Adversarial Vulnerability in Doppler-based Human Activity Recognition

Zhaoyuan Yang, Yang Zhao, Weizhong Yan
GE Research
One Research Circle, Niskayuna, NY 12309
Emails:(zhaoyuan.yang, yang.zhao, and yan)@ge.com

***Abstract*—Human activity recognition (HAR) is an important task in many internet of things (IoT) applications. In recent years, significant efforts have been made towards achieving the highest possible recognition performance (accuracy and robustness) by using advanced machine learning techniques, including deep learning. However, to the best of our knowledge, the adversarial vulnerability of the Doppler sensor-based HAR systems has not been studied. In other domains such as computer vision, the vulnerability of deep learning algorithms to adversarial samples has attracted tremendous research interests in the past few years. In this work, we investigate the adversarial vulnerability of the Doppler-based human activity recognition system. Using a case study we demonstrate that the adversarial examples can significantly degrade the performance of the human activity recognition. Specifically, the basic iterative method (BIM) attack can reduce classification accuracy by as much as 85%. We also discuss different types of attacks, e.g., data poisoning attacks and potential strategies of protecting the Doppler-based HAR systems against adversarial attacks.**

***Index Terms*—Adversarial attack, Activity recognition, Time series classification**

I. Introduction

Human activity recognition (HAR) is an important task in many internet of things (IoT) applications, such as security, home care, and smart facilities. Accurate human activity recognition provides not only context information for prompt service decision making, but also long-term analytics for precision and personalized services. For example, in the pressure ulcer prevention and care application, we can monitor the activity of the patients and send notifications to change their positions on bed if they have not moved for certain amounts of time, in order to prevent pressure ulcer [1]. In the patient vital sign monitoring application, timely notifications to caregivers can save patients' lives [2].

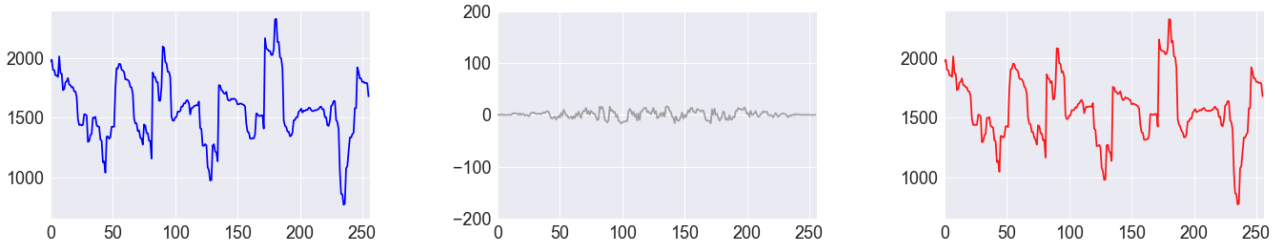
There are many different ways of sensing human activities. Wearable sensors can be attached to human body to monitor their activity and location. However, this type of systems requires user cooperation, and may cause discomfort to users. Cameras can monitor human activity in a non-cooperative way, but they have the privacy issue in many scenarios. As wireless devices become more and more pervasive recently, wireless sensor becomes a cost-effective and promising sensing modality. For example, [3] demonstrates that they can use radio frequency (RF) sensors to estimate human pose accurately through walls. [4] uses channel state information (CSI) from IEEE 802.11 radio chips to recognize human activity in a tag-free way. Work in [2] uses Doppler sensors and received signal strength (RSS) measurements from IEEE 802.15 radio chips to detect occupancy and classify human activity.

From a machine learning perspective, HAR is a multiclass classification problem where each class represents a human activity. Towards achieving the highest possible recognition performance (accuracy and robustness), many machine learning modeling techniques have been explored. As the deep neural networks (DNN)-based machine learning methods have achieved the state-of-the-art performance in the computer vision related applications, researchers have also used convolutional neural networks (CNN) and other deep learning methods in the human activity wireless sensing application [5], [6].

Recent studies have found the vulnerability of machine learning methods, e.g., DNN algorithms, to the adversarial examples [7]–[11]. These *adversarial examples* are small perturbations that cause DNN models to make false predictions with high confidence scores, as illustrated in Figure 1. After [7] first demonstrated the effect of the adversarial examples on the image classification problem, researchers have been proposing different new attack and defense methods for various natural language processing (NLP), reinforcement learning and other applications [12], [13]. More recently, [14] explores the adversarial vulnerability for the HAR using the smartphone data. However, to the best of our knowledge, the adversarial vulnerability of the deep learning models has not been studied for the non-intrusive Doppler sensor-based HAR systems [2]. We argue that adversarial attacks to the non-contact HAR systems can have significant economic and social impact, and thus the study of the adversarial vulnerability for the Doppler-based HAR systems is critical and urgent. For the Doppler sensor-based HAR system concerned in this paper, if the data acquisition system or the WiFi connection was attacked and adversarial examples were fed into the data during the inference stage, as shown in Figure 2, a machine learning model could misclassify the life-threatening "no vital sign" case as the "lying on bed" case, which would cause life-critical issues for a real patient monitoring system.

Aiming for exploring the adversarial vulnerability of HAR system, in this paper, we adapt the adversarial attack methods popularly used for the computer vision and natural language processing applications to a specific IoT time series classification [16] problem, i.e., Doppler-based human activity recognition [15]. We demonstrate that the adversarial exam-

978-1-7281-6926-2/20/$31.00 ©2020 IEEE



(a) Original Doppler time series from "turning on bed" activity.

(b) Crafted adversarial perturbation from C&W method.

(c) Adversarial examples from (a) + (b) that cause false prediction.

Fig. 1: Adversarial examples for Doppler-based human activity recognition: CNN model prediction is correct - "turning on bed" for (a) with 0.99 confidence score, but incorrect - "lying on bed" for (c) with 0.96 confidence score.

ples is a practical challenge for applying machine learning algorithms to the human activity Doppler sensing problem, as shown in Figure 1. Specifically, we apply three adversarial attack methods: fast gradient sign method (FGSM), basic iterative method (BIM) and C&W method, to the Doppler time series data. From our experiments, we demonstrate that the BIM attack method can reduce the recognition/classification accuracy by up to 85%. We also discuss the potential risks of data poisoning attack, and propose research topics and directions to investigate and improve adversarial robustness of human activity recognition.

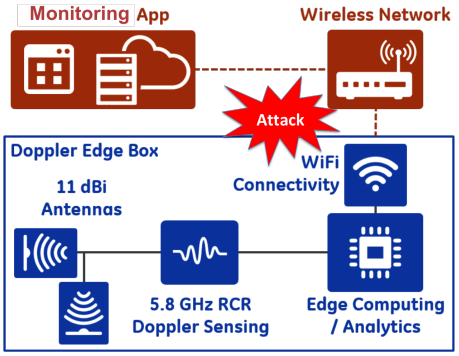


Fig. 2: Architecture of Doppler-based human activity monitoring system.

To the best of our knowledge, our paper is the first one in exploring the adversarial vulnerability of the non-intrusive Doppler-based HAR systems. Our initial results obtained in this paper reveal a potential security problem in HAR. We hope our work here can stimulate more research interests in this direction. The rest of this paper is organized as follows. Section II reviews related work. Attack strategies are discussed in Section III. In section IV we present our experiments and their results. Section V concludes the paper.

II. Related Work

For the human activity recognition problem, various wireless sensing techniques have been proposed and developed recently [5]. [4] uses channel state information (CSI) of wireless devices and proposed a sparse representation classification-based method to recognize lying, sitting, standing and walking activities. [2] fuses the received signal strength (RSS) measurements from a wireless network with the Doppler signal and uses the support vector machine (SVM) method to classify four activities. In [6], four wireless testbeds including WiFi, ultrasound, mmWave and visible light are used to extract environment-independent features from convonlutional neural networks (CNN) for device-free human activity recognition. In this work, we target on the adversarial robustness of the DNN-based algorithms on the Doppler sensor data.

There are various adversarial attack methods for a deep learning model. Two common adversarial attacks are the inference-time attack and the training-time attack. For the inference time-attack, an adversary adds small perturbations to the measurements so that the machine learning model produces incorrect predictions with high confidence [7], [8], [17], [18]. Later, [9] demonstrate a way of generating an universal adversarial perturbation for a trained classifier, [10] show a approach of generating one-pixel attack against a classifier. Most of attack are generated in the digital domains by manipulating the digits of an image, [11] demonstrate that this type of attack are also feasible in the physical world. For the training-time attack, training data are corrupted with carefully designed backdoors or triggers [19]. Through injecting the backdoor into the training data, the poisoned model will make false predictions [20]. In this paper, we demonstrate the inference-time attack using experimental data from a real-world human activity recognition application, and we discuss the training-time attack as a future research topic.

Most of the current adversarial attacks are demonstrated in the computer vision and natural language processing related applications [7], [8], [21]. For example, [8] uses the fast gradient sign method, and [21] uses the forward derivative method to craft adversarial examples. More recently, [22] uses the FGSM and BIM methods on time series classification to investigate the adversarial attacks on the vehicle sensor and food data classification problems. However, we are not aware of any research on the adversarial attacks on the Doppler based human activity recognition problem. We present

research effort on this direction and demonstrate the challenge from adversarial attacks on the state-of-the-art human activity recognition algorithms.

III. Adversarial attack strategies

In this section, we describe our strategies of attacking HAR systems. Specifically we describe different approaches of generating adversarial examples for inference-time attacks. The inference-time attack refers to an adversarial attack in the inference stage after a model is built and deployed. It includes targeted attack and non-targeted attack [12]. We do not limit our attack to a particular class, thus we generate adversarial examples with the more general non-targeted attack. In this work, we perform three type of inference-time attacks: FGSM [8], BIM [23] and C&W [17], which will be described in detail as follows. For description convenience, we denote the input time series as $\mathbf{x} = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$, the class label as $l \in \mathbb{Z}^+$, the perturbed data, i.e., the adversarial examples, as $\mathbf{x}' \in \mathbb{R}^n$, adversary targeted label as $l' \in \mathbb{Z}^+$, and the deep learning model as a function $f_\theta(\cdot)$, which maps input data $\mathbf{x}$ to a label l. Then, generation of adversarial examples can be formulated as following:

$$\min_{\mathbf{x}'} \|\mathbf{x}' - \mathbf{x}\|_p$$
$$s.t.\ f_\theta(\mathbf{x}') = l', f_\theta(\mathbf{x}) = l, l \neq l'$$

A. FGSM attack

Fast Gradient Sign Method (FGSM) was first proposed in [8]. Objective of FGSM method is to reduce the time complexity for generation of the adversarial examples. The adversarial examples from the **FGSM attack** can be generated as:

$$\mathbf{x}' = \mathbf{x} + \epsilon \text{sign}(\nabla_{\mathbf{x}} J_\theta(\mathbf{x}, l))$$

where $J_\theta(\cdot, \cdot)$ is a loss function, e.g., cross entropy of the model f_θ. FGSM only calculates the gradient once. After the gradient is obtained, it takes sign of the gradient and multiplied by a small perturbation ϵ, a hyper parameter controlling the perturbation magnitude, to generate the adversarial examples. Compared with other methods, FGSM is efficient in terms of computational complexity.

B. BIM attack

Basic Iterative Method (BIM) was first introduced in [23]. It extend FGSM method into a multi-step process. The adversarial examples from the **BIM attack** can be formulated as:

$$\mathbf{x}'_0 = \mathbf{x}, \quad \mathbf{x}'_{i+1} = Clip_{\mathbf{x},\eta}\Big\{\mathbf{x}'_i + \alpha \text{sign}\big(\nabla_x J_\theta(\mathbf{x}'_i, l)\big)\Big\}$$

where $Clip_{x,\eta}\{\mathbf{x}'\} = \min\Big\{\mathbf{x} + \eta, \max\{\mathbf{x} - \eta, \mathbf{x}'\}\Big\}$, and α controls the size of the update. Compared with FGSM, BIM attack needs multiple iterations to obtained adversarial examples. During each iteration, new $\mathbf{x}'$ will be clipped by η, which is a hyper parameter controlling the strength of the perturbation. To adapt from the image-based adversarial examples to the time series data, we remove the constrains of $\mathbf{x} \in [0, 255]$ from the formulation [23] .

C. C&W attack

Carlini & Wagner's adversarial attack (C&W) was first proposed in [17]. It formulate the problem as an alternative of a constrained optimization problem. The **C&W** adversarial attack can be formulated as:

$$\min_{\delta} D(\mathbf{x}, \mathbf{x} + \delta) + c \cdot g(\mathbf{x} + \delta)$$
$$s.t.\ \mathbf{x} + \delta \in [0, 1]^n$$

where $D(\cdot, \cdot)$ is the distance metric, $g(\cdot) \leq 0$ if and only if $f_\theta(\mathbf{x} + \delta) \neq l$. C&W attack also requires multiple iterations to craft adversarial examples, and it is effective for most of existing adversarial detecting defenses [12]. Note that all the above hyper parameters, e.g., ϵ in FGSM, can be chosen based on the range of the data. We will discuss this in more details in Section IV.

IV. Experiments and Results

A. Data description

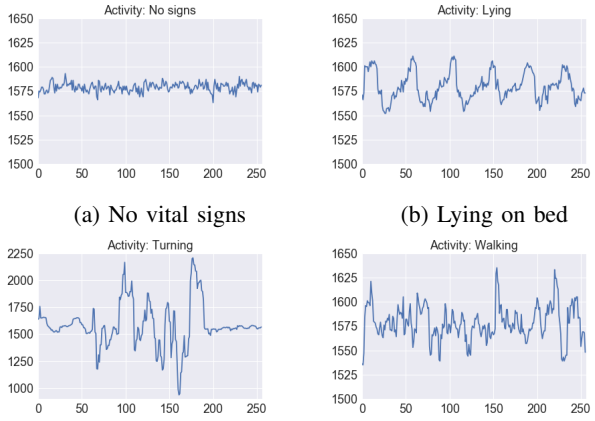
In this work, we present and discuss the adversarial robustness of the state-of-the-art machine learning techniques applied to the Doppler motion sensor-based human activity recognition problem [15]. The architecture and components of the Doppler-based human activity recognition system are shown in Figure 2. Two Doppler sensors are used to capture the vital signs and activities of the patients either on the hospital bed or in the room. The human sensing, edge computing and wireless connectivity components are all in a sensor box, as illustrated in the blue block in Figure 2 [15].

In [2], human subjects were recruited to perform over forty trials of four activities: (1) *walking in room*, (2) *lying on bed*, (3) *body turning on a bed*, and (4) *no vital signs (empty room)*. We use the data from the human activity experiments performed in [2] to evaluate the adversarial robustness of the CNN networks.

B. Data Processing

To capture the temporal features of the Doppler time series from different activities, we use a window of length 256 based on the sampling rate of the Doppler system, as mentioned in Section IV-C. All time series data from 42 trials are split into 21 training cases and 21 testing cases, with 1200 samples in each trial. Through the moving window approach [16], we obtained 19845 samples for training and 19845 samples for testing. Four samples from four activity states are shown in Figure 3. Note that the Doppler time series are from the analog-to-digital converter of the HAR system shown in Figure 2

From Figure 3, we see that the Doppler time series from different activity states clearly show different patterns. The time series from the "no vital signs" state have much smaller variation compared with the other activity states. The Doppler data from the "lying on bed" state shows a clear pattern of periodic changes due to the respiration motion of the person. The "turning on bed" activity has the largest magnitude change in time series, while the "walking in room" activity shows random variations.



(a) No vital signs (b) Lying on bed

(c) Turning on bed (d) Walking in room

Fig. 3: Doppler time series for four activity states.

C. Deep Learning Model Generation

For human activity recognition, CNN algorithms have achieved the state-of-the-art performance [5], [6]. CNN also provides an end-to-end solution, which does not require hand-crafted features. We build a CNN network to classify four activity states described in Section IV using the Doppler sensor data [2]. As shown in Figure 5, we construct two convolutional layers with the rectified linear unit (ReLU) as the activation function, and a fully connected layer with the softmax function before the output layer. The overall architecture of the CNN network is shown in Figure 5.

For training the deep learning model, we use the Doppler data collected in [2]. We use the overlapping moving window approach [16], to group time series into many blocks, as illustrated in Figure 4, to create large training and testing datasets. The window size should be chosen based on the sampling rate of the sensing system such that the model can learn the temporal features of the time series data.

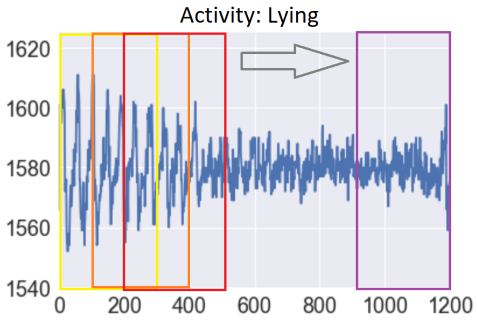


Fig. 4: Doppler time series with overlapping time windows (a window size of 256 is used to capture the temporal correlation of the Doppler time series).

D. Model training and performance measures

We implement the CNN networks mentioned above with Python deep learning package Keras [24]. Since dimension-

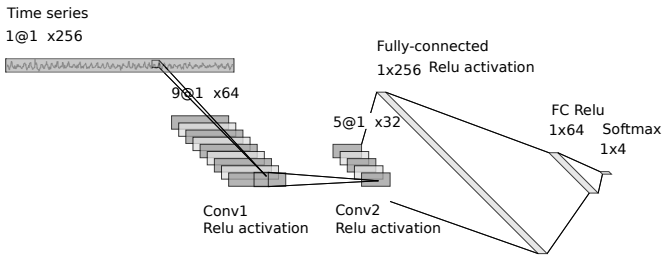


Fig. 5: Architecture of CNN networks for Doppler-based human activity recognition.

ality and size of the data is not large, we perform the model training on a local machine with Intel Core i7-7820HQ CPU.

For evaluating classification/recognition performance, we use confusion matrix to capture the misclassification counts. We calculate the window-based classification accuracy and case-based classification accuracy. To be specific, window-based classification performance evaluates the classifier's performance on all the sliding windows (total 19845 samples). Case-based classification performance evaluates the classifier's performance on all the test cases (total 21 test cases) by majority voting of the sliding windows in the same test case.

E. Attack-free Performance

We first evaluate the attack-free performance of the CNN networks as the baseline. Since we divide the Doppler time series into windows to generate enough samples for training, we perform two evaluations: the first evaluation is the window-based evaluation, where the CNN is evaluated on all 19845 testing samples; the second evaluation is the trial or case-based evaluation, where the evaluation is performed for 21 testing cases. The evaluation results, the confusion matrices, are shown in figure 6. We see that the CNN algorithm achieves high classification accuracy without any adversarial attacks: the average classification accuracy for the window-based evaluation is 92.2%, and the accuracy for the cased-based evaluation is 95.2%.

F. Adversarial Attack

Now we apply the adversarial attack methods discussed in Section III to generate adversarial examples and then feed them into the CNN networks as the testing data to evaluate the performance of CNN under different attacks.

First, we generate the adversarial examples for the white box attack using the IBM ART package [25]. We choose the hyper parameters, i.e., the bound parameters, based on the range of Doppler data. Table I shows the range of the testing data as well as the bounds of the adversarial examples from different activities. Note that since the time series from the"turning on bed" activity have much larger variations compared with those from other activities, we assign relatively larger bound when generating its adversarial examples. However, the bound is still less than 1.5% of the magnitude difference between its

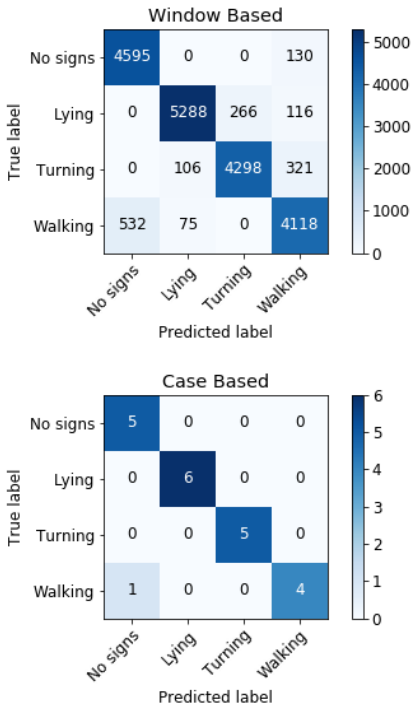


Fig. 6: Confusion matrices of attack-free CNN using window-based and case-based evaluation methods

maximum and minimum values, as shown in Table I. Thus, all the adversarial examples have small perturbations, as shown in Figure 7.

Activity	Min	Max	Diff	Bound ϵ
No Vital Signs	1561	1641	80	$\epsilon = 1.70$
Lying on Bed	1400	1964	564	$\epsilon = 1.70$
Turning on Bed	590	2469	1879	$\epsilon = 25.43$
Walking in Room	1518	1647	129	$\epsilon = 4.23$

TABLE I: Information of testing data and the designed bounds for adversarial noise ϵ.

Then, we feed generated adversarial examples into the CNN networks, and evaluate its performance. We list the window-based classification accuracy of CNN with three adversarial attacks in table II. Also listed is the accuracy of the attack-free CNN. We see that the adversarial examples from FGSM, BIM and C&W methods significantly reduce the performance of CNN. CNN with adversarial examples from FGSM only achieves 39% classification rate for the "no vital signs" class. The iterative attack methods BIM and C&W reduce the classification accuracy of CNN even more than FGSM. The average classification accuracy of CNN with adversarial examples from the BIM method is 11.28% for the "No Vital Signs" class, a 85.9% accuracy reduction from the attack-free CNN algorithm. In the meantime, we also show the standard deviation σ of the attack-free signals as well as adversarial-attacked signals in the table II. According to the table, the raw attack-free signals and the adversarial-attacked signals have similar values of standard deviation; however, their classification results are quite different from each other.

Finally, Figure 7 shows the comparison of normal samples and the adversarial samples from the C&W method as one example. We see that the adversarial examples have no obvious human-observable differences compared to the original signals. However, the adversarial examples cause the human activity recognition deep learning model to misclassify activity. For example, small perturbations to the "no vital sign" data make the CNN algorithm classify the data as "lying on bed", which could cause life-critical issues for a real patient monitoring system.

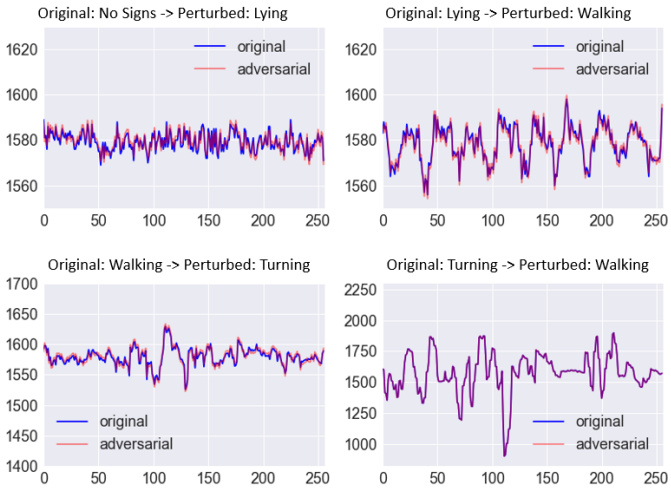


Fig. 7: Adversarial examples that cause false predictions.

G. Future Topics

Recent studies have investigated the time series inference-time attacks for different deep learning models, such as recurrent neural networks (RNN) [21] and CNN [22]. In this paper, we demonstrate the white box adversarial attack on the CNN algorithm. In the future, we plan to investigate another type of attack method, the black box attack. That is, we can attack a deep learning model without knowing its architecture and model parameters. Another interesting research topic is to investigate the transfer ability of various adversarial attack methods. While certain adversarial examples can be used to compromise CNNs or RNNs with different architectures, they may not be able to attack the conventional feature-based machine learning methods. Finally, the training-time attack for long short-term memory (LSTM) has been studied in [26], [27]. It is also an interesting and important topic to investigate the data poisoning attack for the human activity recognition problem.

Another research topic we plan to investigate is to add defense mechanism to the trained classifier. There are multiple research works related to defense of adversarial attacks in computer vision domains. [8] proposed an idea of using generated adversarial examples for adversarial training. [28]

Activity	Attack-free	FGSM	BIM	CW	σ of Attack-free	σ of FGSM	σ of BIM	σ of CW
No Vital Signs	97.25%	39.17%	**11.28%**	14.98%	4.76	4.85	4.84	4.71
Lying on Bed	93.26%	70.83%	**68.71%**	68.78%	42.57	42.64	42.63	42.58
Turning on Bed	90.96%	69.59%	55.98%	**41.82%**	166.42	171.83	170.47	165.78
Walking in Room	87.15%	43.81%	**31.70%**	43.07%	12.56	12.92	12.90	12.46

TABLE II: Class-wise accuracy of attack-free signal, FGSM, BIM and CW (left) and standard deviation of the corresponding attack-free and adversarial signals (right).

demonstrate an idea of using convex outer approximation to provide defense to the adversarial examples. Later, [29] proposed a min-max neural network training formulation to provide robustness to the trained models. [30], [31] demonstrate an approach of building robust and interpretable deep neural network model with k-nearest neighbor. Most of the research works are demonstrated using image data; therefore, we plan to extend the work into the time series domain as well as investigating the difference between defense in time series classification models and image classification models.

V. Conclusion

Human activity recognition is an important task in numerous IoT applications. Many advanced machine learning techniques including deep learning has been explored in recent years aiming for achieving the highest possible recognition performance. However, in the community of HAR, adversarial vulnerability of deep learning models has not been well recognized and studied, while such vulnerability has attracted tremendous research attention in other domains, e.g., computer vision. In this paper we explore the adversarial vulnerability of HAR systems by leveraging attack sampling techniques from computer vision. Using the Doppler-based HAR system as a case study, we demonstrate that the adversarial examples can be generated and are effective in attacking the deep learning models of the HAR system. More specifically, we build a CNN-based HAR system, and apply three adversarial attack methods on the Doppler time series data. Our experimental results show that such generated adversarial examples can significantly reduce the classification accuracy of the HAR system by as much as 85% from its originally designed and attack-free classification accuracy. To the best of our knowledge, our work in this paper is the first one addressing adversarial vulnerability of HAR systems. We hope that our initial work here would stimulate more research interests in the community of HAR.

Acknowledgment

This work was partially supported by the pressure ulcer prevention and care project funded by the Department of Veterans Affairs under Grant No VA118-13-C-0049.

References

[1] M. Chang, T. Yu, J. Luo, K. Duan, P. Tu, Y. Zhao, N. Nagraj, V. Rajiv, M. Priebe, E. A. Wood, and M. Stachura, "Multimodal sensor system for pressure ulcer wound assessment and care," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 3, pp. 1186–1196, March 2018.

[2] Y. Zhao, T. Yu, and J. Ashe, "Poster: Non-invasive human activity monitoring using a low-cost doppler sensor and an RF link," in *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '15. New York, NY, USA: ACM, 2015, pp. 397–398.

[3] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] B. Wei, W. Hu, M. Yang, and C. T. Chou, "Radio-based device-free activity recognition with radio frequency interference," in *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*, ser. IPSN '15. New York, NY, USA: ACM, 2015, pp. 154–165. [Online]. Available: http://doi.acm.org/10.1145/2737095.2737117

[5] Y. Ma, G. Zhou, and S. Wang, "Wifi sensing with channel state information: A survey," *ACM Comput. Surv.*, vol. 52, no. 3, pp. 46:1–46:36, Jun. 2019. [Online]. Available: http://doi.acm.org/10.1145/3310194

[6] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, W. Xu, and L. Su, "Towards environment independent device free human activity recognition," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '18. New York, NY, USA: ACM, 2018, pp. 289–304. [Online]. Available: http://doi.acm.org/10.1145/3241539.3241548

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *CoRR*, vol. abs/1312.6199, 2013. [Online]. Available: http://arxiv.org/abs/1312.6199

[8] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[9] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773.

[10] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[11] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[12] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–20, 2019.

[13] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," *arXiv preprint arXiv:1702.02284*, 2017.

[14] J. Z. Kolter and E. Wong, "Provable defenses against adversarial examples via the convex outer adversarial polytope," *CoRR*, vol. abs/1711.00851, 2017. [Online]. Available: http://arxiv.org/abs/1711.00851

[15] Y. Zhao, J. Ashe, D. Toledano, B. Good, L. Zhang, and A. McCann, "Occupancy and activity monitoring with doppler sensing and edge analytics: Demo abstract," in *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*, ser. SenSys '16. New York, NY, USA: ACM, 2016, pp. 322–323.

[16] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 international joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 1578–1585.

[17] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 39–57.
[18] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.
[19] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," 2017.
[20] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *CoRR*, vol. abs/1708.06733, 2017. [Online]. Available: http://arxiv.org/abs/1708.06733
[21] N. Papernot, P. McDaniel, A. Swami, and R. Harang, "Crafting adversarial input sequences for recurrent neural networks," in *MILCOM 2016-2016 IEEE Military Communications Conference*. IEEE, 2016, pp. 49–54.
[22] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. Muller, "Adversarial attacks on deep neural networks for time series classification," *arXiv preprint arXiv:1903.07054*, 2019.
[23] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *CoRR*, vol. abs/1607.02533, 2016. [Online]. Available: http://arxiv.org/abs/1607.02533
[24] K. Team. (2019) Keras: Deep learning for humans. [Online]. Available: https://github.com/fchollet/keras
[25] M.-I. Nicolae, M. Sinn, M. N. Tran, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. M. Molloy *et al.*, "Adversarial robustness toolbox v0. 4.0," *arXiv preprint arXiv:1807.01069*, 2018.
[26] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," *arXiv preprint arXiv:1807.00459*, 2018.
[27] J. Dai, C. Chen, and Y. Guo, "A backdoor attack against lstm-based text classification systems," *arXiv preprint arXiv:1905.12457*, 2019.
[28] E. Wong and J. Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," *arXiv preprint arXiv:1711.00851*, 2017.
[29] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.
[30] N. Papernot and P. McDaniel, "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning," *arXiv preprint arXiv:1803.04765*, 2018.
[31] N. Virani, N. Iyer, and Z. Yang, "Justification-based reliability in machine learning," *arXiv preprint arXiv:1911.07391*, 2019.