use classification layer. The middle layer is the vector mapping of face images, and then the classification layer is used as the output layer to output the results of face classification. The disadvantage of this method is that it is not direct and inefficient\[8\]. FaceNet outputs 128-dimensional space vectors directly, and calculates the accuracy of current classification through Euclidean distance. FaceNet is classified according to the distance calculation of the final 128-dimensional space vector.

The Deepfool algorithm is used to attack the FaceNet network in white-box attack\[19\]. Traditional methods use category correctly to evaluate the effect of Deepfool algorithm on image classifier. Because of the characteristics of classification layer, it is impossible to obtain the distance between current countermeasure samples and target classes, which makes Deepfool algorithm unable to update step by step when generating adversarial samples, and can only obtain a mutation result. Therefore, the embedding layer of FaceNet is proposed as the classification criterion, and the embedding layer of the adversarial sample and the target class is used as the Euclidean distance to calculate the distance between the adversarial sample and the target class. The loss function as follows:

\[ \text{loss} = \sqrt{(\text{adv}_\text{emb} - \text{target}_\text{emb})^2} \]  \hspace{1cm} (5)

where adv_emb in the formula represents the output of the embedding layer of the adversarial sample, and target_emb represents the output of the embedding layer of the target class. The final loss is a scalar. The flow chart is shown as Figure 3.

![Figure 3. Illustration of procedure of deceiving face recognition neural network with samples](image-url)
In Figure 3, the process of input image represents the input image, Target image represents the target category image, Input image generates Adv image of the adversarial sample through Deepfool, and then makes the Euclidean distance between the embedding layer of the adversarial sample and the embedding layer of the target category image as the loss function of Deepfool, continuously optimizes the parameters of Deepfool, and finally generates the qualified adversarial sample Adv image. The algorithm of deceiving face recognition neural network is as algorithm 1.

Algorithm 1: Deceiving face recognition neural network
Input: Input_pic Ip, Target_pic Tc
Output: Adv_image Ai
Initialize: Ip, Tc
1. loss_limit ← 0
2. adv_x ← Deepfool(Ip)
3. While num_iter do:
   Ai, adv_loss ← FaceNet(adv_x)
   if (adv_loss – last_adv_loss) < loss_limit
      break
   last_adv_loss ← adv_loss
4. end while
5. return Ai

4. Experiment and Analysis
Through the research of the method, we use Deepfool algorithm for FaceNet. To prove the effectiveness of this method, the traditional image perturbation method FGSM is selected as a comparison. Because of the characteristics of FGSM and Deepfool algorithms, only one picture can be disturbed to another picture at a time, and the image can be misclassified into target categories through FaceNet through the loss function optimization algorithm.

The data set is based on 5% of the test set of VGGFace 2, which contains 20 identities. The data set is cross validated with 400(20*20). Therefore, the data set for this experiment contains 400 identities, training data to accumulate the loss, running time and image bias in each optimization process as validation indexs.

In the experiment, in order to ensure the similarity of image deviation between the two methods, the parameters are set as follows. The threshold of the loss function based on Deepfool method is 0.008. When the loss is less than this value, the condition can be satisfied and the iterative optimization can be withdrawn. The maximum number of iterations for Deepfool is 2000. The threshold of loss function based on FGSM is 0.006. The maximum number of iterations for FGSM is also 2000. The experimental results are shown as Table 1.

Table 1. Experimental result of Deepfool and FGSM

|        | Distance | Bias |
|--------|----------|------|
| Deepfool | 0.00418  | 22.33|
| FGSM    | 0.06615  | 21.05|

There are two methods in the table. One is deceiving FaceNet based on Deepfool and the other is deceiving FaceNet based on FGSM. DISTANCE represents the Euclidean distance between the input image and the embedded layer of the target category image, which is used to represent the accuracy of classification. BIAS represents the average distance between the adversarial sample and the original image.

As shown in Figure 4, it represents the time taken to process each image in the Deepfool and FGSM algorithms. The method based on Deepfool costs 45.1 seconds to process each picture, and the method based on FGSM costs 14.2 seconds to process each picture.
Figure 4. Time cost of our method vs FGSM

It can be seen from the Table 1 and Figure 4 that although the Deepfool-based method spends more time on each image processing, the accuracy of the method based on Deepfool is obviously better than FGSM in the case of similar image disturbance. The method based on Deepfool needs to find the smallest disturbance, which makes Deepfool-based methods take longer.

5. Conclusion
The development of face recognition technology brings convenience to life, but at the same time it also has the hidden danger of privacy leakage. At present, the classical image perturbation algorithm is FGSM, but because of the shortcomings of FGSM in accuracy and perturbation factor, FGSM is seldom used in face recognition.

To solve these problems, we propose deceiving face recognition method based on Deepfool. FaceNet is used to generate adversarial samples, which makes it impossible for FaceNet to distinguish the identity of people in pictures. Experiments show that the accuracy of the deceiving face recognition method based on Deepfool is due to the method based on FGSM. With the acceleration of the process of social intelligence, the protection of privacy must be paid more and more attention. The possession of deceiving face recognition method is bound to play an important role in the future ethics of science and technology.

6. References
[1] Goodfellow I J, Shlens J, Szegedy C, et al. Explaining and Harnessing Adversarial Examples[J]. international conference on learning representations, 2015: 1-8.
[2] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. international conference on learning representations, 2014: 1-10.
[3] Kurakin A, Goodfellow I J, Bengio S, et al. Adversarial examples in the physical world[J]. international conference on learning representations, 2017: 1-14
[4] Papernot N, McDaniel P D, Jha S, et al. The Limitations of Deep Learning in Adversarial Settings[J]. ieee european symposium on security and privacy, 2016: 372-387.
[5] Su J, Vargas D V, Sakurai K, et al. One Pixel Attack for Fooling Deep Neural Networks[J]. IEEE Transactions on Evolutionary Computation, 2019: 1-1.
[6] Jean S, Cho K, Memisevic R, et al. On Using Very Large Target Vocabulary for Neural Machine Translation[J]. international joint conference on natural language processing, 2015: 1-10.
[7] Moosavidezfouli S, Fawzi A, Frossard P, et al. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks[J]. computer vision and pattern recognition, 2016: 2574-2582.
[8] Schroff F, Kalenichenko D, Philbin J, et al. FaceNet: A unified embedding for face recognition and clustering[J]. computer vision and pattern recognition, 2015: 815-823.
[9] Legendre P, Gallagher E D. Ecologically meaningful transformations for ordination of species data[J]. Oecologia, 2001, 129(2): 271-280.
[10] Jacob Y, Denoyer L, Gallinari P, et al. Learning latent representations of nodes for classifying in heterogeneous social networks[C]. web search and data mining, 2014: 373-382.

[11] Wu D, Zheng S, Bao W, et al. A novel deep model with multi-loss and efficient training for person re-identification[J]. Neurocomputing, 2019: 69-75.

[12] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[J]. national conference on artificial intelligence, 2016: 4278-4284.

[13] Krizhevsky A, Sutskever I, Hinton G E, et al. ImageNet Classification with Deep Convolutional Neural Networks[J]. neural information processing systems, 2012, 141(5): 1097-1105.

[14] Weinberger K Q, Saul L K. Distance Metric Learning for Large Margin Nearest Neighbor Classification[J]. Journal of Machine Learning Research, 2009: 207-244.

[15] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[J]. computer vision and pattern recognition, 2015: 1-9.

[16] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[J]. computer vision and pattern recognition, 2016: 2818-2826.

[17] Glorot X, Bordes A, Bengio Y, et al. Deep Sparse Rectifier Neural Networks[C]. international conference on artificial intelligence and statistics, 2011: 315-323.

[18] Cao Q, Shen L, Xie W, et al. VGGFace2: A Dataset for Recognising Faces across Pose and Age[J]. ieee international conference on automatic face gesture recognition, 2018: 67-74.

[19] Dong Y, Liao F, Pang T, et al. Boosting Adversarial Attacks with Momentum[J]. computer vision and pattern recognition, 2018: 9185-9193.