Black-box Adversarial ML Attack on Modulation Classification

Muhammad Usama
muhammad.usama@itu.edu.pk
Information Technology University, Punjab, Pakistan

Junaid Qadir
junaid.qadir@itu.edu.pk
Information Technology University, Punjab, Pakistan

Ala Al-Fuqaha
aalfuqaha@hbku.edu.qa
Hamad Bin Khalifa University, Qatar

ABSTRACT

Recently, many deep neural network (DNN) based modulation classification schemes have been proposed in the literature. We have evaluated the robustness of two famous such modulation classifiers (based on the techniques of convolutional neural networks and long short term memory) against adversarial machine learning attacks in black-box settings. We have used Carlini & Wagner (C-W) attack for performing the adversarial attack. To the best of our knowledge, the robustness of these modulation classifiers have not been evaluated through C-W attack before. Our results clearly indicate that state-of-art deep machine learning based modulation classifiers are not robust against adversarial attacks.

KEYWORDS

Adversarial ML, Modulation Classification, Deep Learning

ACM Reference Format:
Muhammad Usama, Junaid Qadir, and Ala Al-Fuqaha. 2019. Black-box Adversarial ML Attack on Modulation Classification. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Machine learning (ML) especially deep ML schemes have beaten human-level performance in many computer vision, language, and speech processing tasks which were considered impossible a decade ago. This success of ML schemes has inspired the ideas of self-driving networks [2] and knowledge defined networking [5] where ML schemes are profoundly utilized to ensure automation and control of networking tasks such as dynamic resource allocation, modulation classification, network traffic classification, etc.

Despite the success of ML in different modern communication and data networking applications, there are some pitfalls in the fundamental assumptions of ML schemes which can be exploited by the adversaries to craft adversarial examples in order to compromise the ML-based system. An *adversarial example* is defined as an input to the ML model specially crafted by an adversary by adding a small imperceptible perturbation to the input sample to compromise the performance of the ML model. Mathematically, an adversarial example x^* can be formed by adding a typically-imperceptible perturbation δ to the legitimate test example x of the deployed trained classifier $f(.)$. The perturbation δ is computed by approximating the following nonlinear optimization problem provided in equation 1 where t is the targeted class in case of a targeted attack or any other wrong class is the case of untargeted attack.

$$x^* = x + \arg\min_{\eta_x}\{\|\eta\| : f(x+\eta) = t\} \quad (1)$$

Adversarial examples are possible because of two major faulty assumptions in ML schemes. *Firstly*, the underlying data distribution experienced during the training phase of the ML model will also be encountered in the testing phase. This data stationarity is not valid for most of the real world cases and the void created by following this assumption is exploited by the adversary for crafting the adversarial examples. *Secondly*, most of the ML schemes are based on the empirical risk minimization (ERM), which is an approximation of the actual unknown probability distribution. The ERM has an associated error with it which can be exploited by the adversary to make an adversarial example.

Adversarial attacks can be classified broadly into *white-box* and *black-box* attacks based on the knowledge of the adversary about the deployed ML model. In a *white-box attack*, it is assumed that adversary has complete knowledge (hyperparameters, test data, etc.) of the deployed model whereas in a *black-box attack* no such knowledge is assumed and it is assumed that the adversary can only act as a standard user and query the system for a response.

In this paper, we have taken modulation classification (which is an important component of modern communication and data networks) as a proxy of functional areas of cognitive self-driving networks. We have performed a black-box adversarial attack on DNN-based modulation classification to highlight the brittleness of ML schemes utilized in cognitive self-driving networks.

2 RELATED WORK

There does not exist much literature on adversarial attacks on modulation classification. Recently, Sadeghi et al. [7] used a variant of fast gradient sign method (FGSM) attack [3] on modulation classification on CNN-based modulation classification to highlight the threat of the adversarial examples. FGSM is an adversarial sample crafting algorithm where the adversarial perturbation is calculated by taking a gradient step in the direction of the sign of the gradient of test example. Kokalj et al. [4] also crafted the adversarial examples for modulation classification by using the FGSM perturbation generation algorithm. Most of the available results on the application of the adversarial attacks are reported by using the FGSM attack.

A shortcoming with the FGSM attack is its lack of optimality in adversarial perturbation generation as FGSM was designed to quickly craft adversarial examples irrespective of the optimality

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

and the size of the perturbation in the test example. To overcome the lack of optimality and to highlight that optimal adversarial example for modulation classification can be crafted we have used Carlini & Wagner (C-W) attack [1] where the adversarial examples are crafted using the following optimization process provided in equation 2.

$$\begin{aligned} \underset{\eta}{\text{minimize}} \quad & \|\eta\|_{\mathcal{P}} + c.g(x^*) \\ \text{such that} \quad & x^* \in [0,1]^n \end{aligned} \tag{2}$$

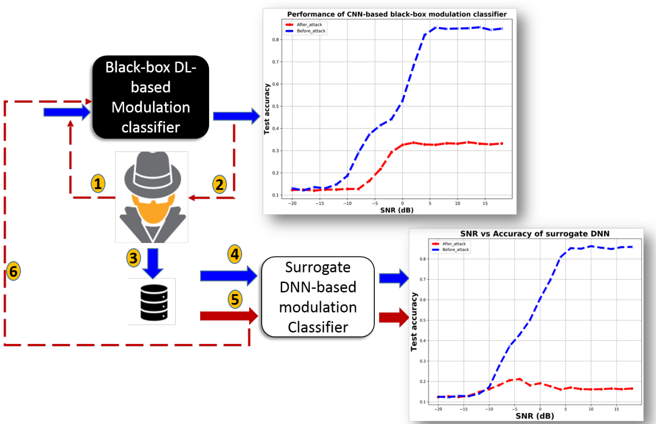

Figure 1: The step by step procedure followed for crafting black-box adversarial attack against DL-based modulation classification is depicted in the figure.

3 BLACK-BOX ADVERSARIAL ATTACK PROCEDURE

In this section, we will provide our black-box adversarial attack procedure (illustrated in Figure 1). The steps followed are: **1)** the adversary queries the deployed modulation classifier with test examples; **2)** the deployed modulation classifier provides a labeled response to the adversary considering the adversary as a normal user; **3)** the adversary stores the query-response pair in a database (which is later used as a substitute dataset for training a surrogate DNN); **4)** once sufficient data is collected in the adversarial database, the adversary constructs a fully connected DNN model and trains it for suitable classification performance; **5)** once the surrogate DNN is trained, the adversary launches a C-W attack on the surrogate DNN for crafting adversarial examples that compromises the performance of the surrogate DNN model; **6)** adversarial examples that compromises the performance of surrogate DNN-model are then transferred to black-box DL-based modulation classifier which according to the transferability property of adversarial examples will compromise the performance of DL-based modulation classifier.

Since we are performing this experiment in lab settings, we have opted for training two modulation classifiers based on CNN and LSTM and then considered them as black-box models. We have used highly-cited GNU radio ML RML2016.10a dataset [6] which provides 11 digital and analog modulation schemes on the SNR ranging from -20 dB to 18dB. We have used only 10% of the test examples to construct the surrogate classifier and then performed C-W attack the performance of the surrogate DNN model before and after the attack is provided in Figure 1. Once the adversarial attack on surrogate DNN is completed, we have transferred the adversarial examples that evaded the surrogate DNN to black-box modulation classifier by leveraging the transferability property of adversarial ML. The performance impact of the adversarial attack is provided in Figures 1 and 2. A clear drop in the accuracy of the modulation classifier after the adversarial attack highlights that our method of performing black-box adversarial attack has successfully compromised the performance crafted adversarial examples.

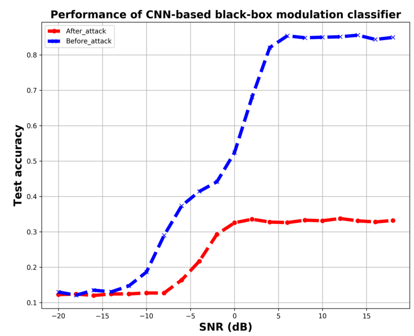

Figure 2: Performance of black-box adversarial attack on CNN-based modulation classification.

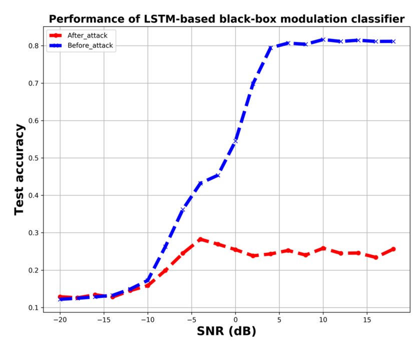

Figure 3: Performance of black-box adversarial attack on LSTM-based modulation classification.

4 CONCLUSIONS

In this paper, we have highlighted the lack of robustness in deep learning based modulation classification by performing a black-box adversarial attack on CNN and LSTM based modulation classifiers. We have used a surrogate deep neural network for crafting adversarial examples and then showed that adversarial examples crafted for modulation classification are transferable to other deep learning based models. We have achieved a 60% performance drop in both CNN and LSTM based modulation classification.

REFERENCES

[1] Carlini and Wagner. 2017. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, 39–57.

[2] Nick Feamster and Jennifer Rexford. 2017. Why (and how) networks should run themselves. *arXiv preprint arXiv:1710.11583* (2017).

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[4] Silvija Kokalj-Filipovic and Rob Miller. 2019. Adversarial Examples in RF Deep Learning: Detection of the Attack and its Physical Robustness. *arXiv preprint arXiv:1902.06044* (2019).

[5] Albert Mestres, Alberto Rodriguez-Natal, Josep Carner, Pere Barlet-Ros, Eduard Alarcón, Marc Solé, Victor Muntés-Mulero, David Meyer, Sharon Barkai, Mike J Hibbett, et al. 2017. Knowledge-defined networking. *ACM SIGCOMM Computer Communication Review* 47, 3 (2017), 2–10.

[6] Timothy J O'shea and Nathan West. 2016. Radio machine learning dataset generation with gnu radio. In *Proceedings of the GNU Radio Conference*, Vol. 1.

[7] Meysam Sadeghi and Erik G Larsson. 2018. Adversarial attacks on deep-learning based radio signal classification. *IEEE Wireless Communications Letters* 8, 1 (2018), 213–216.