Vulnerability of deep learning-based gait biometric recognition to adversarial perturbations

Vinay Uday Prabhu
UnifyID
San Francisco, CA, 94107

John Whaley
UnifyID
San Francisco, CA, 94107

Abstract

In this paper, we would like to draw attention towards the vulnerability of the motion sensor-based gait biometric in deep learning-based implicit authentication solutions, when attacked with adversarial perturbations, obtained via the simple fast-gradient sign method. We also showcase the improvement expected by incorporating these synthetically-generated adversarial samples into the training data.

1. Introduction

In recent times, password entry-based user-authentication methods have increasingly drawn the ire of the security community [1], especially when it comes to its prevalence in the world of mobile telephony. Researchers [1] recently showcased that creating passwords on mobile devices not only takes significantly more time, but it is also more error prone, frustrating, and, worst of all, the created passwords were inherently *weaker*. One of the promising solutions that has emerged entails *implicit authentication* [2] of users based on behavioral patterns that are *sensed* without the active participation of the user. In this domain of implicit authentication, measurement of gait-cycle [3] signatures, mined using the on-phone Inertial Measurement Unit - MicroElectroMechanical Systems (IMU-MEMS) sensors, such as accelerometers and gyroscopes, has emerged as an extremely promising passive biometric [4, 5, 6]. As stated in [7, 5], gait patterns can not only be collected passively, at a distance, and unobtrusively (unlike iris, face, fingerprint, or palm veins), they are also extremely difficult to replicate due to their dynamic nature.

Inspired by the immense success that Deep Learning (DL) has enjoyed in recent times across disparate domains, such as speech recognition, visual object recognition, and object detection [8], researchers in the field of gait-based implicit authentication are increasingly embracing DL-based machine-learning solutions [4, 5, 6, 9], thus replacing the more traditional hand-crafted-feature-engineering-driven shallow machine-learning approaches [10]. Besides circumventing the oft-contentious process of hand-engineering the features, these DL-based approaches are also more robust to noise [8], which bodes well for the implicit-authentication solutions that will be deployed on mainstream commercial hardware. As evinced in [4, 5], these classifiers have already attained extremely high accuracy ($\sim 96\%$), when trained under the k-class supervised classification framework (where k pertains to the number of individuals). While these impressive numbers give the impression that gait-based deep implicit authentication is ripe for immediate commercial implementation, we would like to draw the attention of the community towards a crucial shortcoming. In 2014, Szegedy *et al*. [11] discovered that, quite like *shallow* machine-learning models, the state-of-the-art deep neural networks were vulnerable to adversarial examples that can be synthetically generated by strategically introducing small perturbations that make the resultant adversarial input example only slightly different from correctly classified examples drawn from the data distribution, but at the same time resulting in a potentially controlled misclassification. To make things worse, a large plethora of models with disparate architectures, trained on different subsets of the training data, have been found to misclassify the same adversarial example, uncovering the presence of *fundamental blind spots* in our DL frameworks. After this discovery, several works have emerged ([12, 13]), addressing both means of *defence* against adversarial examples, as well as novel attacks. Recently, the `cleverhans` software library [13] was released. It provides standardized reference implementations of adversarial example-construction techniques and adversarial training, thereby facilitating rapid development of machine-learning models, robust to adversarial attacks, as well as providing standardized benchmarks of model performance in the adversarial setting explained above. In this paper, we focus on harnessing the simplest of all adversarial attack methods, *i.e.* the *fast gradient sign method* (FGSM) to attack the IDNet *deep convolutional neural network* (DCNN)-based gait classifier introduced in

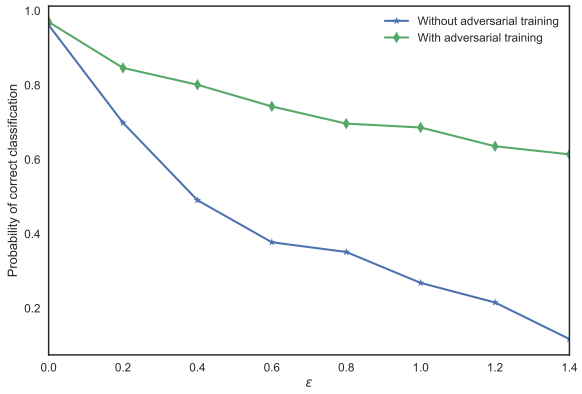

Figure 1. Variation in the probability of correct classification (37 classes) with and without adversarial training for varying ε

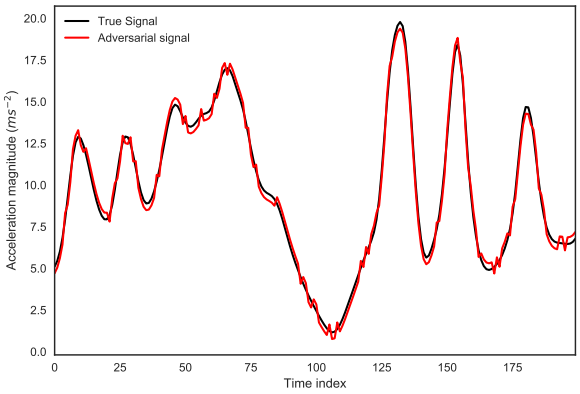

Figure 2. The true accelerometer amplitude signal and its adversarial counterpart for $\varepsilon = 0.4$

[4]. Our main contributions are as follows: 1: This is, to the best of our knowledge, the first paper that introduces deep adversarial attacks into this non-computer vision setting, specifically, the gait-driven implicit-authentication domain. In doing so, we hope to draw the attention of the community towards this crucial issue in the hope that further publications will incorporate adversarial training as a default part of their training pipelines. 2: One of the enduring images that is widely circulated in adversarial training literature is that of the *panda+nematode = gibbon* adversarial-attack example on GoogleNet in [14], which was instrumental in vividly showcasing the potency of the *blindspot*. In this paper, we do the same with accelerometric data to illustrate how a small and seemingly imperceptible perturbation to the original signal can cause the DCNN to make a completely wrong inference with high probability. 3: We empirically characterize the degradation of classification accuracy, when subjected to an FGSM attack, and also highlight the improvement in the same, upon introducing adversarial training. 4: Lastly, we have open-sourced the code at: https://github.com/vinayprabhu/Gait_FGSM

2. Methodology and Results

In this paper, we focus on the DCNN-based IDNet [4] framework, which entails harnessing low-pass-filtered tri-axial accelerometer and gyroscope readings (plus the sensor-specific *magnitude* signals), to, firstly, extract the *gait template*, of dimension 8×200, which is then used to train a DCNN in a supervised-classification setting. In the original paper, the model identified users in real time by using the DCNN as a deep-feature extractor and further training an outlier detector (*one-class support vector machine*-SVM), whose individual gait-wise outputs were finally combined into a Wald's probability-ratio-test-based framework. Here, we focus on the trained IDNet-DCNN and characterize its performance in the adversarial-training regime. To this end, we harness the FGSM introduced in [14], where the adversarial example, $\tilde{\mathbf{x}}$, for a given input sample, $\mathbf{x}$, is generated by

$$\tilde{\mathbf{x}} = \mathbf{x} + \varepsilon \operatorname{sign}\left(\nabla_{\mathbf{x}} \mathrm{J}\left(\theta, \mathbf{x}\right)\right),$$

where θ represents the parameter vector of the DCNN, $J\left(\theta, \mathbf{x}\right)$ is the cost function used to train the DCNN, and $\nabla_{\mathbf{x}}()$ is the gradient function.

As seen, this method is parametrized by ε, which controls the magnitude of the inflicted perturbations. Fig. 2 showcases the true and adversarial gait-cycle signals for the accelerometer magnitude signal (given by $a_{mag}(t) = \sqrt{a_x^2(t) + a_y^2(t) + a_z^2(t)}$) for $\varepsilon = 0.4$. Fig. 1 captures the drop in the probability of correct classification (37 classes) with increasing ε. First, we see that in the absence of any adversarial example, we were able to get about 96% accuracy on a 37 class classification problem, which is very close to what is claimed in [4]. However, with even *mild* perturbations ($\varepsilon = 0.4$), we see a sharp decrease of nearly 40% in accuracy. Fig. 1 also captures the effect of including the synthetically generated adversarial examples in this scenario. We see that, for $\varepsilon = 0.4$, we manage to achieve about 82% accuracy, which is a *vast improvement* of $\sim 25\%$.

3. Future Work

This brief paper is part of an ongoing research endeavor. We are currently currently extending this work to other adversarial-attack approaches, such as *Jacobian-based Saliency-Map Approach* (JSMA) and *Black-Box-Attack* (BBA) approach [15]. We are also investigating the effect of these attacks within the deep-feature-extraction+SVM approach of [4], and we are comparing other architectures, such as [6] and [5].

References

[1] W. Melicher, D. Kurilova, S. M. Segreti, P. Kalvani, R. Shay, B. Ur, L. Bauer, N. Christin, L. F. Cranor, and M. L. Mazurek, "Usability and security of text passwords on mobile devices," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 527–539, ACM, 2016. 1

[2] E. Shi, Y. Niu, M. Jakobsson, and R. Chow, "Implicit authentication through learning user behavior," in *International Conference on Information Security*, pp. 99–113, Springer, 2010. 1

[3] J. Perry, J. R. Davids, *et al.*, "Gait analysis: normal and pathological function.," *Journal of Pediatric Orthopaedics*, vol. 12, no. 6, p. 815, 1992. 1

[4] M. Gadaleta and M. Rossi, "Idnet: Smartphone-based gait recognition with convolutional neural networks," *arXiv preprint arXiv:1606.03238*, 2016. 1, 2

[5] Y. Zhao and S. Zhou, "Wearable device-based gait recognition using angle embedded gait dynamic images and a convolutional neural network," *Sensors*, vol. 17, no. 3, p. 478, 2017. 1, 2

[6] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," *arXiv preprint arXiv:1611.01942*, 2016. 1, 2

[7] S. Wang and J. Liu, *Biometrics on mobile phone*. INTECH Open Access Publisher, 2011. 1

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. 1

[9] N. Neverova, C. Wolf, G. Lacey, L. Fridman, D. Chandra, B. Barbello, and G. Taylor, "Learning human identity from motion patterns," *IEEE Access*, vol. 4, pp. 1810–1820, 2016. 1

[10] C. Nickel, C. Busch, S. Rangarajan, and M. Möbius, "Using hidden markov models for accelerometer-based biometric gait recognition," in *Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium on*, pp. 58–63, IEEE, 2011. 1

[11] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013. 1

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015. 1

[13] N. Papernot, I. Goodfellow, R. Sheatsley, R. Feinman, and P. McDaniel, "cleverhans v1.0.0: an adversarial machine learning library," *arXiv preprint arXiv:1610.00768*, 2016. 1

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

[15] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against deep learning systems using adversarial examples," *arXiv preprint arXiv:1602.02697*, 2016. 2