Multiuser Adversarial Attack on Deep Learning for OFDM Detection

Youjie Ye, Yunfei Chen, *Senior Member, IEEE*, Mingqian Liu

***Abstract*—Adversarial attack has been widely used to degrade the performance of deep learning (DL), especially in the field of communications. In this letter, we evaluate different white-box and black-box adversarial attack algorithms for a DL-based multiuser orthogonal frequency division multiplexing (OFDM) detector subject to multiuser adversarial attack. The bit error rates under different adversarial attacks are compared. The results show that, the perturbation efficiency of adversarial attack is higher than conventional multiuser interference. Virtual adversarial methods (VAM) and zeroth-order-optimization (ZOO) attacks perform the best among white-box and black-box methods, respectively. They are also effective when the attack changes the starting time. Additionally, adding the number of attackers is found useful to improve the VAM attack but not for ZOO. This work shows that adversarial attack is powerful to generate adversarial against multiuser OFDM communications.**

***Index Terms*—Adversarial attack, deep learning, multiuser, OFDM, signal detection.**

I. Introduction

Orthogonal frequency division multiplexing (OFDM) modulation has been widely used to combat multipath fading in wireless channels [1]. In recent years, deep learning (DL) methods are popular in channel estimation and signal detection due to their high accuracy [2]. In [3], a convolutional neural network-minimum mean squared error (CNN-MMSE) based estimator was proposed to estimate the wireless channel. In [4], wireless channel response was treated as 2D images and restored by DL methods. Many such works consider OFDM systems. For example, in [5], residual learning was applied to OFDM channel estimation. In [6], a deep-neural-network(DNN)-aided channel estimation scheme was presented for multiple-input multiple-output-OFDM system. In [7], a channel estimation network and a channel-conditioned recovery network were presented for signal processing.In [8], fully connected DNN (FCDNN) was used to detect the transmitted symbols.

However, as deep learning based communications systems evolve, many researchers have found that it is not stable under targeted perturbation. In some applications, the perturbed model could have disastrous consequences for the safety and human life [9]. In [10], perturbation imperceptible to human was generated on images to fool DNN models, called adversarial attack. In [11], [12], and [13], adversarial attack was added to voice controllable system, object recognition system, and automatic speech recognition models, respectively. In recent years, researchers have paid more attention on adversarial attack of DL models in wireless systems [14]. In [15], adversarial samples were generated for learning based modulation classifiers. In [16], a generative-adversarial-network-based spoofing attack was proposed to fool DL-based signal classifier. In [17], adversarial attack and jamming attack were tested on DL-based autoencoder communication systems. The result shows that adversarial attacks are more destructive than jamming. In [18], white-box and black-box attacks were designed for DL-based signal classification. The attack leads to more misclassification than the conventional random noise. In [19], a perturbation generator against DNN-based wireless communication was tested. Thus, it is of great interest to examine different attack methods in OFDM.

In this letter, different adversarial attack methods will be evaluated against the DL based OFDM signal detector. Instead of adding the perturbation to the transmitted signal directly as in previous works [19], perturbation signal in this work will be regarded as an interfering user in the multiuser OFDM system, or as a multiuser adversarial attack. The DL model at the receiver recovers the transmitted signal against the attack disguised as a legitimate user. In this work, the adversarial robustness toolbox (ART) [20] is used to apply adversarial methods. For white-box methods, projected gradient descent (PGD) [22], virtual adversarial methods (VAM) [21], and elastic net attack (ENA) [23] are studied, while for black-box methods, boundary attack (BoA) [24], HopSkipJump (HSJ) attack [25], and zeroth-order-optimization (ZOO) attack [26] are used. Their performance are tested in different signal-to-interference ratios (SIRs) and channels. To the best of our knowledge, this is the first time that different attack methods are evaluated for multiuser OFDM communications systems. The novelty of the work includes the following:

- Adversarial signals are added as multi-user interference instead of transmitted samples.
- Three white-box methods and three black-box methods are evaluated and compared for OFDM DL detection.
- Useful guidance on the choice of the most efficient attacks in different conditions is given.

II. System Model

A. System architecture

The architecture of the attacked multiuser OFDM communication system is illustrated in Fig 1. For the desired

Youjie Ye and Yunfei Chen are with the School of Engineering, University of Warwick, Coventry, U.K. CV4 7AL. (e-mail: Youjie.Ye@warwick.ac.uk, Yunfei.Chen@warwick.ac.uk)

Mingqian Liu is with State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China. (e-mail: mqliu@mail.xidian.edu.cn)

This work is supported in part by EC H2020 DAWN4IoE-Data Aware Wireless Network for Internet-of-Everything under Grant 778305.

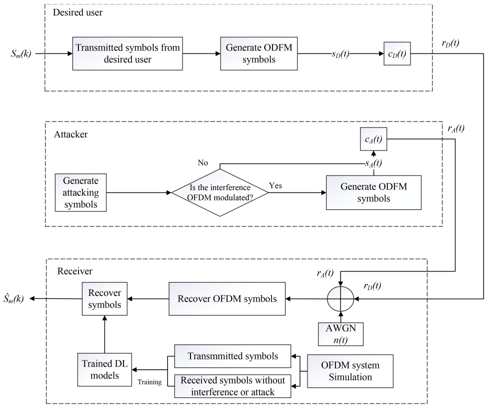

Fig. 1. System structure.

user, the transmitted symbols $S_m(t)$ with pilots are converted to parallel streams from a serial one, and then from the frequency domain to the time domain by inverse discrete Fourier transform (IDFT). Then, cyclic prefix (CP) is added and the signal is converted back to a serial stream and sent over the wireless channel. In the simulation, the attack size is the same as the number of transmitted symbols. The attacker may not know whether the desired user uses OFDM or not. Thus, two cases are considered. In the first case, the attack signal is also OFDM modulated before being transmitted. In the second case, the symbols are added directly to the channel without OFDM modulation. Signal transmitted from the attacker will interfere the desired signal as multiuser interference. In either case, the interfered signal will be OFDM demodulated at the receiver. The received signal $r(t)$ at the receiver can be represented as

$$\begin{aligned} r(t) &= r_D(t) + r_A(t) + n(t) \\ &= c_D(t) \otimes s_D(t) + c_A(t) \otimes s_A(t) + n(t), \end{aligned} \tag{1}$$

where $r_D(t)$ and $r_A(t)$ are the signals received from the desired user and the attacker, $\otimes$ represents the convolution operator, $c_D(t)$ and $c_A(t)$ are channel gains as time-varying complex Gaussian random processes, $s_D(t)$ and $s_A(t)$ are transmitted signals, respectively, and $n(t)$ is the additive white Gaussian noise (AWGN) with mean zero and variance σ^2.

At the receiver, the CP is removed, and the symbols are converted back to the frequency domain by discrete Fourier transform (DFT). Finally, $S_m(k)$ is recovered as $\hat{S}_m(k)$ using the pre-trained DL models. In this work, there are 128 symbols in a frame, 64 of which are pilots. The CP length is 16. The wireless channel is a multipath fading channel defined in [27] using MATLAB [28].

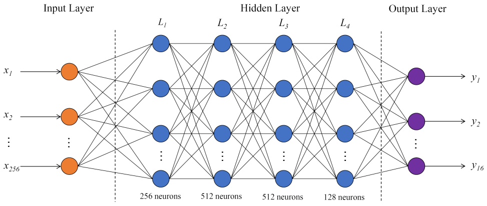

Fig. 2. Architecture of FCDNN.

In the physical layer of a wireless system, the transmitted symbols are 0 or 1. Therefore, there is a binary classification problem that the DL network try to solve at the detector. The attacked DL model uses FCDNN, details of which can be found in [8]. At the input layer, the 128 symbols in every frame will be divided into real and imaginary parts, and used as inputs separately to 8 parallel DL networks, which means each network detects 16 of them. This is shown in Figure 2. Compared to [8], the number of neurons in hidden layer has been increased to fit the multiuser condition. The results in $[0, 1]$ will be binarized and given as a size of 8×16 symbols at the output layer. Bit error rate (BER) is used to measure the detection error as $P_e = \mathrm{P}\left[\widehat{b_i} \neq b_i\right]$.

B. Adversarial attack methods

Adversarial attack methods can be divided into two types, white-box attack and black-box attack. For white-box attack, the attacker knows all the information and parameters of the DL model and it generates adversarial samples based the known model to attack the network. Black-box attack only interacts with the DL model to generate attacking samples, and then attacks the network without knowing the parameters and structure of the model. As mentioned before, VAM, PGD, ENA are white-box attacks, while BoA, HSJ, ZOO are black-box attacks.

PGD is an iterative extension of the widely-used gradient-based attack method fast gradient sign method (FGSM) [20]. The gradient-based attack intends to find the perturbation $\boldsymbol{\eta}$ to maximize the loss function $L(\mathbf{x} + \boldsymbol{\eta}, \mathbf{y}, \boldsymbol{\theta}))$ based on the constraint Δ and the optimization as

$$\max_{\boldsymbol{\eta} \in \Delta} L(\mathbf{x} + \boldsymbol{\eta}, \mathbf{y}, \boldsymbol{\theta}), \tag{2}$$

where $\mathbf{x}$ denotes the input of neural network, $\mathbf{y}$ denotes the true label of $\mathbf{x}$, and $\boldsymbol{\theta}$ denotes the parameter of the DL model. FGSM [29] generates attacks by using the sign of the gradient function as

$$\mathbf{x}' = \mathbf{x} + \epsilon \cdot sign(\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})), \tag{3}$$

where ϵ is the attack strength. Different from FGSM, PGD projects the adversarial perturbations on the $\epsilon - L_\infty$-norm ball around $\mathbf{x}$ at each iteration as

$$\mathbf{x}'_{t+1} = Proj(\mathbf{x}_t + \alpha \cdot sign(\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}))), \tag{4}$$

where $Proj$ is the constrained projection operation in a PGD standard optimization, and α is the step size of the gradient descent update.

Carlini and Wagner's (C&W) attack [30] is a popular gradient-based attack method. It can generate attack samples of different norms, such as ℓ_1, ℓ_2, and ℓ_∞. ENA is an advanced version of C&W, which generates perturbation by using the elastic-net regularization method and minimizing the ℓ_1 norm [23]. ZOO attack is a black-box version of C&W(ℓ_2) attack. It queries the gradient of the objective function to the input in each iteration based on stochastic coordinate descent method [26]. VAM maximizes the Kullback-Leibler (KL) divergence between output distributions to find the ℓ_2 norm bounded perturbation [21]. Different from the methods above, VAM aims to generate the adversarial sample which can affect the trained model local distributional smoothness instead of the sample [20].

BoA and HSJ are decision based attacks. BoA is the earliest successful decision based attack. It only needs to query output classes, and perturbs an adversarial sample along the decision boundary between the non-adversarial and the adversarial region until the ℓ_2 difference from the original input to the perturbed input is minimized [24]. HSJ is an improved version of BoA by optimizing ℓ_2 or ℓ_∞ distances for attacks [20]. In each iteration, binary search is used to approach the decision boundary iteratively. Then, the gradient direction is estimated. Finally, the updating step size is initialized and is decreased until the perturbation is successful [25].

III. Numerical Results and Discussion

In this section, the BERs of the DL model under different attacks are compared to measure their attack efficiency. We use 4-ary quadrature amplitude modulation (4-QAM) signalling for the desired user. For ZOO, BoA, and HSJ, we set the maximum number of iterations as 200, 200, and 50, respectively, which have been tested to have the best attack. The norm of HSJ is set as ℓ_∞. Other methods use the default settings of ART functions. The test is done when the multiuser SIR changes from 0 dB to 50dB, where SIR represents the ratio of the desired signal power to the attack or general multiuser interference power. The SNR is set to 15dB in all the tests. There are two baselines. The first one is the BER of 0.148 when there is no multiuser interference or attack at SNR = 15dB. This is called the no-attack error floor. The second is the BER with a general multiuser interference, as a general QAM signal being received at the receiver with or without OFDM modulation. A QAM signal source similar to the desired user is simulated as the general interference to the receiver so that one can compare the performance degradation caused by adversarial attack with that caused by general multiuser interference.

A. Comparison of Attack Methods

Fig. 3 gives the BER comparison of the DL model under different white-box and black-box attacks. All the methods have been tested both with and without OFDM modulation. However, for better readability of the figure, results for white-box are only shown with OFDM modulation, while results for black-box are only shown without OFDM modulation. As the SIR increases, the BERs under different attacks decrease and approach the no-attack error floor as a lower limit, because when SIR is higher, the proportion of attack signal is lower, which leads to less interference to the receiver. For white-box methods, most are more efficient than the general interference. When SIR $\geq$ 25dB, the BERs of DL model with general interference (with or without OFDM modulation) reach the no-attack error floor. Similarly, BERs under PGD attack also reach the error floor when SIR $>$ 25dB. The error floor under PGD attack is about 0.0198. ENA and VAM attacks perform better than PGD, especially in the high SIR region. When SIR = 50dB, their BERs are still 0.045 and 0.06, respectively, which are 200% and 300% higher than the no-attack error floor. It shows that these two attack methods are still efficient even when the attack signal is weak at the receiver. VAM is the most powerful white-box attack method, which leads to the largest degradation of BERs when SIR>5dB.

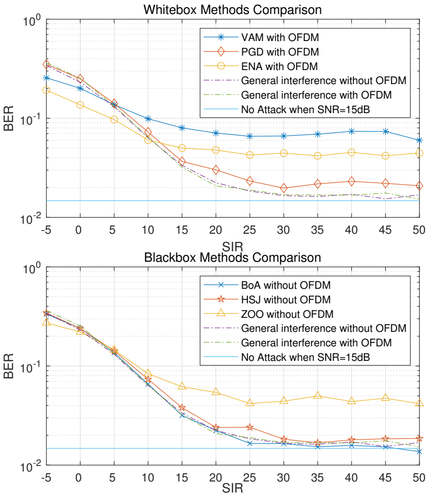

Fig. 3. BER comparison of different attack methods.

Since the parameter and gradient inside the DL model is not available to black-box attacks, the performances of black-box attacks are not as good as white-box ones. However, most of them can still cause more misclassification than the general interference without any intentional attack designs. BoA is the least powerful method among black-box attacks. The BERs under BoA are near that of general interference. The performance of HSJ attack is slightly better than BoA. The BERs under HSJ are about 0.05 higher than BoA on average. ZOO attack is the most efficient, especially at high SIR. BERs under ZOO are at a stable level of 0.042 to 0.05, which are much higher than the BERs with other methods. But when SIR $\leq$0 dB, ZOO is not powerful. When SIR = -5dB, BER under ZOO attack is 0.273, which is 0.07 lower than that of HSJ.

In summary, the order of the performances of white-box attacks from high to low within the SIR range considered is: VAM, ENA, and PGD, and that of black-box attacks is: ZOO, HSJ, and BoA.White-box attacks are more efficient than black-box ones.

B. Random starting time

In practice, due to asynchronous operations, frames from the desired user and the attacker usually cannot achieve synchronization. To investigate its potential effect on attack efficiency, the BERs with asynchronous users are studied. Each user has its own random starting time, following a uniform distribution between 1 and the frame size 128. As the VAM and ZOO attacks give the best performances among white-box and black-box methods, respectively, they are used in following study.

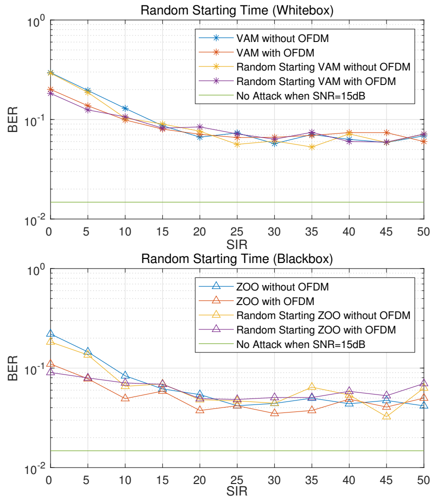

Fig. 4. BER comparison with uniformly distributed starting time for the frame.

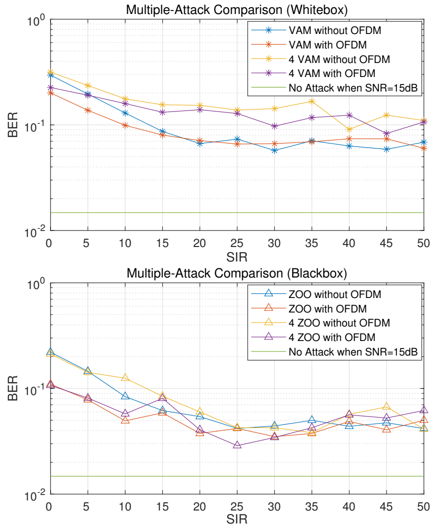

Fig. 5. BER comparison when multi-attack.

Fig. 4 shows the BER comparison. Both non-OFDM-modulated VAM and ZOO outperforms OFDM-modulated ones at high SIR region. For VAM attacks, no matter whether the perturbation is OFDM modulated or not, the same level of BER is achieved. It can be seen that, even in a high SIR region, the attack still has a stable efficiency. Similarly, although the parameters of the DL model are not known, the ZOO attack is not affected by asynchronous users. The results show that, the random starting time has almost no impact on the attack efficiency of VAM and ZOO methods.

C. Multi-attacker experiment

In order to further improve the attack efficiency of adversarial methods, several attackers are used to generate multiple attacks. Fig. 5 shows the BER comparison between one and four attackers for VAM and ZOO. In this case, the SIR is still the receiving SIR, which is the same for one attacker and four attackers. From Fig. 5, for the VAM attack, the BERs under quadruple attack are higher than that with one attacker, which means multiple VAM attacks can increase the attack performance in perturbation capability. When the VAM attack is OFDM-modulated, quadruple attack increases the BER by 2.85% to 33.74% over the single attack. For non-OFDM-modulated VAM attack, the gap between quadruple attack and single attack is even larger. The BER increases by 7.49% to 47.74% over the single attack. However, adding the number of attackers has no impact for ZOO. There is no significant increase of BER when using multi-ZOO-attack. Therefore, multi-attack is effective to enhance the performance of VAM attack, but not ZOO.

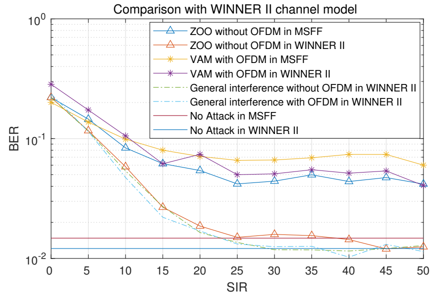

Fig. 6. Comparison with BERs under WINNER II channel.

D. Realistic channel experiment

To evaluation the attack methods under realistic channel conditions, they are studied in the WINNER II channel model [31]. As [8], we used the typical urban channels with a maximum delay of 16 sampling period. The frequency of the carrier is 2.6 GHz, and the number of paths is 24. The DL model is re-trained for this channel. Fig. 6 compares the MATLAB simulated frequency-selective fading (MSFF) channel in [28] with the WINNER II channel. BERs under VAM and ZOO attacks in WINNER II channel are both lower than those in MSFF. In low SIR region, VAM has the same level of attack efficiency. Although ZOO still outperforms the general interference case, its BER in WINNER II is much lower than in MSFF. It is because the DL model trained in WINNER II channel has stronger anti-interference ability and better protection from the black-box method. Thus, VAM attack is more effective.

IV. Conclusion

In this letter, the performances of adversarial attack algorithms have been compared.VAM and ZOO are the most effective in white-box and black-box methods, respectively.The experiments have also shown that, when there is random starting time, attack efficiencies of these two methods will not be affected. In addition, VAM's performance can be improved by adding attackers to perform multi-attack and VAM is proved efficient in WINNER II channel.

References

[1] J.G.Proakis, *Digital Communications*, 4th ed. New York: McGraw-Hill, 2001.

[2] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[3] D. Neumann, T. Wiese and W. Utschick, ”Learning the MMSE Channel Estimator,” in *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2905-2917, 1 June1, 2018.

[4] M. Soltani, V. Pourahmadi, A. Mirzaei and H. Sheikhzadeh, ”Deep Learning-Based Channel Estimation,” in *IEEE Communications Letters*, vol. 23, no. 4, pp. 652-655, April 2019.

[5] L. Li, H. Chen, H. Chang and L. Liu, ”Deep Residual Learning Meets OFDM Channel Estimation,” in *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 615-618, May 2020.

[6] A. L. Ha, T. Van Chien, T. H. Nguyen, W. Choi and V. D. Nguyen, ”Deep Learning-Aided 5G Channel Estimation,” *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2021, pp. 1-7.

[7] X. Yi and C. Zhong, ”Deep Learning for Joint Channel Estimation and Signal Detection in OFDM Systems,” in *IEEE Communications Letters*, vol. 24, no. 12, pp. 2780-2784, Dec. 2020.

[8] H. Ye, G. Y. Li and B. Juang, ”Power of Deep Learning for Channel Estimation and Signal Detection in OFDM Systems,” in *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 114-117, Feb. 2018.

[9] X. Yuan, P. He, Q. Zhu and X. Li, ”Adversarial Examples: Attacks and Defenses for Deep Learning,” in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805-2824, Sept. 2019.

[10] C. Szegedy *et al.* (2013). “Intriguing properties of neural networks.” [Online]. Available: https://arxiv.org/abs/1312.6199

[11] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, “DolphinAttack: Inaudible voice commands,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, New York, NY, USA, Oct. 2017, pp. 103–117.

[12] C. Xie, J. Wang, Z. Zhang, Y . Zhou, L. Xie, and A. Y uille, “Adversarial examples for semantic segmentation and object detection,” in *Proc. Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1378–1387.

[13] N. Carlini et al., “Hidden voice commands,” in *Proc. USENIX Security Symp.*, 2016, pp. 513–530.

[14] J. Liu, M. Nogueira, J. Fernandes and B. Kantarci, ”Adversarial Machine Learning: A Multilayer Review of the State-of-the-Art and Challenges for Wireless and Mobile Systems,” in *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 123-159, Firstquarter 2022.

[15] M. Usama, M. Asim, J. Qadir, A. Al-Fuqaha and M. A. Imran, ”Adversarial Machine Learning Attack on Modulation Classification,” *2019 UK/ China Emerging Technologies (UCET)*, 2019, pp. 1-4.

[16] Y. Shi, K. Davaslioglu and Y. E. Sagduyu, ”Generative Adversarial Network in the Air: Deep Adversarial Learning for Wireless Signal Spoofing,” in *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 1, pp. 294-303, March 2021.

[17] M. Sadeghi and E. G. Larsson, ”Physical Adversarial Attacks Against End-to-End Autoencoder Communication Systems,” in *IEEE Communications Letters*, vol. 23, no. 5, pp. 847-850, May 2019.

[18] M. Sadeghi and E. G. Larsson, ”Adversarial Attacks on Deep-Learning Based Radio Signal Classification,” in *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 213-216, Feb. 2019.

[19] Alireza Bahramali, Milad Nasr, Amir Houmansadr, Dennis Goeckel, and Don Towsley, ”Robust Adversarial Attacks Against DNN-Based Wireless Communication Systems,” in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS ’21)*, Association for Computing Machinery, New York, NY, USA, 126–140, 2021.

[20] Nicolae, M.-I., M. Sinn, M. N. Tran, et al. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018.

[21] T. Miyato, S. Maeda, M. Koyama and S. Ishii, ”Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979-1993, 1 Aug. 2019.

[22] 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.

[23] P. Y. Chen, Y. Sharma, H. Zhang, J. F. Yi, and C. Hsieh, “EAD: Elastic-net attacks to deep neural networks via adversarial examples,” in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, Louisiana, USA, 2018, pp. 10–17.

[24] W. Brendel, J. Rauber, and M. Bethge “Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models,” *arXiv preprint arXiv:1712.04248v2*, 2018.

[25] W. Brendel, J. Chen, M. I. Jordan and M. J. Wainwright, ”HopSkipJumpAttack: A Query-Efficient Decision-Based Attack,” *2020 IEEE Symposium on Security and Privacy (SP)*, 2020, pp. 1277-1294.

[26] P. Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C. J. Hsieh, “ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models,” *arXiv preprint arXiv:1708.03999*, 2017.

[27] O. Edfors, M. Sandell, J. . -J. van de Beek, S. K. Wilson and P. O. Borjesson, ”OFDM channel estimation by singular value decomposition,” in *IEEE Transactions on Communications*, vol. 46, no. 7, pp. 931-939, July 1998.

[28] V. K. Veludandi, “LMMSE based channel estimation for OFDM systems,” *https://github.com/vineel49/lmmse,*

[29] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y . Bengio and Y . LeCun, Eds., 2015.

[30] N. Carlini and D. A. Wagner, “Towards evaluating the robustness of neural networks,” in *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017.* IEEE Computer Society, 2017, pp. 39–57.

[31] P. Kyosti *et al.*, “WINNER II channel models,” Eur. Commission, Brussels, Belgium, Tech. Rep. D1.1.2 IST-4-027756-WINNER, Sep. 2007.