Adversarial Attacks on Deep Learning-based Floor Classification and Indoor Localization

Mohini Patil, Xuyu Wang
Department of Computer Science, California State University, Sacramento
Sacramento, CA, USA
mpatil@csus.edu,xuyu.wang@csus.edu

Xiangyu Wang, Shiwen Mao
Department of Electrical and Computer Engineering, Auburn University
Auburn, AL, USA
xzw0042@tigermail.auburn.edu,smao@ieee.org

ABSTRACT

With the great advances in location-based services (LBS), Wi-Fi localization has attracted great interest due to its ubiquitous availability in indoor environments. Deep neural network (DNN) is a powerful method to achieve high localization performance using Wi-Fi signals. However, DNN models are shown vulnerable to adversarial examples generated by introducing a subtle perturbation. In this paper, we propose adversarial deep learning for indoor localization system using Wi-Fi received signal strength indicator (RSSI). In particular, we study the impact of adversarial attacks on floor classification and location prediction with Wi-Fi RSSI. Three white-box attacks methods are examined, including fast gradient sign attack (FGSM), projected gradient descent (PGD), and momentum iterative method (MIM). We validate the performance of DNN-based floor classification and location prediction using a public dataset and show that the DNN models are highly vulnerable to the three white-box adversarial attacks.

CCS CONCEPTS

• Human-centered computing → Ubiquitous and mobile computing systems and tools; **• Computing methodologies → Machine learning**; **• Security and privacy → Mobile and wireless security**.

KEYWORDS

Adversarial Attacks, Deep Learning, Floor Classification, Indoor Localization, Received Signal Strength Indicator (RSSI), Wi-Fi

ACM Reference Format:
Mohini Patil, Xuyu Wang and Xiangyu Wang, Shiwen Mao. 2021. Adversarial Attacks on Deep Learning-based Floor Classification and Indoor Localization. In *3rd ACM Workshop on Wireless Security and Machine Learning (WiseML '21), June 28–July 2, 2021, Abu Dhabi, United Arab Emirates.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3468218.3469052

WiseML '21, June 28–July 2, 2021, Abu Dhabi, United Arab Emirates

ACM ISBN 978-1-4503-8561-9/21/06…$15.00
https://doi.org/10.1145/3468218.3469052

1 INTRODUCTION

With the rapid growth of mobile wireless systems, there has been great interest in location-based service (LBS). Wi-Fi based indoor localization has become a mainstreaming technique for LBS because of its ubiquitous availability. Wi-Fi fingerprinting based methods can estimate an unknown location by comparing the stored fingerprints with the newly received fingerprints and finding the best match. Existing Wi-Fi fingerprinting-based indoor localization systems mainly utilize two types of wireless signals, i.e., Received Signal Strength Indicator (RSSI) and Channel State Information (CSI). RADAR is the first RSSI-based scheme with a deterministic location estimation method [1]. Horus improves the localization accuracy by using a probabilistic, K-nearest-neighbor (KNN) approach [2]. RSSI is a coarse-grained representation of the Wi-Fi channel, which is usually significantly affected by the complex indoor propagation environment. Compared with RSSI, CSI is more stable and effective to capture the multipath effect. FIFS utilizes the weighted average of CSI amplitudes over multiple antennas for location estimation [3], while SpotFi achieves centimeter-level localization by estimating angle of arrival(AoA) using CSI [4].

Deep learning also greatly benefits indoor localization systems. DeepFi is the first work to apply a deep autoencoder to extract location features from CSI amplitudes as fingerprints [5]. Further, PhaseFi uses calibrated phase values to train a deep autoencoder [6], while BiLoc exploits bimodal CSI data [7]. Moreover, deep convolutional neural network (DCNN) is introduced as a classifier to simplify the fingerprinting based localization system. CiFi is the first work that utilizes DCNN with CSI data [8]. Unlike previous works, only one group of weights is required in CiFi for localization. Similarly, ConFi also uses the DCNN classifier with images generated from CSI amplitude data [9]. The recent work ResLoc leverages a deep residual sharing learning model with bimodal CSI tensors for improved localization accuracy [10].

Although deep learning has been effectively on improving indoor localization accuracy, Szegedy et al. [11] found that the state-of-the-art deep neural networks (DNN) are likely to be misled by adversarial examples including unrecognizable perturbations to the human eye. To defend against and understand the adversarial example, Goodfellow et al. [12] also proposed the Fast Gradient Sign Method (FGSM) attack method, and the corresponding defending strategy where the robustness of the network can be effectively enhanced by the modified adversarial objective function. Several adversarial attack methods were proposed based on FGSM. For example, Projected Gradient Descent (PGD) is a multiple-step variant of FGSM, which leverages the local first order information about the network [13]. The Momentum Iterative Method (MIM) attack

strengthens the effectiveness of the FGSM attack by introducing the momentum term [14]. Adversarial attacks are being studied under various wireless communication systems (e.g., modulation recognition, end-to-end communication system, and wireless security) [15–22].

Motivated by the previous adversarial attack works, in this paper, we propose adversarial deep learning models for indoor localization and floor classification with Wi-Fi RSSI values. The idea is to use adversarial examples to test the models' indoor localization and floor classification performance by adding a subtle perturbation to Wi-Fi RSSI. We first present the system model, where floor recognition is modeled as a classification problem and location prediction is modeled as a regression problem in deep learning. We then present the problem formulation and introduce three white-box attack methods (i.e., FGSM, PGD, and MIM). We examine the performances of DNN-based floor classification and location prediction using a public dataset, and show that the DNN models are vulnerable to the three white-box attacks.

The main contributions of this paper are summarized as follows.

- To the best of our knowledge, this is the first work to study adversarial attacks on DNN and Wi-Fi RSSI-based floor classification and indoor localization.
- We model floor recognition as a classification problem and location prediction as a regression problem in deep learning.
- Using a public, large-scale indoor localization dataset that includes three buildings and 13 floors, our experimental study demonstrates that the three white-box attack methods greatly influence DNN models' performance on floor classification and localization .

The remainder of this paper is organized as follows. The system model is described in Section 2. We present the problem formulation and adversarial attacks in Section 3. Our experimental study is presented in Section 4. Section 5 summarizes this paper.

2 SYSTEM MODEL

Most existing Wi-Fi based localization works are fingerprinting based methods, where fingerprinting data (i.e., pairs of Wi-Fi RSSI and location) are stored in a database in the offline training stage, and the distance matching method is used to find the location in the online test. To reduce the data storage and improve the location estimation accuracy, fingerprinting based localization can be formulated as a classification problem (e.g., in a small area) or a regression problem (e.g., in a large area). Naturally, compared with traditional machine learning methods, DNN becomes a powerful method to solve such localization problem when a large dataset is available. However, Szegedy et al. [11] found that DNN models are vulnerable to the adversarial examples that are only slightly different from the original data. Specifically, adversarial examples can be leveraged to fool a well-trained DNN model, thus generating incorrect predictions. Consequently, DNN based localization systems are also vulnerable to adversarial attacks.

2.1 System Architecture

Fig. 1 shows the architecture of the proposed floor classification and indoor localization system. Similar to traditional DNN based indoor localization systems, the proposed system also includes an offline training stage and an online test stage. In the data preprocessing stage, the input data will be normalized to improve the performance of DNN models. In the offline stage, the training dataset is used to train two DNN models for floor classification and location prediction, respectively. In the online stage, adversarial perturbations generated by three white-box attacks will be introduced to the new Wi-Fi RSSI data. Then, we use the trained models to validate the performances of floor classification and location prediction. To the best of our knowledge, this is the first work to study the impact of adversarial attacks on a deep regression model in the wireless field, while most of the previous works mainly focus on classification problems (e.g., wireless modulation recognition).

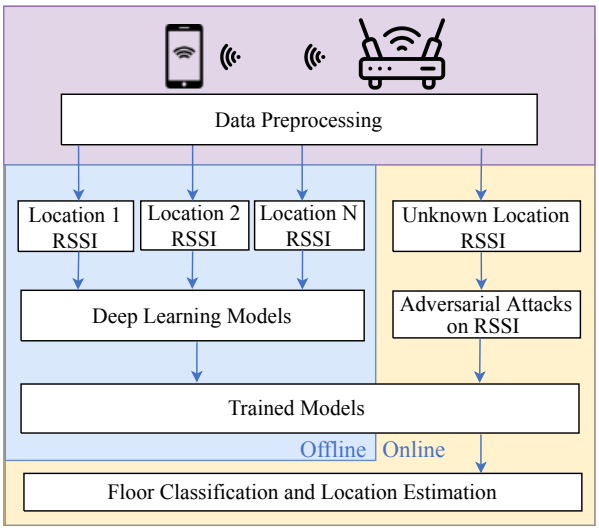


Figure 1: System architecture.

We use the public UJIIndoorLoc dataset (i.e., a multi-building and multi-floor database) [23] for Wi-Fi fingerprinting based indoor localization and floor classification. This database was collected from more than 20 users and 25 devices in three buildings (totally 13 floors) with 520 wireless access points (WAPs). The UJIIndoorLoc database has 19,937 reference records in the training and validation sets (e.g., used in the offline training stage) and 1,111 reference records in the test set (e.g., used in the online test stage) in Fig. 2. From every location, we can extract 520 RSSI values from all the WAPs, which can be used as input to the proposed DNN models for floor classification and location prediction. In the next section we will present the DNN models for floor classification and location prediction used in this work.

2.2 DNN Models for Floor Classification and Location Prediction

To study the effect of adversarial attacks on DNN based floor classification and indoor localization, two DNN models are adopted in the proposed system, which are shown in Fig. 3 and Fig. 4, respectively.

For floor classification, the building and floor attribute values are concatenated to format the output values. There are three unique building identification (ID) values in the range of [0, 1, 2], and five unique floor ID values in the range of [0, 1, 2, 3, 4]. We then

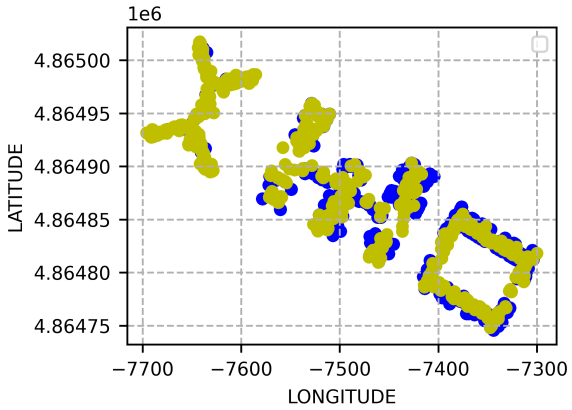


Figure 2: Campus map for the training and test datasets.

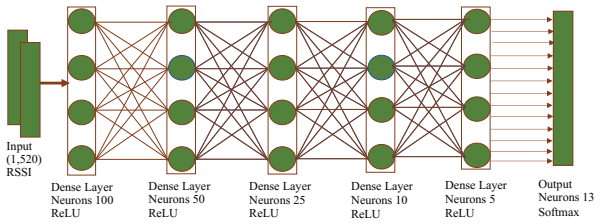


Figure 3: The DNN classification model.

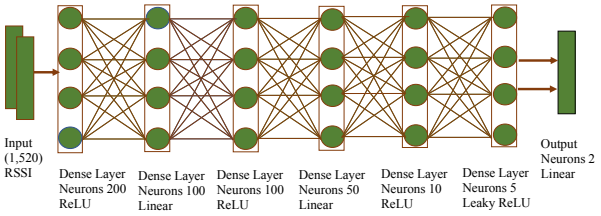


Figure 4: The DNN regression model.

create 13 corresponding unique IDs as floor labels, as shown in Table 1, where each label represents both building ID and floor ID (e.g., "12" indicates building ID "1" and floor ID "2"). Thus, our floor classification problem also includes building recognition. Fig. 3 presents the DNN classification model, which is used for floor classification, where 520 RSSI values and 13 labels are the input and output of DNN, respectively. The DNN classification model consists of five dense layers with 100, 50, 25, 10, and 5 neurons, respectively, where `ReLU` is used as the activation function. The last layer has 13 neurons using the `Softmax` function for classification. Cross-entropy is used as the loss function for training.

For location estimation, we use 520 RSSI values from the WAPs as input, and two values (i.e., longitude and latitude) are the output to train the DNN regression model in the offline stage, and then use the trained DNN model to predict location using 520 new RSSI samples. Fig. 4 shows the DNN regression model, where six dense layers are used with 200, 100, 100, 50, 10, and 5 neurons, respectively. `ReLu`, `linear`, and `Leaky ReLU` are used as activation functions.

Table 1: Labels for Floor Classification

Building ID	Floor ID	Labels
0	0, 1, 2, 3	{00, 01, 02, 03}
1	0, 1, 2, 3	{10, 11, 12, 13}
2	0, 1, 2, 3, 4	{20, 21, 22, 23, 24}

The last layer has two neurons with linear activation function, where the loss function for the DNN regression model is the mean squared error (MSE). In addition, we use Adam as the optimizer for both DNN models, with a batch size of 128.

3 PROBLEM FORMULATION AND ADVERSARIAL ATTACK MODELS

In this section, we provide a general formulation for floor classification and location prediction using Wi-Fi RSSI values. We will also introduce three white-box adversarial attack methods, including FGSM, PGD, and MIM, used to evaluate the resilience performance of the proposed system.

3.1 Problem Formulation

Let x denote the input data (i.e., 520 RSSI values), and y denote the output (e.g., 13 neurons for the floor classification problem or 2 neurons for the indoor localization regression problem). Further, we use f to denote the DNN model function, $\mathcal{L}$ as the loss function of the DNN model (i.e. cross-entropy for classification and MSE for regression), and θ as the weight parameters of the DNN model. For floor classification and location prediction problems, our objective is to minimize the loss function by finding the optimal weight parameters of the DNN model, which can be formulated as

$$\arg\min_{\theta} \mathcal{L}(f(x, \theta), y). \tag{1}$$

By training the DNN model to minimize the loss function, we can identify the optimal weight parameters θ^*, which will be used for floor classification and location prediction in the online stage with new collected RSSI values.

However, adversarial examples can mislead the DNN model by introducing a small perturbation to the new RSSI data. In fact, the objective of the adversary is to degrade the performance the DNN model by maximizing the following loss function.

$$\arg\max_{x_{adv}} \mathcal{L}(f(x_{adv}, \theta^*), y), \tag{2}$$

where x_{adv} is the adversarial example. The adversarial example x_{adv} can be generated as $x_{adv} = x + \eta$, where η is the perturbation. Given a trained DNN model f with parameter θ^*, generating an adversarial example x_{adv} can be formulated as a box-constrained optimization problem (e.g., L-BFGS attack needs to use a binary search to find the optimal parameter value) [11]. However, this could be time-consuming and impractical. Thus, in this paper, we focus on the one-step attack method (i.e., FGSM) and two iterative attack methods (PGD and MIM) as the follows.

3.2 Fast Gradient Sign Method

The FGSM attack method is to obtain a perturbation by calculating the gradient of the loss function with a given input [12]. The

perturbation η is generated by

$$\eta = \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(f(x, \theta^*), y)), \tag{3}$$

where ϵ is a hyper-parameter, which controls the magnitude of the perturbation. Since $\mathcal{L}$ is the loss function of the model, the perturbation η could be calculated by using the first derivative of $\mathcal{L}(f(x, \theta^*), y)$ through the backpropagation algorithm. The Fast Gradient Method (FGM) [24] is a generalization of FGSM, where the perturbation is given by

$$\eta = \epsilon \cdot \frac{\nabla_x \mathcal{L}(f(x, \theta^*), y)}{\|\nabla_x \mathcal{L}(f(x, \theta^*), y)\|_2}. \tag{4}$$

Using (4), the perturbation can be conveniently generated.

3.3 Projected Gradient Descent Attack

Based on the one-step FGM, the iterative version of FGM (i.e., the PGD attack) was proposed to improve the attack performance [13]. PGD can also help to enhance the robustness of DNN model against first-order attacks. By using iterative methods, the adversarial examples are created as follows.

$$x_0^{adv} = x, \tag{5}$$

$$x_{N+1}^{adv} = \text{Clip}_{x,\epsilon}\left\{ x_N^{adv} + \alpha \cdot \frac{\nabla_x \mathcal{L}(f(x_N^{adv}, \theta^*), y)}{\left\|\nabla_x \mathcal{L}(f(x_N^{adv}, \theta^*), y)\right\|_2} \right\}, \tag{6}$$

where α is a hyper-parameter in each iteration; when ϵ is given, it could be set to ϵ/N. The perturbation is always small and around the original input x in the L^p ball. Moreover, $\text{Clip}_{x,\epsilon}$ is used to project the perturbation back into the L^p ball. PGD has been recognized as a stronger adversarial attack method than the one-step FGM/FGSM.

3.4 Momentum Iterative Method

Instead of using the gradient in one iteration to update the perturbation, MIM uses the gradient of the previous iterations to help update the perturbation. To create adversarial examples using the MIM method, we first obtain the gradient in the $(N+1)$th iteration, which is given by

$$g_{(N+1)} = \mu \cdot g_N + \frac{\nabla_x \mathcal{L}(f(x_N^{adv}, \theta^*), y)}{\left\|\nabla_x \mathcal{L}(f(x_N^{adv}, \theta^*), y)\right\|_2}, \tag{7}$$

where g_N includes the gradients from previous $N-1$ iterations with a decay factor μ. The adversarial examples are computed by

$$x_{(N+1)}^{adv} = x_N^{adv} + \alpha \cdot \text{sign}(g_{(N+1)}), \tag{8}$$

where α could be set to ϵ/N when ϵ is provided.

4 EXPERIMENTAL STUDY

4.1 Experiment Configuration

Experiments have been performed using the public UJIndoorLoc dataset. This dataset consists of a training dataset with the 19,937 reference records and a test dataset with 1,111 reference records (that were collected four months later) in three buildings with totally 13 floors. Wi-Fi RSSI values were measured at Wi-Fi receivers (e.g., smartphones, laptops, and other devices) of more than 20 users with 25 devices, which are used as features in floor classification and indoor localization. In the training and validation datasets, there are totally 529 features for each sample, where the first 520 features are from Wi-Fi WAPs corresponding to the RSSI values. The normal range is from -104 dBm to 0 dBm. For example, -109 dBm means that there is no signal or an extremely weak signal. There are nine other features, including Latitude, Longitude, Floor ID, Building ID, Space ID, Relative Position, User ID, Phone ID, and Timestamp. In this paper, we consider location coordinates (i.e., latitude and longitude) in the location estimation model, and use floor ID and building ID in the floor classification model.

Data pre-processing is very important for the proposed DNN models to achieve satisfactory results. Scaling of the input features has been achieved with normalization. After implementing normalization there are few access points that were undetectable from all the locations. These access points could be removed to accelerate the training process. In addition, the training dataset with the 19,937 reference records for floor classification and location prediction is divided into two subsets: 70% for training and 30% for validation, respectively. Then, we use the test dataset with 1,111 reference records to verify the proposed methods.

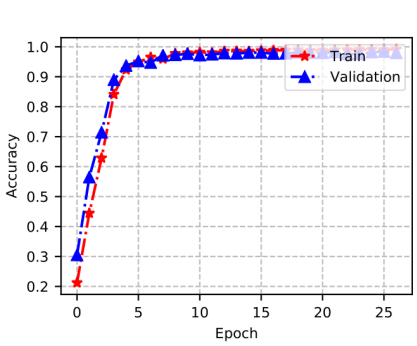
Three types of adversarial attacks (i.e., FGSM, PGD, and MIM) have been performed independently for floor classification as well as location prediction. All adversarial attacks have been performed in the validation and testing stages using different epsilon values. In addition, all the experiments are conducted using Python and the Tensorflow, Keras, and Scikit-Learn libraries for training both models. Further, Google Colab Pro is used as a cloud service to train these models.

In the following section, we will present the performance of DNN-based floor classification and location prediction, and evaluate their performances under three types of adversarial attacks.

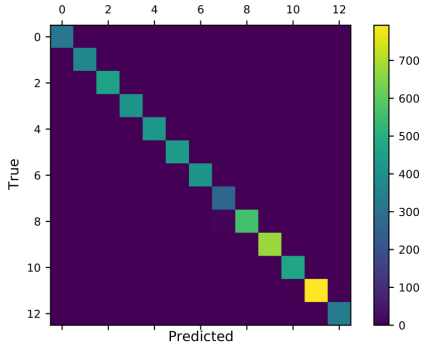
4.2 Results and Discussions

Fig. 5 presents the experiment results for floor classification. Specifically, Fig. 5(a) shows the accuracy over different epochs for training and validation of the floor classification model. The DNN classification model is trained over 30 epochs in the offline stage. The accuracy quickly increases with the increase of epochs. Moreover, the accuracy curves for training and validation both converge when the number of epochs is over 10. Fig. 5(b) shows the confusion matrix for the floor classification model using the validation data and Fig. 5(c) presents the confusion matrix using the test data. It is easy to see that the classification accuracy for the validation data is better than that for the test data, because the test data is collected at a different time (i.e., four months later).

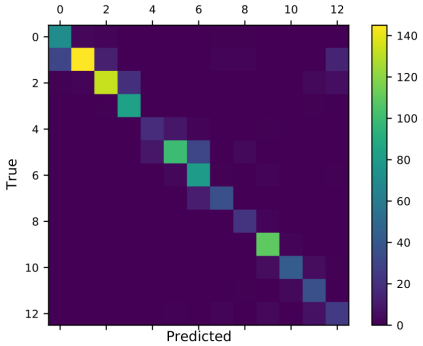
Fig. 5(d) shows the accuracy of floor classification over different ϵ values under FGSM attack. The adversarial examples are generated by adding a small perturbation under different ϵ values, which determine the strength of the noise added to the original input data. The perturbation is calculated for each ϵ value in the range of [0.001, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09]. In Fig. 5(d), the accuracy with clean data is denoted by "Validation" and "Test," while the accuracy with adversarial examples is expressed by "Validation (FGSM)" and "Test (FGSM)," respectively. We find that after adding the perturbation data, the accuracy over all epsilons becomes very poor. For example, when $\epsilon = 0.02$, the accuracy of validation data drops from 0.98 to 0.093 and the test data accuracy drops from



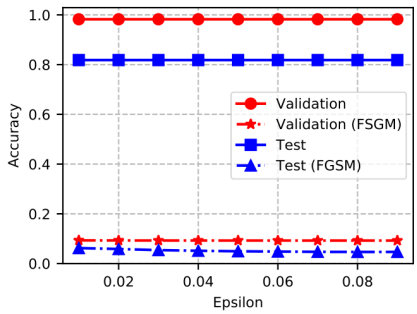
(a) Accuracy versus epochs for training and validation.



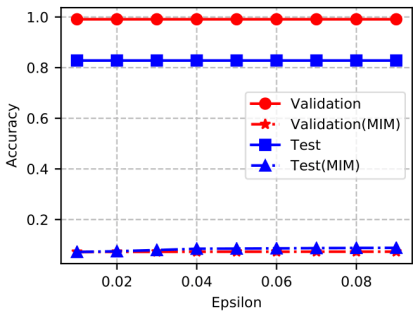
(b) Confusion matrix for the validation data.



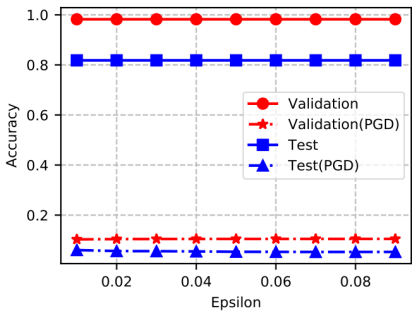
(c) Confusion matrix for the test data.



(d) Accuracy of floor classification versus epsilon under the FGSM attack.



(e) Accuracy of floor classification versus epsilon under the MIM attack.



(f) Accuracy of floor classification versus epsilon under the PGD attack.

Figure 5: Floor classification results with and without FGSM, MIM and PGD attacks.

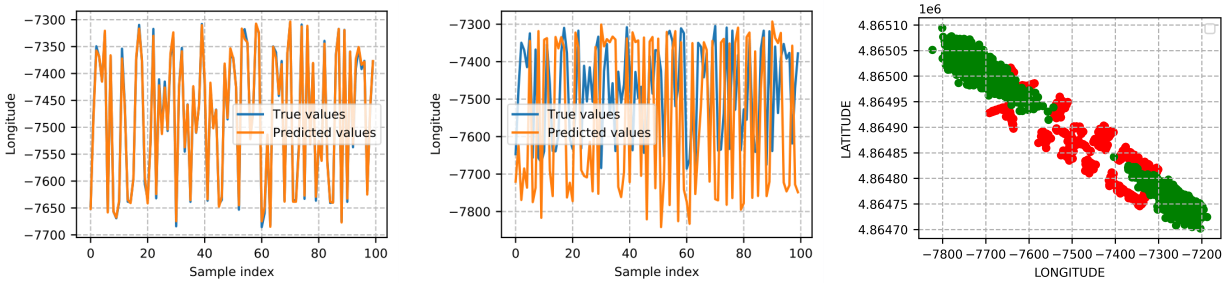
0.81 to 0.059. Fig. 5(e) shows the accuracy of the floor classification model under MIM attacks. When $\epsilon = 0.02$, the accuracy using validation data drops from 0.98 to 0.072 and the test data accuracy decreases to 0.067 from 0.81. Fig. 5(f) shows the accuracy under PGD attacks. when $\epsilon = 0.02$, the validation data accuracy drops from 0.98 to 0.072. Thus, we conclude that all three adversarial attacks can easily fool the floor classification model, and the impact is relatively independent to epsilon values.

Fig. 6 presents the experiment results of location prediction. Specifically, Fig. 6(a) shows the longitude predictions without adversarial attacks on the validation data using 100 samples. We can see that the predicated longitude values are very close to the ground truth. Fig. 6(b) presents the true longitude values and predicted values under the FGSM attack when $\epsilon = 0.05$. We can see that the predicted values do not match the ground truth anymore due to the small perturbation introduced to the validation data with the same 100 samples. Fig. 6(c) illustrates the campus map in latitude and longitude for predicted values using clear validation data (i.e., red color) and predicted values using the perturbed validation data under FGSM when $\epsilon = 0.05$ (i.e., green color). We can conclude that adversarial attacks (e.g., FGSM) can greatly degrade the location prediction performance.

We also apply these three types of attacks to the localization estimation model. Fig. 6(d) plots the localization errors under FGSM attacks for different ϵ values. The mean localization error with clear validation data is 7.62 meters and that for clean test data is 11.83 meters. Under FGSM attacks, with increased ϵ, the localization error also increases. Specifically, with $\epsilon = 0.02$, the localization error drastically increases from 7.62 meters to 168.72 meters with the validation data, and to 170.41 meters with the test data. Fig. 6(e) and Fig. 6(f) present the localization errors under MIM and PGD attacks, respectively, with different ϵ values. The localization error increases with ϵ under both attacks. Compared with FGSM, MIM and PGD generate much larger errors. For example, when $\epsilon = 0.02$, the errors are 208.63 meters with the validation data and 204.63 meters with the test data under MIM attacks. Under PGD attacks when $\epsilon = 0.02$, the error increases to 211.27 meters with the validation data and 207.70 meters with the test data.

5 CONCLUSIONS

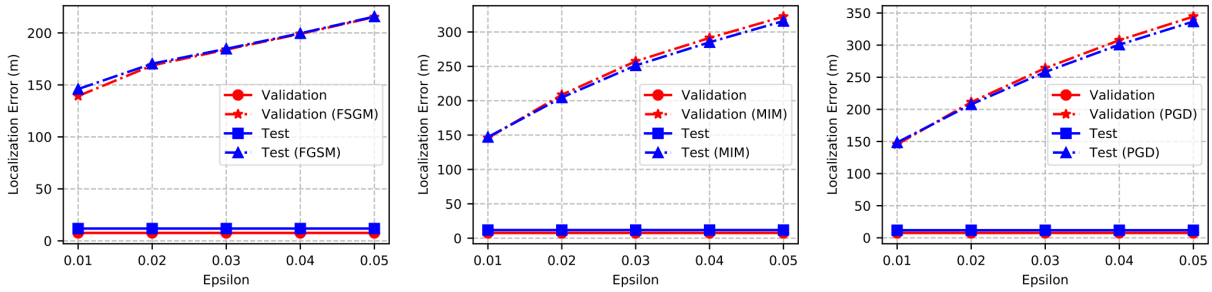
In this paper, we studied adversarial attacks on DNN-based floor classification and location estimation using Wi-Fi RSSI. We introduced the system model including floor recognition as a classification problem and location prediction as a regression problem in deep learning. Then, three white-box attack methods (i.e., FGSM, PGD, and MIM) were discussed. Through experiments with a public and large-scale indoor localization dataset, our results demonstrated that the performances of floor classification and indoor localization are highly susceptible to adversarial attacks.



(a) Longitude prediction without adversarial attacks.

(b) Longitude prediction under FGSM attacks when $\epsilon = 0.05$.

(c) Predicated campus map under FGSM attacks when $\epsilon = 0.05$.



(d) Localization errors over different ϵ values under FSGM attacks.

(e) Localization errors over different ϵ values under MIM attacks.

(f) Localization errors over different ϵ values under PGD attacks.

Figure 6: Localization results with and without FSGM, MIM, and PGD attacks.

ACKNOWLEDGMENTS

This work is supported in part by the NSF under Grants ECCS-1923163, CNS-1822055, CNS-2105416, and the Wireless Engineering Research and Education Center at Auburn University, Auburn, AL, USA.

REFERENCES

[1] P. Bahl and V. N. Padmanabhan, "Radar: An in-building RF-based user location and tracking system," in *Proc. IEEE INFOCOM'00*, Tel Aviv, Israel, Mar. 2000, pp. 775–784.

[2] M. Youssef and A. Agrawala, "The Horus WLAN location determination system," in *Proc. ACM MobiSys'05*, Seattle, WA, June 2005, pp. 205–218.

[3] J. Xiao, K. Wu., Y. Yi, and L. Ni, "FIFS: Fine-grained indoor fingerprinting system," in *Proc. IEEE ICCCN'12*, Munich, Germany, Aug. 2012, pp. 1–7.

[4] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "SpotFi: Decimeter level localization using WiFi," in *Proc. ACM SIGCOMM'15*, London, UK, Aug. 2015, pp. 269–282.

[5] X. Wang, L. Gao, S. Mao, and S. Pandey, "CSI-based fingerprinting for indoor localization: A deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 763–776, 2016.

[6] X. Wang, L. Gao, and S. Mao, "CSI phase fingerprinting for indoor localization with a deep learning approach," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 1113–1123, 2016.

[7] X. Wang, L. Gao, S. Mao, and S. Pandey, "BiLoc: Bi-modal deep learning for indoor localization with commodity 5GHz WiFi," *IEEE Access*, vol. 5, pp. 4209–4220, Mar. 2017.

[8] X. Wang, X. Wang, and S. Mao, "Deep convolutional neural networks for indoor localization with CSI images," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 316–327, 2018.

[9] H. Chen, Y. Zhang, W. Li, X. Tao, and P. Zhang, "Confi: Convolutional neural networks based indoor wi-fi localization using channel state information," *IEEE Access*, vol. 5, pp. 18 066–180 747, Sept. 2017.

[10] X. Wang, X. Wang, and S. Mao, "Indoor fingerprinting with bimodal CSI tensors: A deep residual sharing learning approach," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4498–4513, 2021.

[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*, Dec. 2013.

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

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, June 2017.

[14] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. IEEE CVPR'18*, Salt Lake City, UT, June 2018, pp. 9185–9193.

[15] S. Kokalj-Filipovic, R. Miller, and J. Morman, "Targeted adversarial examples against RF deep classifiers," in *Proceedings of the ACM Workshop on Wireless Security and Machine Learning*, 2019, pp. 6–11.

[16] Y. Lin, H. Zhao, Y. Tu, S. Mao, and Z. Dou, "Threats of adversarial attacks in DNN-based modulation recognition," in *Proc. IEEE INFOCOM'19*, Toronto, Canada, July 2020, pp. 1–10.

[17] M. Patel, X. Wang, and S. Mao, "Data augmentation with conditional GAN for automatic modulation classification," in *Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning*, 2020, pp. 31–36.

[18] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Channel-aware adversarial attacks against deep learning-based wireless signal classifiers," *arXiv preprint arXiv:2005.05321*, 2020.

[19] D. Adesina, C.-C. Hsieh, Y. E. Sagduyu, and L. Qian, "Adversarial machine learning in wireless communications using RF data: A review," *arXiv preprint arXiv:2012.14392*, 2020.

[20] M. Sadeghi and E. G. Larsson, "Physical adversarial attacks against end-to-end autoencoder communication systems," *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 847–850, 2019.

[21] Y. Sagduyu, Y. Shi, and T. Erpek, "Adversarial deep learning for over-the-air spectrum poisoning attacks," *IEEE Trans. Mobile Comput.*, vol. 20, no. 2, pp. 306–319, 2020.

[22] Y. E. Sagduyu, Y. Shi, T. Erpek, W. Headley, B. Flowers, G. Stantchev, and Z. Lu, "When wireless security meets machine learning: Motivation, challenges, and research directions," *arXiv preprint arXiv:2001.08883*, 2020.

[23] J. Torres-Sospedra, R. Montoliu, A. Martínez-Usó, J. P. Avariento, T. J. Arnau, M. Benedito-Bordonau, and J. Huerta, "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems," in *Proc. IEEE IPIN'14*, 2014, pp. 261–270.

[24] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial training methods for semi-supervised text classification," *arXiv preprint arXiv:1605.07725*, 2016.