A Robust Adversarial Network-Based End-to-End Communications System with Strong Generalization Ability Against Adversarial Attacks

Yudi Dong*, Huaxia Wang†, and Yu-Dong Yao‡

*Department of Computer & Information Science, Gannon University
†College of Engineering, Architecture and Technology (CEAT), Oklahoma State University
‡Department of Electrical & Computer Engineering, Stevens Institute of Technology

Abstract—End-to-end learning of communications systems is a promising new paradigm for future communications, in which deep neural networks (DNNs) are implemented in the transmitter and receiver as an autoencoder architecture. However, due to DNN’s natural vulnerability to adversarial perturbations, the end-to-end communications system exhibits security and robustness issues in terms of adversarial attacks over the air. The common defensive method, known as adversarial training, is to augment training data with adversarial perturbations, but it is hard to cover all possible perturbations and also hurt the system generalization. In this paper, we propose a novel and defensive mechanism based on a generative adversarial network (GAN) framework to achieve robust end-to-end learning of a communications system. We utilize a generative network to model a powerful adversary and enable the end-to-end communications system to combat the generative attack network via a minimax game. We show that the proposed system not only works well against white-box and black-box adversarial attacks but also possesses excellent generalization capabilities to maintain good performance under no attacks. The results also show that our GAN-based system outperforms the conventional communications system and the autoencoder communications system with/without adversarial training.

Index Terms—Adversarial networks, Wireless communications security, Adversarial attacks, Robust end-to-end learning

I. INTRODUCTION

Deep neural networks (DNNs) bring wireless communications into a new era of deep learning and artificial intelligence. One of the insightful ideas is end-to-end learning of communications systems [1], which re-designs the physical layer by employing a neural network instead of multiple independent blocks at the transmitter and the receiver. Particularly, an autoencoder architecture [2] is utilized for end-to-end communications, where an encoder neural network (NN) and a decoder NN are respectively utilized in the transmitter and receiver to replace signal processing tasks. Through jointly training the transmitter NN and the receiver NN, the end-to-end communications system can achieve global optimization and considerable performance improvements [1].

However, neural networks hold an inherent/natural vulnerability to adversarial attacks [3], [4]. That is, a neural network model easily leads to a false output by adding a small perturbation into the input of a neural network. Such perturbation, called adversarial perturbation, is an elaborate vector designed based on the receptive fields of inputs in the neural network model. This vulnerability threatens deep learning based communications systems in terms of robustness and security. A recent work [5] investigates adversarial attacks against autoencoder end-to-end communications systems, which crafts universal adversarial perturbations using a fast gradient method (FGM) [4]. By leveraging the broadcast nature of the wireless channel, attackers can inject adversarial perturbations into the input of the receiver NN [5], which causes a more significant impact on the end-to-end learning based systems than classical communications systems [5]. In addition to the end-to-end communications systems, deep learning models deployed in separate communications blocks also have the same concerns, where adversarial perturbations are designed to cause performance degradation on a neural network-based modulation classification system [6], [7] and a neural network-based channel state information (CSI) feedback [8] system.

A direct defensive method against adversarial attacks is to train the deep learning model with adversarial perturbations, which is called adversarial training [4], [9]. However, adversarial training is a passive defense mechanism. It only works for some specific adversarial perturbations that have been added to the training. For other various and new adversarial perturbations, adversarial training may be incapable of defense [5]. Also, adversarial training hurts the generalization ability of neural networks [10], [11], which can reduce the performance of neural networks on unperturbed/clean inputs. Therefore, it is difficult to apply adversarial training to the increasingly complex communications environments/scenarios with a variety of adversarial perturbations. In addition to adversarial training, some recent works [12], [13] use generative adversarial networks (GAN) to actively defend against adversarial attacks. However, these GAN methods are used for a DNN-based image classifier, which are neither suitable for an autoencoder-based end-to-end system nor valid for a communications system. Therefore, a more effective and active defense mechanism is desired for robust deep learning of end-to-end communications systems.

To this end, in this paper, we propose to integrate the GAN...
framework [14] into the autoencoder based end-to-end communications system for defending against various adversarial attacks. We utilize a generative network as an adversary to generate adversarial perturbations that can fool the receiver NN into recovering the false message. By leveraging the great computational capacity of the neural network, the generative network can generate various and powerful perturbations. Meanwhile, the discriminative network is the receiver/decoder NN of the end-to-end system, which is responsible to recover the correct message from both clean signal and perturbed signal with adversarial perturbations generated by the generative network. Triple networks, including the generative network, the encoder NN, and the discriminative network (i.e., decoder NN) are jointly trained in a confrontation game, where the generative network becomes a powerful adversary while the discriminative network becomes a robust defender.

The main contributions of our paper are as follows.

- This work is the first to resolve the security and robustness issue induced by adversarial attacks in the end-to-end communications system, where we build a robust and defensive GAN-based end-to-end communications system through triple-training, where we jointly and adversarially training an encoder communications NN and an decoder communications NN against a generative attack NN.

- Unlike the adversarial training method that is hard to gain simultaneously defense and generalization capacity, the proposed approach can effectively defend against various adversarial attacks including white-box attacks and black-box attacks for communications systems and, meanwhile, it has excellent generalization performance to remain in low error rates on clean inputs of communications systems.

- Consensus optimization is utilized in the triple-training of the proposed end-to-end communication system, which ensures a stable and impartial minimax game to train a defensive end-to-end communications system.

- Extensive experiments show that our GAN-based end-to-end communications system is more robust than both the classical communications system and the autoencoder end-to-end communications system with regular training or adversarial training.

II. ADVERSARIAL ATTACKS ON END-TO-END COMMUNICATION SYSTEMS

In the section, we introduce the preliminary knowledge regarding autoencoder based end-to-end communications system and adversarial attacks. Also, we talk about the attack model and the method of crafting adversarial perturbations for attacking a end-to-end communications system.

A. AUTOENCODER BASED END-TO-END COMMUNICATIONS SYSTEM

A conventional communications system is composed of many processing blocks (e.g., channel encoding/decoding, modulation, and signal detection) that are designed with mathematical models. Due to its complicated block structure and fixed mathematical expressions, the conventional communications framework is getting harder to achieve performance improvements and system optimization [15]. To this end, deep learning technologies are introduced to achieve an end-to-end communications system [1], which enables robust communications and overall optimization by jointly training two neural networks. In particular, an encoder NN and a decoder NN are respectively equipped in the transmitter and receiver, which can be seen as an autoencoder architecture from the perspective of deep learning [2].

Fig. 1 illustrates a typical end-to-end autoencoder communication system [1], which is implemented in this paper. Specifically, the message $s$ that needs to be transmitted is chosen from a message set $\mathcal{M} = \{1, 2, \cdots, M\}$, where $M = 2^k$ and $k$ is the bit number of a message. The message $s$ is first preprocessed as a one-hot binary vector $o \in \mathbb{R}^M$ where the $s^{th}$ element of $o$ is equal to one and all the others are zero. Then the one-hot message goes through the encoder NN to perform a mapping: $E_{\alpha} : \mathcal{O} \mapsto \mathbb{R}^{2^n}$, which generates the output signal $x = E_{\alpha}(o) \in \mathbb{R}^{2^n}$, where $E_{\alpha}$ refers the encoder model with parameters $\alpha$, $\mathcal{O}$ is the message set via one-hot calculation, $n$ refers to the number of channel uses and $x$ is a concatenation of the real and imaginary parts of the transmitted signal. Consider the hardware constraints of a transmitter, we restrict the energy of the transmitted signal as $\|x\|_2^2 \leq \frac{P}{2}$. Next, a additive white Gaussian noise (AWGN) channel is used for the transmission of $x$ to obtain the received signal $y \in \mathbb{R}^{2^n}$, where $y$ involves $x$ and noise. We assign the fixed variance $\sigma^2 = (2RE_{b0}/N_0)^{-1}$ in the AWGN channel, where $R = k/n$, computed by bit number $k$ and $n$ channel uses, is the data rate in our communications system, and $E_{b0}/N_0$ is the energy per bit to noise power spectral density ratio. Finally, the decoder NN performs a mapping $D_{\theta} : \mathbb{R}^{2^n} \mapsto \mathcal{M}$ to recover the estimated message $\hat{s} = D_{\theta}(y)$, where $D_{\theta}$ is the decoder model parameterized by $\theta$. In particular, the softmax layer of the decoder NN generates the vector $(0, 1)^M$. The estimated message $\hat{s}$ is determined as the index of the highest value in the output vector $(0, 1)^M$.

B. ADVERSARIAL ATTACKS

Neural networks have a natural vulnerability to adversarial attacks, where input with adversarial perturbations can lead a well-trained neural network to output a wrong answer with high confidence [3], [4]. An adversarial perturbation is a carefully crafted vector or matrix with small values, which are imperceptible but sensitive to neural networks. Due to this property of neural networks, the security and robustness of deep learning-based systems are compromised.
by adversarial attacks. In our case, an autoencoder based end-to-end communications system can be easily fooled by using adversarial attacks. As shown in Fig. 2, attackers can leverage the broadcast nature of the channel and emit an interference signal of adversarial perturbation $p$ to the channel. The perturbed received signal $y + p$ forces the decoder NN to provide an incorrect output. Under adversarial attacks, autoencoder communications systems have more significant performance degradation than conventional communications systems [5].

According to the knowledge of attackers, adversarial attacks can be divided into white-box attacks and black-box attacks [16]. In white-box attacks, an attacker has total knowledge of the NN model $D_\theta$, including the architecture of the NN, the hyper-parameters for training, and the distribution of training data. In black-box attacks, attackers only knows the output of the decoder model but have no information about the NN model.

C. Attack Model: Crafting Adversarial Perturbation

To perform white-box attacks on the decoder NN model $D_\theta$ that generates the estimated message $\hat{s} = D_\theta(y)$, we need to find an adversarial perturbation $p$ such that $y + p$ results in a wrong output, which is described as

$$
\arg \min_p \|p\|_2 \quad \text{s.t.} \quad \arg \max D_\theta(y + p) \neq D_\theta(y). \quad (1)
$$

To solve the problem (1) of generating adversarial perturbations, the FGM method [17], [4] is commonly used to get an optimal $\ell_2$-norm constrained perturbation,

$$
p = \epsilon \cdot \frac{\nabla_y l(D_\theta(y), s)}{||\nabla_y l(D_\theta(y), s)||_2}, \quad (2)
$$

where $\epsilon$ is a scaling coefficient, $l$ denotes the loss function, and $\nabla_y$ is the gradient of the loss function $l$ with respect to the input $y$. However, FGM requires the knowledge of the message $s$ that is unknown to the transmission process. Therefore, Sadeghi et al. introduce an input-agnostic FGM [5] to generate an universal perturbation $p$ that works for all messages from $\mathcal{M}$. This method is used in this paper for crafting adversarial perturbations. For the black-box attacks, attackers cannot obtain any information about our autoencoder. Thus, attackers need to design white-box perturbations based on a substitute autoencoder system that is fully open to attackers. These adversarial perturbations are also effective for other unknown autoencoder systems due to the transferability of adversarial attacks [18]. We use this general approach to perform black-box attacks on the autoencoder system and the proposed system.

Fig. 3 shows the block-error-rate (BLER) of an autoencoder end-to-end communications system with $n = 7$ channel uses and $k = 4$ bits per channel, and the BLER of a conventional communications system using binary phase-shift keying (BPSK) modulation and Hamming (7,4) code with hard-decision (HD) decoding [1], [5]. The BLER is calculated as the ratio of $\hat{s} \neq s$. The smaller BLER indicates the better system performance. We can see that the autoencoder outperforms the conventional scheme if there is no attack. However, by performing the adversarial attacks using the input-agnostic FGM, the performance of the autoencoder is degraded more significantly, where the performance of the autoencoder is worse than the conventional scheme. This paper is to address this issue induced by adversarial attacks in the autoencoder end-to-end communications system.

III. ROBUST END-TO-END LEARNING OF COMMUNICATIONS SYSTEMS USING ADVERSARIAL NETWORKS

We adopt the GAN approach [14] to defend against adversarial attacks. The original goal of the GAN is to learn a generative neural network based on a discriminative network. The generative network generates samples as similar as possible to real training samples to fool the discriminative network. Meanwhile, the discriminative network needs to distinguish the real training samples against samples synthesized by the generative network. These two networks are trained jointly to counter each other until the generated samples cannot be distinguished from real samples by the discriminative network.

We integrate the GAN framework into an autoencoder communications system. As shown in Fig. 4, one more neural network is added as the generative network $G_{\phi}$ parameterized by $\phi$. The decoder NN $D_\theta$ serves as a discriminative network. In our case, the purposes of the generative and discriminative networks are different from the original GAN. Here the generative network acts as an adversary to model and generate adversarial perturbation (i.e., $\epsilon G_{\phi}(x)$) based on the input $x$ and a scaling factor $\epsilon$. The discriminative network tries to
estimate the correct message from both clean signal $y$ and perturbed signal $y + \epsilon G_\phi(x)$. The generative and discriminative networks are trained jointly and adversarially, where the generative network generates more and more powerful adversarial perturbation but the discriminative network still correctly estimates messages from the more perturbed signals. With the proposed adversarial network-based approach, the autoencoder communications system can obtain a strong capability to defend against adversarial attacks.

### A. Objective Function

The intuition of an ideal defensive method is to find a solution $\theta$ of the decoder NN that simultaneously has the small loss $L(\theta)$ on the clean inputs and the small loss $L_p(\theta)$ on the inputs with adversarial perturbations,

$$L(\theta) = l(D_\theta(y), s)$$

$$L_p(\theta) = l(D_\theta(y + p), s).$$

However, the solution of $L(\theta)$ and the solution of $L_p(\theta)$ are hard to be the same. There is a trade-off between $L(\theta)$ and $L_p(\theta)$. The traditional adversarial training usually satisfies either the small loss $L(\theta)$ of clean input or the small loss $L_p(\theta)$ of the perturbed inputs, which causes the model $D_\theta$ to lose either defense ability or generalization ability.

To satisfy the above two conditions, we try to find a optimal parameter $\theta$ of the decoder NN $D_\theta$ to minimize the loss between the output of the clear signal $y$ and the groundtruth $s$, as well as minimize the loss between the output of the perturbed signal $D_\theta(y + p)$ and the groundtruth $s$, where the objective of the decoder NN is

$$\arg\min_{\theta} \left[ l(D_\theta(y), s) + l(D_\theta(y + p), s) \right].$$

In our approach, we model the adversarial perturbation using the generative neural network $G_\phi$, then the objective of the decoder NN becomes

$$\arg\min_{\theta} \left[ l(D_\theta(y), s) + l(D_\theta(y + \epsilon G_\phi(x)), s) \right],$$

where $\epsilon G_\phi(x)$ is the generated adversarial perturbation. In order to enable the decoder NN to handle as various perturbations as possible, we want the generative neural network $G_\phi$ to be a powerful adversary, where we need to find a network parameter $\phi$ to maximize the loss between the output of the perturbed signal $D_\theta(y + \epsilon G_\phi(x))$ and the groundtruth $s$,

$$\arg\max_{\phi} l(D_\theta(y + \epsilon G_\phi(x)), s).$$

Finally, we jointly train the decoder NN (i.e., discriminative network) and the generative network to find a solution of a minimax game between $D_\theta$ and $G_\phi$,

$$\arg\min_{\theta} \max_{\phi} \left[ l(D_\theta(y), s) + l(D_\theta(y + \epsilon G_\phi(x)), s) \right]$$

$$+ l(D_\theta(y + \epsilon G_\phi(x)), s),$$

where our final objective is realized that the discriminative network is capable of countering against a powerful adversary while has a good generalization performance.

### B. Consensus Optimization For GAN Training

GAN training is always unstable, which suffers from problems of non-convergence, mode collapse, and diminished gradient. In this paper, we adopt a consensus optimization approach [19] to regularize gradients to stabilize the GAN training.

We can denote the objective of the discriminative network (i.e., Eq. 6) as $d(\theta, \phi)$ and denote the objective of the generative network (i.e., Eq. 7) as $g(\theta, \phi)$. The gradient vector field $v(\theta, \phi)$ of this minimax game is defined as

$$v(\theta, \phi) = \left( \frac{\nabla_\theta d(\theta, \phi)}{\nabla_\phi g(\theta, \phi)} \right).$$

The GAN training is to find a solution of $v(\theta, \phi) = 0$. However, the eigenvalues of the Jacobian of $v(\theta, \phi)$ could be zero in real part or be big in imaginary part [19], which results in the convergence failure of GAN training. To this end, we respectively add a regularization factor $\mathcal{L}(\theta, \phi) = \frac{1}{2} \|v(\theta, \phi)\|^2$ to the objectives of the discriminative network and the generative network. Then, the new gradient vector field $v_s(\theta, \phi)$ is obtained [19]

$$v_s(\theta, \phi) = \left( \frac{\nabla_\theta d(\theta, \phi) - \gamma \mathcal{L}(\theta, \phi)}{\nabla_\phi g(\theta, \phi) - \gamma \mathcal{L}(\theta, \phi)} \right),$$

where $\gamma$ is a constant parameter for regularization. This added regularization factor $\mathcal{L}(\theta, \phi)$ can help two networks to reach a consensus optimization wit a better convergence.

### IV. Evaluation Results

#### A. Neural Network Architecture

We implement our adversarial network based approach into two different end-to-end communications systems: One is a multilayer perceptron (MLP) based end-to-end communications system and the other is a convolutional neural network (CNN) based end-to-end communications system, which are shown in Table I and Table II, respectively.

For the MLP based system, the architectures of neural networks are shown in Table I. The encoder NN and decoder NN both consist of two FC layers. The generative network is a two-layer CNN without BN layers and a $l_2$ normalization layer. For the CNN based system, the encoder NN consists of one fully connected (FC) layer with eLU activation function, one convolutional layer with ReLu activation function, one FC layer with linear activation, and one normalization layer. The decoder NN consists of two convolutional layers with ReLu
TABLE I: NN Architectures used in our approach (MLP based).

| Layer                  | Encoder NN               | Decoder NN               | Generative Network          |
|------------------------|--------------------------|--------------------------|-----------------------------|
| FC+elu                 | FC+relu                  | Conv1d+ReLU (Flatten)    |
| FC+linear+ $\ell_2$ Norm | FC+softmax               | FC+linear                |
| Normalization ($\ell_2$) | FC+relu                  | FC+linear                |

TABLE II: NN Architectures used in our approach (CNN-based).

| Layer                  | Encoder NN               | Decoder NN               | Generative Network          |
|------------------------|--------------------------|--------------------------|-----------------------------|
| Conv1d+relu+Flatten    | Conv2d+relu+Flatten      | Conv1d+relu+ BN+Flatten  |
| FC+Linear              | FC+relu                  | FC+linear                |
| Normalization ($\ell_2$) | FC+softmax               | FC+linear                |

activation function, one FC layer with ReLU activation function, and one FC layer with softmax function. The generative network is composed of two convolutional layers with ReLu activation function and batch normalization (BN) layer, one FC layer with linear activation, and one $\ell_2$ normalization layer. The encoder NN and decoder NN used in these two systems are the same in [1], [5].

For the design of the generative network, one noticed rule is that the depth (i.e., number of layers) of the generative network and the decoder NN (i.e., discriminative network) should be similar, which can reach equal competition between the generative network and the discriminative network to result in better performance. Our approach is easily applied to various end-to-end communications systems with different architectures of neural networks. In addition, since the adversarial perturbations used to attack the systems are with $\ell_2$-norm, a $\ell_2$ normalization layer is added to the generative network to generate $\ell_2$-norm perturbations.

B. Experiment Setup

In the experiments under white-box attacks, the proposed system uses the network architecture listed in Table I. The autoencoder system uses the same MLP encoder and MLP decoder listed in Table I. The conventional communications system uses BPSK modulation and Hamming coding with HD decoding. The adversarial perturbations for attacking these three systems are generated using FGM [5] based on the MLP decoder. In the experiments under black-box attacks, the proposed system uses the network architecture listed in Table II. The autoencoder system uses the same CNN encoder and CNN decoder in Table II. The conventional communications system also uses BPSK modulation and Hamming coding with HD decoding. The adversarial perturbations for black-box attacks are generated from the MLP decoder.

In addition, the proposed system and the autoencoder system are established using TensorFlow on a NVIDIA Quadro P4000 GPU. The systems are all sufficiently trained using stochastic gradient descent (SGD) with two batch sizes 1,000 and 10,000, three learning rates $10^{-3}$, $10^{-4}$, and $10^{-5}$ for total 31,000 iterations.

C. Proposed Approach versus Conventional Communications System

We first compare our proposed GAN-based communications system with the conventional communications system under adversarial attacks and no attack. In Figure 5, we can see that the performance of our proposed system is better than the conventional communications system under no attacks. When we attack these two systems using white-box attacks, as shown in Figure 5(a), our system can mitigate the effect of attacks and has a better performance than the conventional communications system. While performing black-box attacks in Figure 5(b), our system shows a considerable defense capacity, where the performance of our system significantly outperforms the conventional one, which is very close to the performance under no attacks.

D. Proposed Approach versus Autoencoder End-to-End Communications System

Next, we compare our proposed system with the autoencoder end-to-end system that uses regular training and adversarial training, respectively. Regular training means that we train the autoencoder end-to-end system using clean inputs. Adversarial training means that we training the autoencoder system with both the clean inputs and the inputs with adversarial perturbations. For the results of white-box attacks shown in Figure 6(a), we can see that the regular training based autoencoder system has no capability to defend against white-box attacks, where the regular training has the highest error rate. The adversarial training based autoencoder system achieves successful defense against white-box attacks, which obtains large performance improvements compared with the
regular training based autoencoder system. The adversarial training is effective for defending against white-box perturbations because it augments the training data with the same perturbations beforehand. Our proposed system also achieves a good performance similar to the adversarial training, indicating a good defense against white-box attacks. Notably, adversarial training causes considerable performance degradation when there is no attack, indicating that adversarial training degrades the generalization ability of the autoencoder. In contrast, our proposed system still remains in a good performance under no attacks with a strong generalization ability. For the results of black-box attacks shown in Figure 6(b), our proposed system still shows a good defensive ability against black-box perturbations, but the adversarial training leads to a defense failure where the adversarial training has a high error rate. This is because the perturbations used for black-box attacking are different from the perturbations used in the adversarial training. Adversarial training does not work well for unknown perturbations. In contrast, our system can defend against various unknown perturbations. Similarly, Figure 6(b) also indicates that the adversarial training shows a performance degradation under no attacks while our system shows good generalization performance.

V. CONCLUSIONS

This paper presents a novel GAN-based defense approach for end-to-end learning of communications systems, which uses a generative network to model powerful adversarial perturbations and jointly train the end-to-end communications system against the generative attack network. Our approach can learn an end-to-end communications system robust to various adversarial perturbations including both white-box and black-box attacks, but without hurting the generalization performance of the system. In evaluation results, our GAN-based communications system shows better performance and defense capability than the classical communications scheme and the end-to-end communications system with regular training or adversarial training.

REFERENCES

[1] T. O’Shea and J. Hoydis, “An introduction to deep learning for the physical layer,” IEEE Transactions on Cognitive Communications and Networking, vol. 3, no. 4, pp. 563–575, 2017.
[2] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. Deep learning. MIT Press Cambridge, 2016, vol. 1.
[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.
[4] G. Ian, S. Jonathon, and S. Christian, “Explaining and harnessing adversarial examples,” in International Conference on Learning Representations, 2015.
[5] M. Sadeghi and E. G. Larsson, “Physical adversarial attacks against end-to-end autoencoder communication systems,” IEEE Communications Letters, vol. 23, no. 5, pp. 847–850, 2019.
[6] ———, “Adversarial attacks on deep-learning based radio signal classification,” IEEE Wireless Communications Letters, vol. 8, no. 1, pp. 213–216, 2018.
[7] B. Kim, Y. E. Sagduyu, K. Davasioglu, T. Erpek, and S. Ulukus, “Over-the-air adversarial attacks on deep learning based modulation classifier over wireless channels,” in 2020 54th Annual Conference on Information Sciences and Systems (CISS). IEEE, 2020, pp. 1–6.
[8] Q. Liu, J. Guo, C.-K. Wen, and S. Jin, “Adversarial attack on dl-based massive mimo csi feedback,” Journal of Communications and Networks, vol. 22, pp. 230–235, 2020.
[9] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” in ICLR, 2017.
[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in International Conference on Learning Representations, 2018.
[11] A. Raghunathan, S. M. Xie, F. Yang, J. C. Duchi, and P. Liang, “Adversarial training can hurt generalization,” in International Conference on Learning Representations, 2019.
[12] P. Samangouei, M. Kabkab, and R. Chellappa, “Defense-gan: Protecting classifiers against adversarial attacks using generative models,” in International Conference on Learning Representations, 2018.
[13] H. Wang and C.-N. Yu, “A direct approach to robust deep learning using adversarial networks,” in International Conference on Learning Representations, 2019.
[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in Neural Information Processing Systems, vol. 27. Curran Associates, Inc., 2014, pp. 2672–2680.
[15] Z. Qin, H. Ye, G. Y. Li, and B.-H. F. Juang, “Deep learning in physical layer communications,” IEEE Wireless Communications, vol. 26, no. 2, pp. 93–99, 2019.
[16] 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, vol. 30, no. 9, pp. 2805–2824, 2019.
[17] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in International Conference on Learning Representations, 2014.
[18] N. Papernot, P. McDaniel, and I. Goodfellow, “Transferability in machine learning: from phenomena to black-box attacks using adversarial samples,” arXiv preprint arXiv:1605.07277, 2016.
[19] L. Mescheder, S. Nowozin, and A. Geiger, “The numerics of gans,” in Advances in Neural Information Processing Systems, 2017, pp. 1825–1835.