Adversarial Examples for Silhouette-based Gait Recognition Networks

Cong Lin, ZiJia Chen, Yifan Xu
{cl3954,zc2521,yx2502}@columbia.edu

Abstract

The development of deep neural networks for gait recognition in recent years provides a new approach to building gait recognition security systems. However, it has been demonstrated that deep neural networks are vulnerable to adversarial example attacks, which will jeopardize the whole security system. In this paper, we particularly study and design adversarial examples against silhouette-based gait recognition neural networks typical of models in the research field. By adopting Fast Gradient Sign Method, we produce imperceptible perturbation added to the source silhouette frames or gait energy images and cause the neural network to return the target prediction. We also try to combine adversarial example methods for attacking segmentation neural networks to break down the entire gait recognition security system. We carry out experiments on the widely used gait dataset CASIA and our own gait video, with current results justifying the practicability of our methods.

1. Introduction

Gait recognition has been frequently recognized as a step beyond facial recognition. Models for this task are being built in multiple crucial safety fields such as recognizing criminals from a footage video, identifying people that look similar in airport security, etc. In the past few years, with the fast development of deep learning, deep neural network models have been introduced to gait recognition, resulting in many novel gait recognition frameworks. Due to the powerful expressiveness of neural networks, performance in gait recognition tasks has been pushed to a new higher level, which is a guarantee for the reliability of neural-network-based gait recognition security systems.

Among different types of gait recognition neural network models, models that take silhouette frames or the deduced averaged gait energy images (GEI) as inputs are very common, and in this paper, we refer to them as silhouette-based gait recognition network. Silhouette frames are cropped frames of gait record video in which figure of the person being recorded is labeled in white while the background is black. They are generated using image segmentation models. By averaging several consecutive silhouette frames, we can get the gait energy image which describes the cumulative spatial distribution of the person's movement. Take the silhouette frames or GEI as input, silhouette-based gait recognition networks extract latent features underneath and return the corresponding feature vector. In convention, the feature vector will be compared with those in the database in metrics like Euclidean distance or Chebyshev distance for estimating identity. By elaborating objective function in the training phase, the recognition network can learn to generate feature vectors that are sufficient for identifying and distinguishing people, which is promising for applications in practice.

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

However, despite their performance and unprecedented progress made in a variety of fields, neural networks are justified to be lack of robustness under the attack of adversarial examples. That is, there exist some slightly modified source inputs that can cause the neural network to incorrect outputs or even any desired outputs. This indicates that, in the case of gait recognition, we can generate adversarial silhouette frame examples or adversarial GEI examples to hack a neural-network-based gait recognition security system. If this is practicable, then it will be a devastating flaw in such a security system.

The key points in this adversarial example attack task are that: 1) the adversarial example should be indistinguishable from the original image for human eyes, namely making the perturbation imperceptible, 2) the adversarial example is not restricted to only be generated for a single image, but rather for a sequence of frame images, 3) in order to hack the entire system, the adversarial example in real-life source video level needs to fool both the segmentation network and the gait recognition network, which can be two non-end-to-end models.

In this paper, we focus on these problems and study the adversarial examples against gait recognition networks. To achieve the final target of generating real-life adversarial gait record video examples, we divide the task into the following two parts and tackle them one by one:

- Based on the silhouette frames or the derived GEI of a source person and that of a target person, generate an adversarial example for the source person in which he will be identified as the target person by the neural network.
- Based on a gait record video of a source person and the silhouette frames or the derived GEI of a target person, generate an adversarial example for the source person in which he will be identified as the target person by the neural network.

Our algorithm for generating adversarial examples mainly based on Fast Gradient Sign Method (FGSM). We first use a desirable target to train the adversarial examples against

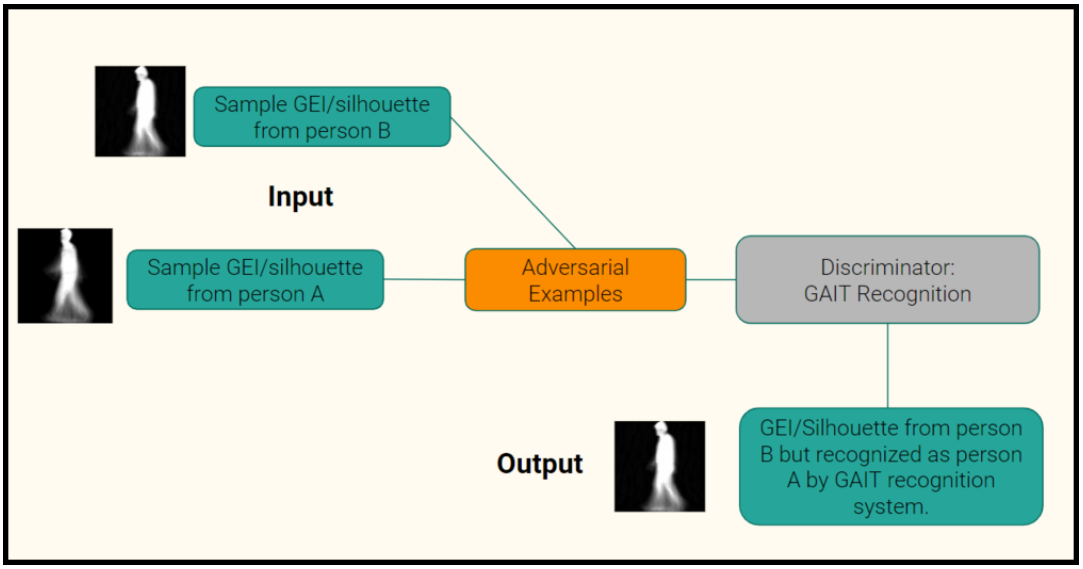

Figure 1: Adversarial Examples to GAIT Recognition Algorithm

the recognition network and then take the resulting examples as a target to train the adversarial examples against the segmentation network.

To sum up, our contributions in this paper are listed below:

- We point out the danger of adversarial examples to neural-network-based gait recognition security system.
- We generate adversarial examples against the silhouette-based gait recognition.
- We devise an algorithm for generating adversarial examples against the non-end-to-end segmentation-recognition models.

2. Related Works

2.1. Gait Recognition Networks The cross-view gait methods can be categorized into three types: features invariant model, 3D construction model, and View Transformation Model (VTM). [10] proposed an updated feature selection solution which reduced misclassification for the view-invariant gait recognition problem. [11] proposed an arbitrary view gait recognition method where the gait recognition is performed in 3D to be robust to variation in speed, inclined plane and clothing, and in the presence of a carried item. The most prevailing one now is View Transformation Model which can project a large, dimensional, multi-view gait data into a lower-dimensional feature space such as silhouette and GEI image (Gait energy image) that has sufficient discriminative capability. [4] developed a specialized deep CNN architecture for Gait Recognition which is less sensitive to several cases of common variations and occlusions that affect and degrade gait recognition performance.

2.2. General Adversarial Examples Generating adversarial examples for classification has been extensively studied. In [7], Adversarial examples were broadly discussed. An adversarial example is a sample of input data which has been modified very slightly in a way that is intended to cause a machine learning classifier to misclassify it. In many cases, these modifications can be so subtle that a human observer can not notice it at all, yet the machine learning classifier still makes a mistake. [7] majorly demonstrated how to feed adversarial images obtained from a cell-phone camera to an ImageNet Inception classifier and measure the classification accuracy of the system. [8] proposed a simple and fast yet more accurate gradient sign method to generate adversarial examples based on the linear nature of CNN.[9] efficiently trained feed-forward neural networks in a self-supervised manner to generate adversarial examples for a particular target model or a particular set of networks.

2.3 Adversarial Examples for Semantic Segmentation and Object Detection [1] proposed an algorithm named Dense Adversary Generation (DAG) to generate a large family of adversarial examples and apply them to a wide range of deep networks for segmentation and detection. [1] elaborated a possibility of adversarial perturbations being transferred across networks with different training data, based on different architectures, and even for different recognition tasks. [2] proposed a way of how existing adversarial attackers can be transferred to the task of semantic segmentation and the possibility of creating imperceptible adversarial perturbations that lead a deep network to misclassify almost all pixels of a chosen class while leaving network prediction nearly unchanged outside of that class.

2.4 Adversarial Examples on Gait Recognition system There are some prior experiments on Adversarial Examples to Gait Recognition system before. [12] proposed the vulnerability of the motion sensor-based gait authentication algorithm with adversarial perturbations, obtained via the simple fast-gradient sign method. However, its mainly fo-

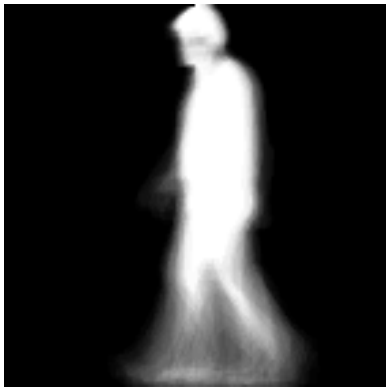

Figure 2: GEI Image from person B

Figure 3: GEI Image from person A

Figure 4: GEI Image from person B but considered to be person A by gait recognition system

Figure 5: Noise we added to person B GEI image

cusing on mainly sensor-based Gait mechanism which has been tackled long while ago. [13] proposed a spoof attack on sensor-based gait authentication system on 2007 and it demonstrated that, in sensor-based gait recognition, attackers with knowledge of their closest person in the database can be a serious threat to the authentication system. On the contrary, within this article, we will be focusing on silhouette based gait recognition system with segmentation which claimed to be more advanced and accurate.

3. Methods

Most silhouette-based gait recognition networks transform the raw input into feature vector and compare feature vectors in terms of distance or similarity metrics to predict the person identity. If we denote by $x_1, ...x_n$ the raw input, f the gait recognition network, g the distance loss function, x a query input and $\hat{y} \in \{1, ...n\}$ is the corresponding predicted identity, then the procedure of the gait recognition network system can be summarized as:

$$\hat{y} = \arg\min_i g(f(x), f(x_i)).$$

Since neural networks are trained using backpropagation and are thereby differentiable, a practicable to generate adversarial example based on a raw input is to fix the network and use the objective function and backpropagation to update the raw input. A typical iterative method is the Fast Gradient Sign Method (FGSM) [8] which can be formulated as

$$\eta = \epsilon \cdot \text{sgn}(\nabla_x J(\theta, x, f(x'))),$$

where η is the update to the current x, x' is the target, ϵ is the learning rate and J is a loss function taking θ as parameters. The flowchart of this procedure is shown in Figure 1, and Figure 2 to Figure 5 are examples for the desirable adversarial examples.

To our concern, given a source image x, a target x' and a gait recognition network f_θ, we are to generate an adversarial exmaple $x + r$ such that $||f_\theta(x + r) - f_\theta(x')||_\infty$ is minimized while $||r||_\infty$ is small. To achieve this, we use algorithm based on FGSM and L_∞ norm constrain:

$$\begin{cases} r_{t+1} = r_t - \epsilon \cdot \text{sgn}(\nabla_{r_t}(||f(x + r_t) - f(x')||_\infty)) \\ \text{truncate elements of } r_{t+1} \text{ within } [-\eta, \eta] \\ t = t + 1 \end{cases}.$$

In each iteration, we first adopt FGSM to update the r on x, but this may cause too much modification as iterations go on. To address this issue, a clip operation is taken right after FGSM which ensure that the L_∞ norm of r is not larger than a given threshold η.

If the inputs to the gait recognition network are GEI images, then we have to use some tricks for obtaining the silhouette adversarial examples. In this case, we cannot directly use backpropagation to update the silhouettes because: (1) silhouettes are binary images. If we adopt FGSM, it may result in decimal pixel value. (2) the GEI image is the average of a sequence of silhouette, so the modification to GEI image can actually be distributed to any of the silhouette, which means there is no unique and robust solution.

A compromise solution is to distribute the noise randomly on corresponding number of silhouette frames. Suppose n silhouette frames are used for generating a GEI image. When generating an adversarial example GEI image, noise of value z is added to a particular pixel. Then we randomly pick nz silhouette frames where the corresponding pixel value is 1 if z is negative or 0 if z is positive, and inverse that pixel value as binary bit. In this way, we transfer the noise added on the GEI image to the silhouette frames and ensure that the resulting silhouette frames will give GEI image close to the target adversarial example.

To further attack the segmentation network so that the entire gait recognition security system can be broken down, again we can just take the resulting adversarial silhouette frames as target and use FGSM to generate adversarial example based on real-life image. The main difference here compared with generating adversarial example against gait recognition network is that cross-entropy loss function, which is usually the objective for segmentation network, is taken as the objective instead of Chebyshev norm. That is, we are to generate an adversarial exmaple $x + r$ such that cross-entropy$(f_\theta(x + r), f_\theta(x'))$ is minimized while $||r||_\infty$ is small, where here f_θ is the segmentation network with parameter θ and $f_\theta(x')$ stands for the target adversarial silhouette frame.

4. Experiments

We carry out experiments on segmentation network FCN and GaitCNN[4] , gait recognition network taking GEI images as inputs, to rudimentarily examine the performance of our algorithm. The architecture of GaitCNN, as is shown in Figure 6, comprises of 8 layers, half of which are convolutional layers while the rest are pooling layers, and each layer consists of 8 feature maps. Those feature maps will eventually be passed through a fully connected layer together with a softmax function to give a prediction on the identity of the input (in Figure 6, the identity is one of the 124 people in the training set).

As for the attack on GaitCNN alone, we use the large multiview gait dataset CASIA-B. In this dataset, three variations, namely view angle, clothing and carrying condition changes, are separately included. We refer to the results in this experiment as experiment-1 and experiement-2.

Since the real-life video or image of the CASIA-B dataset is not provided, to carry out experiment of attacking the entire segmentation-recognition system, we record our own real-life gait videos. We are to use these videos to fool both the segmentation network and the recognition network. This experiment is referred to as experiment-3.

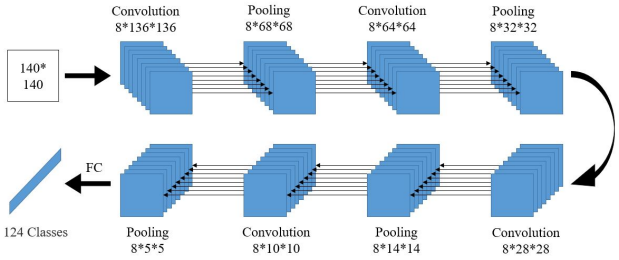

Figure 6: Architecture of GaitCNN

5. Results

The result is divided into three part: adversarial examples on GEI, adversarial examples on silhouette frames and adversarial examples on video frames. The evaluations are based on confidence of target and origin, and l_2norm of noise per pixel, which includes l_2norm_{GEI}: the l_2norm of noise added on GEI, l_2norm_{silh}: the average l_2norm of all the noise added on silhouettes, l_2norm_{video}: the average of l_2norm of all the noise added on video frames. We have conducted three experiments on first two-part, and one experiment on the third part. For experiment-1, the source image label is 13, the target image label is 75. For experiment-2, the source image label is 27, the target image label is 37. For experiment-3, the source image label is 125, the target image label is 127. The data in the first two experiments are from CASIA-B, The data in third experiments are from videos recorded by ourselves.

In the first part, we tried to attack GaitCNN by adding noise to the GEI, the direct input of GaitCNN. We conducted three experiments on it, by back-propagating the loss between Source Image (the adversarial example we want to create from) and Target Image (the target of adversarial example) to the source image, we successfully fooled the Gait Recognition system by letting the system recognize source as the target. Also, the created adversarial examples cannot be discerned by a human. The source image, added noise, adversarial image, and target image are in the below Figure and evaluation are in the below table(Figure 7-Adversarial Examples on GEI, Figure 8: Evaluation-l_2norm_{GEI}).

In the second part, we tried to attack the system by adding noise to the silhouette. Unlike the GEI which is a continuous-value image, silhouette image is a segmentation image whose adversarial examples cannot be simply created by back-propagating the noise of GEI to all the silhouette images. Thus, we randomly distributed the noise to different silhouette, but make the average noise just equal to the noise of GEI. The noise of GEI, the noise of silhouette, the origin silhouette and the adversarial examples of the silhouette are in the below image and the evaluation is in the below table(Figure 7-Adversarial Examples on silhouette frame, Fig-

ure 8: Evaluation-l_2norm_{silh})

In the third part, we tried to attack the system by adding noise to the video frames. We employed the fcn8s-net to segment human from the background from the video frames and used FGSM method to back-propagate the noise from every silhouette to their corresponding video frames, the result is in figure 7-Adversarial Examples on video frames, Figure 8: Evaluation-$l2norm_{video}$. Though we can successfully back-propagate the noise from silhouettes to the video frames, the resulted adversarial video cannot extract high-quality gait silhouette because of the added noise, and thus cannot be used make a good GEI and be recognized as a target by GaitCNN. The extracted silhouette is in Figure 7- Extracted silhouette from the adversarial video. From the figure, we can find that because of the influence of noise, the recognized human silhouettes are distorted, some are even incomplete like no arms or no legs. Thus we need to further implement a more sophisticated method to attack gait recognition on the video frame.

The related codes are posted at our Github:https://github.com/YifanPTAH/GaitRecFooler

6. Conclusions

We propose to generate adversarial examples against silhouette-based gait recognition networks so as to justify the vulnerability of gait recognition systems based on neural networks. Current results have shown that silhouette-based gait recognition network can be easily fooled by adversarial examples. To take one step further, we propose to generate adversarial examples to break down both the segmentation network and the recognition network, which are the two crucial components of a gait recognition security system. Since the two neural networks are two separate components on the pipeline, they are not an end-to-end model where we can do backpropagation. Still, we propose algorithms to generate adversarial examples for this non-end-to-end model, but due to the lack of consistency, the resulting real-life adversarial example videos fail to give a desirable performance. To tackle this problem, in the future, we will focus on devising a consistent algorithm for generating adversarial examples against non-end-to-end models. Besides, in the current phase, we only experiment on relatively simple models such as GaitCNN and FCN. To further justify the algorithm we are to propose, we will also try to attack more complicated and more powerful segmentation models and recognition models.

7.References

[1] Xie, Cihang, et al. *"Adversarial examples for semantic segmentation and object detection."* Proceedings of the IEEE International Conference on Computer Vision. 2017.

[2] Fischer, Volker, et al. *"Adversarial examples for semantic image segmentation."* arXiv preprint arXiv:1703.01101 (2017).

[3] Shiraga, Kohei, et al. *"GEINet: View-invariant gait recognition using a convolutional neural network."* 2016 international conference on biometrics (ICB). IEEE, 2016.

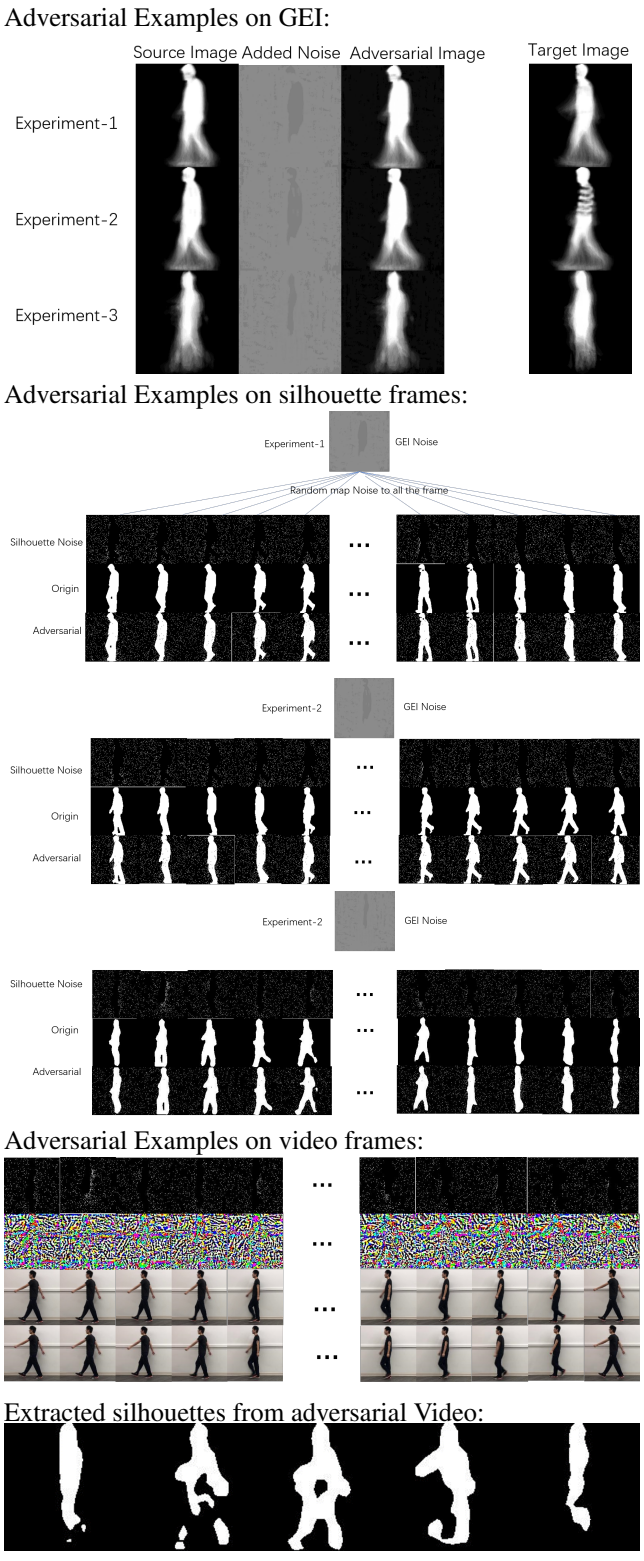

Figure 7: Result Image

Figure 8: Evaluation

$experiment$ l_2norm_{silh}	$confidence_{origin}$ l_2norm_{video}	$confidence_{target}$	l_2norm_{GEI}
1	35.2%	36.4%	0.986
0.0013	-		
2	43.3%	44.9%	0.984
0.0014	-		
3	47.3 %	47.9 %	0.985
0.0012	18.8		

[4] Alotaibi, Munif, and Ausif Mahmood. *"Improved gait recognition based on specialized deep convolutional neural network."* Computer Vision and Image Understanding 164 (2017): 103-110.

[5] Zhang, Cheng, et al. *"Siamese neural network based gait recognition for human identification."* 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[6] Feng, Yang, Yuncheng Li, and Jiebo Luo. *"Learning effective gait features using LSTM."* 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016.

[7] Akhtar, Naveed, and Ajmal Mian. "Threat of adversarial attacks on deep learning in computer vision: A survey." IEEE Access 6 (2018): 14410-14430.

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. InICLR, 2015.

[9] S. Baluja and I. Fischer. Adversarial transformation networks: Learning to generate adversarial examples. arXiv preprint arXiv:1703.09387, 2017.

[10] Jia, Ning et al. On view-invariant gait recognition: a feature selection solution. IET Biometrics 7 (2018): 287-295.

[11] Luo, Jian et al. Robust arbitrary view gait recognition based on parametric 3D human body reconstruction and virtual posture synthesis. Pattern Recognition 60 (2016): 361-377.

[12]V. U. Prabhu and J. Whaley. Vulnerability of deep learning- based gait biometric recognition to adversarial perturbations. In CVPR Workshop on The Bright and Dark Sides of Com- puter Vision: Challenges and Opportunities for Privacy and Security (CV-COPS 2017), 2017. 2

[13]Davrondzhon Gafurov, Einar Snekkenes, and Patrick Bours. Spoof attacks on gait authentication system. IEEE Transactions on Information Forensics and Security, 2(3), 2007. Special Issue on Human Detection and Recognition.

[14] Athalye A, Engstrom L, Ilyas A, et al. Synthesizing robust adversarial examples[J]. arXiv preprint arXiv:1707.07397, 2017.