Towards Defending Adversarial Attacks with Temperature Regularization in Automatic Modulation Recognition

Tao Jiang[1,3], Huabao Xu[2], Linlin Liang[2], and Peihan Qi[1(✉)]

[1] State Key Laboratory of ISN, Xidian University, Xi'an 710071, China
phqi@xidian.edu.cn
[2] School of Cyber Engineering, Xidian University, Xi'an 710126, China
huabaoxu@stu.xidian.edu.cn, llliang@xidian.edu.cn
[3] Guangxi Key Laboratory of Cryptography and Information Security, Guilin, China

Abstract. Deep learning has been shown to perform extremely well at various machine learning tasks. However, these same architectures are highly vulnerable to adversarial examples: malicious inputs carefully crafted by adversaries which can force a neural network to produce erroneous predictions with high confidence. This undermines the security of deep learning algorithms when apply to those security-sensitive applications. Existing works have shown that the Signal Modulation Recognition (SMR) solutions based on deep learning are also susceptible to adversarial attacks. In this paper, we propose a new approach called temperature regularization to defense a deep learning scheme against white-box attacks in signal modulation recognition. Specifically, we introduce different temperatures to the softmax layer during the training of the neural network. Experimental results show that training a neural network with an appropriate high temperature can significantly enhance its robustness to three white-box attacks.

Keywords: Deep learning · Automatic modulation recognition · Adversarial defense · Temperature regularization

1 Introduction

With the rapid growth of the end-devices and wide deployment of wireless communication technology, the scarcity of spectrum resources becomes more and more severe and consequently reduces the network availability. However, a large portion of the assigned spectrum is used sporadically and geographically. The survey [10] shows that the utilization of licensed bands from 0 to 6 GHz is less than 6%. Moreover, according to the Federal Communications Commission (FCC) [8], temporal and geographical variations in the utilization of the assigned spectrum range from 15% to 85%, or even lower, thus wasting a lot of spectrum resources. Because of this, the FCC proposed a new concept, namely Cognitive Radio (CR), to use the spectrum opportunistically and overcome the

Y. Chenggang et al. (Eds.): MobiMedia 2022, LNICST 451, pp. 350–361, 2022.
https://doi.org/10.1007/978-3-031-23902-1_27

problem of low spectrum utilization [8]. CR is an intelligent wireless communication system, which builds upon software-defined radio technology and is aware of its operating environment. It helps to learn and readjust the transmission parameters dynamically based on the statistical variations in the environment. CR has been proved to be effective in using the spectrum holes without causing harmful interference to the primary users while maintaining a good quality of service.

Modulation recognition plays a key role in most intelligent communication systems and is considered as a major task of CR in both civilian and military systems. Generally, the existing wireless signal modulation recognition algorithms are mainly implemented in two approaches, i.e., the maximum likelihood method based on hypothesis testing and the pattern recognition method based on feature extraction. Since the later approach has much lower computational complexity and is applicable to real-time applications, it is considered as a promising alternative of the previous one [13]. Deep learning is one of the typical pattern recognition methods with impressive performance. However, Szegedy et al. [26] has found that Deep Neural Networks (DNNs) are highly vulnerable to adversarial examples. By applying well-crafted but subtle perturbations to natural examples, it is easy to fool a pre-trained model to produce erroneous predictions with high confidence. Thus, if a DNN is deployed in adversarial environment to perform automatic modulation recognition task, the normal modulation recognition may not be implemented.

In this paper, inspired by the defensive distillation [23], we propose a new method, called temperature regularization, to defend the DNN against adversarial attacks. Our empirical findings show that training a DNN with an appropriate high temperature can substantially improve its robustness to three white-box attacks, i.e., FGSM [9], MIM [7] and PGD [17].

Contributions. In this paper, we make the following contributions:

- We provide an intuition of why training the DNN with an appropriate high temperature *without distillation* may enhance its robustness to adversarial attacks.
- We analyze how temperature influences the DNN's sensitivity to subtle variations around the inputs from the prospect of the Jacobian matrix.
- We empirically prove the feasibility of the proposed method on modulation signal dataset and identify an "optimal" temperature for defending the VT-CNN2 model in modulation recognition.

2 Preliminaries

Szegedy et al. [26] first noticed the existence of adversarial examples in the field of computer vision and they proposed the L-BFGS algorithm for crafting adversarial examples. After that, various attack algorithms have been proposed for generating adversarial examples, such as FGSM [9], MIM [7], PGD [17], DeepFool [19], CW [4] and so on. Surprisingly, Moosavi-Dezfooli et al. [18] showed

the existence of a universal and quasi-imperceptible perturbation that causes most natural images to be misclassified with high probability. And they proposed an iterative algorithm for generating such universal perturbations. [12,24] even demonstrated that it is feasible to learn the universal perturbations with a generative model. More troubling, adversarial examples misclassified by one model are often misclassified by another model, even if the two models have different network architectures or were trained on disjoint training sets, so long as both models were trained to perform the same task [28–30].

More recently, Lin et al. [16] showed that the classification accuracy of the modulation recognition model, VT-CNN2 [21], could be decreased by about 50% on average when adding a perturbation level of 0.001 to the natural modulation signals. Thus, when applying deep learning models to modulation recognition in adversarial settings, one must take into account certain vulnerabilities.

In the literature, various defense methods have been proposed to reduce the effects of adversarial examples. Typically, these defense methods can be divided into three categories. The first class is to detect the presence of adversarial perturbations in the input during inference. This is usually done by finding statistical outliers or training separate sub-networks that can distinguish between adversarial and benign inputs [6,14,15]. The second class denotes the various pre-processing methods which aim at removing or destroying structured perturbations on adversarial inputs before passing them to the classifier, including [1,11,25] and so on. This class of defenses can be easily used in conjunction with other defense mechanisms and is more practical due to its model- and attack-agnostic property. The third class is to enhance the robustness of neural networks itself, including Adversarial Training [3,27], Label Smoothing [5] and Defensive Distillation [23].

3 Methodology

3.1 Temperature Regularization

Defensive distillation was proposed in [23] to reduce the effectiveness of adversarial examples on DNNs. Their intuition is that knowledge extracted from teacher neural networks, in the form of probability vectors, and transferred in student neural networks can be beneficial to improving generalization capabilities of DNNs to unseen inputs and therefore enhances their robustness to adversarial examples. The authors also pointed out that an ideal training procedure would result in distilled network F^d converging to the original network F although this is not the case empirically.

We believe that *the distilled network F^d can learn what the original network F can, and vice versa, because their network architectures are the same.* Thus, different from the defensive distillation [23] we discard the distilled network training procedure and train the original network with high temperature. The specific analysis is as follow.

(1) Since softmax operation is to normalizes the logits into a probability vector $F(x)$, each component in $F(x)$ indicates the probability that current input x belongs to the corresponding class. For N classes classification problem (i.e. the output dimension of F is N), the output vector $F(x)$ of the last softmax layer can be expressed as

$$F(x) = \left[\frac{e^{z_i(x)/T}}{\sum_{n=0}^{N-1} e^{z_n(x)/T}} \right]_{i \in \{0 \ldots N-1\}} \tag{1}$$

where $z_i(x), i \in 0 \ldots N-1$ are the logits produced by the last hidden layer of a DNN, and parameter T is called temperature. As $T \to \infty$ the $e^{z_i(x)/T}$ converge to 1. Thus, all components in $F(x)$ are close to $1/N$. From the above analysis, we can see that there exists T_0, when $T \geq T_0$ the model's prediction probability assigning to all classes not greater than $p \in (1/N, 1)$ for all inputs. Thus, training a DNN with a high temperature will force it not to make overly confident predictions in any one class examples and reduce its sensitivity to small variations of its inputs [23]. In addition, from the perspective of the loss function, the higher the temperature is, the more ambiguous its probability distribution will be (i.e. all probabilities of the $F(x)$ are close to $1/N$). Therefore, the more stable the loss function will be (i.e. the cross-entroy for any input is close to $\log N$). This consequently can make the learned model smoother. Note that this change of temperature does not impact the relative ordering of classes.

(2) The neural network's sensitivity to input variations can be quantified by its Jacobian matrix. Let M denote the input dimension, we now consider one element $(i, j) \in [0..N-1] \times [0..M-1]$ of the $N \times M$ Jacobian matrix for a neural network F at temperature T:

$$\left. \frac{\partial F_i(x)}{\partial x_j} \right|_T = \frac{\partial}{\partial x_j} \left(\frac{e^{z_i(x)/T}}{\sum_{n=0}^{N-1} e^{z_n(x)/T}} \right) \tag{2}$$

Let $h(x) = \sum_{n=0}^{N-1} e^{z_n(x)/T}$, we then have:

$$\begin{aligned} \left. \frac{\partial F_i(x)}{\partial x_j} \right|_T &= \frac{\partial}{\partial x_j} \left(\frac{e^{z_i(x)/T}}{h(x)} \right) \\ &= \frac{1}{h^2(x)} \left(\frac{\partial e^{z_i(x)/T}}{\partial x_j} h(x) - e^{z_i(x)/T} \frac{\partial h(x)}{\partial x_j} \right) \\ &= \frac{1}{h^2(x)} \frac{e^{z_i(x)/T}}{T} \left(\sum_{n=0}^{N-1} e^{z_n(x)/T} \frac{\partial z_i(x)}{\partial x_j} - \sum_{n=0}^{N-1} e^{z_n(x)/T} \frac{\partial z_n(x)}{\partial x_j} \right) \\ &= \frac{1}{T} \frac{e^{z_i(x)/T}}{h^2(x)} \left[\sum_{n=0}^{N-1} e^{z_n(x)/T} \left(\frac{\partial z_i(x)}{\partial x_j} - \frac{\partial z_n(x)}{\partial x_j} \right) \right] \end{aligned} \tag{3}$$

Equation 3 shows that increasing the temperature T will reduce the absolute value of all elements of model F's Jacobian matrix when values of the logits

$z_0(x), ..., z_{N-1}(x)$ are fixed. Thus, using a high temperature during training will systematically reduce the model sensitivity to small perturbations of its inputs at test time.

From the above analysis, we can derive that choosing an appropriate high temperature T for training will improve the robustness of a DNN to adversarial attacks and maintain its performance on benign examples at the same time (Fig. 1).

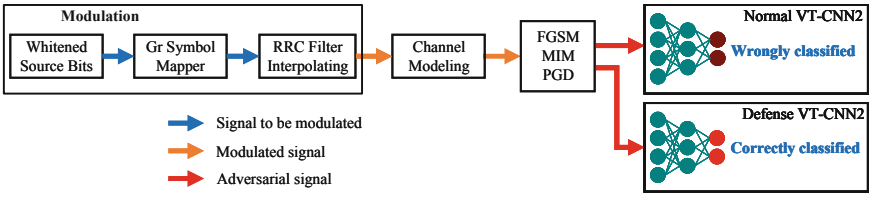


Fig. 1. An overview of defending the VT-CNN2 model against white-box attacks in modulation recognition. The only difference between the Normal VT-CNN2 and Defense VT-CNN2 is the softmax layer, where Defense VT-CNN2 has a much higher temperature than the Normal VT-CNN2.

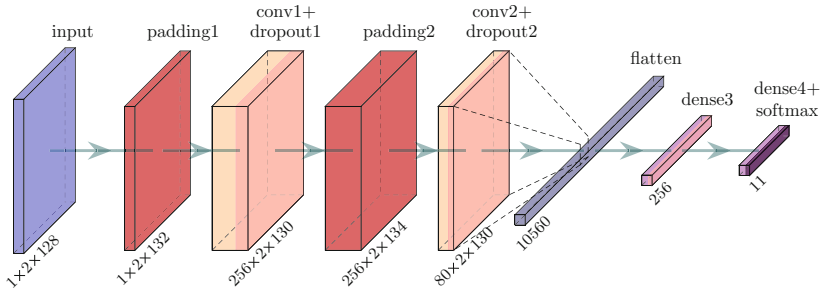


Fig. 2. Structure of VT-CNN2

3.2 Datasets and Target Model

To have a thorough study of our defense technique in the field of wireless communications, we choose the open-source dataset RADIOML 2016.10A generated with GNU Radio [2]. This dataset consists of 11 modulation signals at varying signal-to-noise ratios, including eight kinds of digital signals: 8PSK, BPSK, CPFSK, GFSK, PAM4, QAM16, QAM64, and QPSK, and three kinds of analog signals: AM-DSB, AM-SSB and WBFM. It has 220,000 data samples totally with 20 kinds of SNRs range from −20 dB to 18 dB in steps of 2 dB. Each SNR

category has 11,000 data samples. We use 8 800 samples for training and the remaining 2,200 samples for testing at each SNR. For target model, we use the VT-CNN2 [20,21] which is optimized for the dataset RADIOML 2016.10A [2] to perform the modulation classification task. The network depth of VT-CNN2 is roughly equivalent to neural networks which work well on similar simple datasets in the field of computer vision. The specific architecture of VT-CNN2 is shown in Fig. 2.

3.3 Implementation

All the experiments in this work were performed on an NVIDIA GeForce RTX 2080 Ti and the code for the experiments was written in Python 3 using Tensorflow 2.4.1. An Adam solver and a categorical-crossentropy loss function were used to train the target model VT-CNN2 with a batch size of 1024. The three white-box attacks, FGSM, MIM, and PGD were implemented by an open-source software library, i.e., Cleverhans [22]. After the training converges, we obtain ≈72% and ≈38% test accuracy on benign examples under SNR = 10 dB and SNR = −10 dB respectively when there is no attack and no defense.

4 Experiments

4.1 Temperature Parameter Space Exploration

Firstly, we measure how temperature improves the resilience of the VT-CNN2 to adversarial attacks. Since the softmax mainly normalizes the logits into a probability vector $F(x)$ and does not change the relative ordering of classes, the parameter temperature T only matters during training. In order to produce more discrete distributions of probabilities, the softmax temperature is set back to 1 at test time.

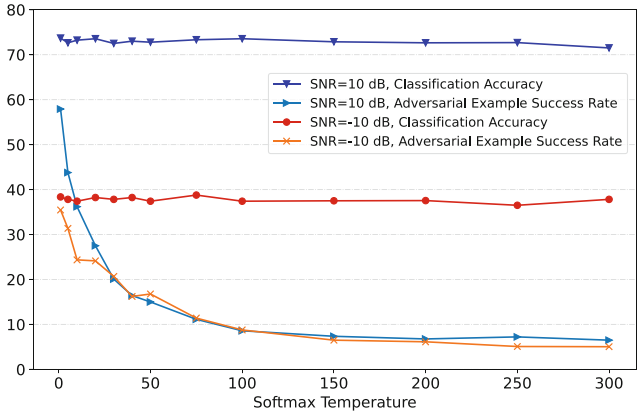


Fig. 3. An exploration of the temperature parameter space. We plot the VT-CNN2's classification accuracy on test set and the success rate of adversarial examples generated with PGD attack. Note that perturbations norm $\varepsilon = 0.0015$.

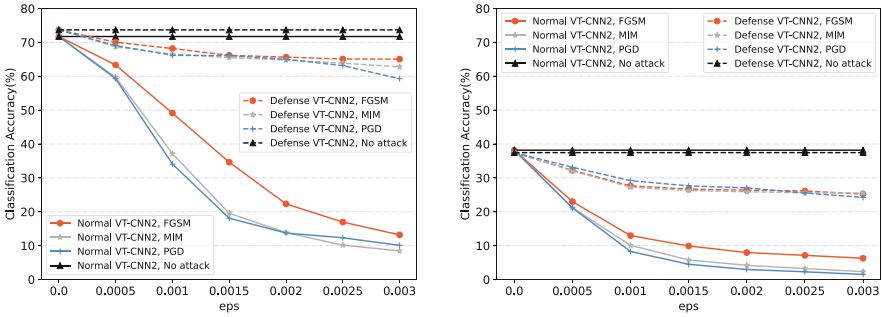
To identify the "optimal" softmax temperature, we train the VT-CNN2 under SNR = 10 dB and SNR = −10 dB with different temperatures {T: 1, 5, 10, 20, 30, 50, 75, 100, 150, 200, 300}. Then we use the attack algorithm PGD to generate adversarial examples. The classification accuracy of benign examples and the success rate of adversarial examples with respect to temperature is shown in Fig. 3. We observe that with the increase of temperature, the adversarial example success rate decreases rapidly and then largely remains constant. Specifically, when the temperature increases from 1 to 100, the adversarial example success rate decreased from 57.91% and 35.45% to 8.59% and 8.77% under SNR = 10 dB and SNR = −10 dB respectively. Based on the above observations, we choose temperature $T = 100$ to train the VT-CNN2 to effectively defense against adversarial attacks.

4.2 Defending Against Adversarial Attacks

To analyze the performance of our defense method, we conduct a series of comparative experiments. In the following sections, we refer to the model without defense as the Normal VT-CNN2 while the model with defense (trained with an appropriate high temperature) as the Defense VT-CNN2. Note that we do not care about the relationship between these different attack methods, but just how our Defense VT-CNN2 behaves under attacks.

Under Different Perturbations Magnitude. Figure 4 shows the classification accuracy of Normal VT-CNN2 and Defense VT-CNN2 under three attack algorithms (i.e., FGSM, MIM, and PGD) with different perturbations magnitude ε at SNR = 10 dB and SNR = −10 dB. We can observe that the proposed defense method can improve the classification accuracy of VT-CNN2 on adversarial examples largely, especially those generated with iterative-step attack algorithms (i.e. MIM and PGD). Concretely, at SNR = 10 dB, the classification accuracy of Defense VT-CNN2 on iterative adversarial examples is about three times of that Normal VT-CNN2 when the perturbations magnitude $\varepsilon > 0.001$. Furthermore, this defense technique has almost no influence on the accuracy of VT-CNN2 on benign examples. In some cases, for example, when SNR = 10 dB, the Defense VT-CNN2 even have higher accuracy than the Normal VT-CNN2.

To get more insight into the defense method further, we plot the confusion matrix of Normal VT-CNN2 and Defense VT-CNN2 for all 11 categories at SNR = 10 dB, as is shown in Fig. 5. It can be seen from Fig. 5a that when there is no attack and no defense, the Normal VT-CNN2 has serious classification confusion between analog signals 8PSK and QPSK, AM-DSB and WBFM, and between digital signals QAM16 and QAM64. In [16], they claimed that analog modulation confusion is difficult to solve, but the QAMs confusion can be improved by better synchronization and reducing channel impairments. Figure 5b shows the confusion matrix of Normal VT-CNN2 under adversarial attack. As can be seen, the MIM attack has a strong negative effect on the classification accuracy and the confusion matrix is extremely chaotic. Particularly, BPSK and PAM4 are



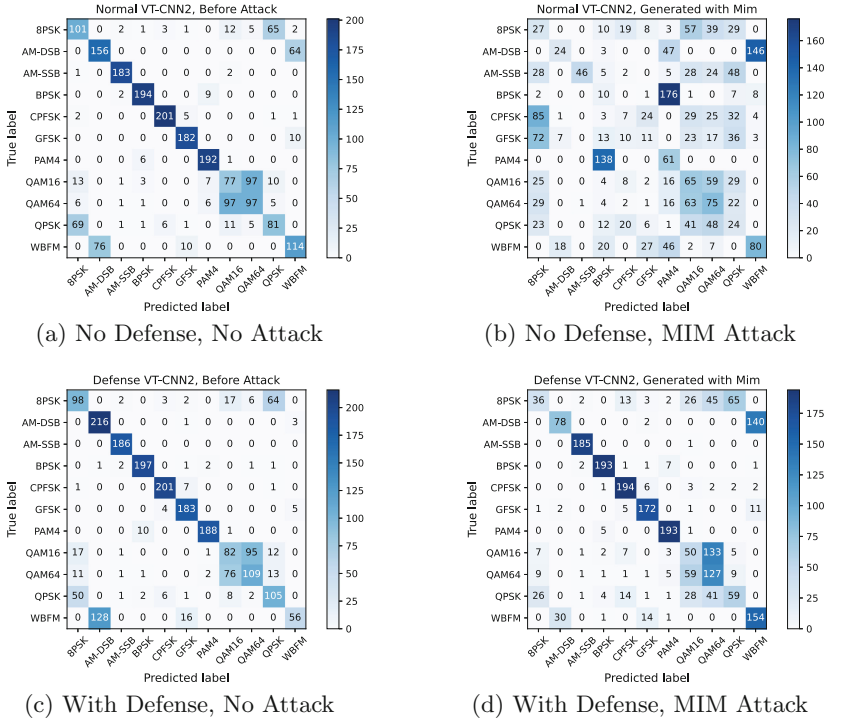
(a) SNR = 10 dB (b) SNR = -10 dB

Fig. 4. Classification accuracy of VT-CNN2 with different perturbations magnitude ε.

misclassified to each other with a higher probability. And CPFSK and GFSK both are more likely misclassified as 8PSK. At the same time, the AM-DSB has a great possibility misclassified as WBFM. That is, compared to other modulation signals, FSKs, PSKs and AM-DSB signals are more vulnerable to adversarial attacks.

Figure 5c shows the confusion matrix of Defense VT-CNN2 before attack. Compare Fig. 5c with Fig. 5a, we notice that the defense method has some impact on the accuracy of 8PSK and WBFM benign examples. But overall, this impact is negligible. Figure 5d shows the confusion matrix of Defense VT-CNN2 under attack. Compare Fig. 5d with Fig. 5b, it can be clearly seen that the defense method can mitigate the adversarial effect significantly with the exception of QAM16. We do not exactly know why the accuracy of Defense VT-CNN2 on QAM16 is lower than that of Normal VT-CNN2 under attack, but this may be caused by the low accuracy of Normal VT-CNN2 on the QAM16 benign examples. Because when the classification accuracy is low, attack or defense does not make much sense.

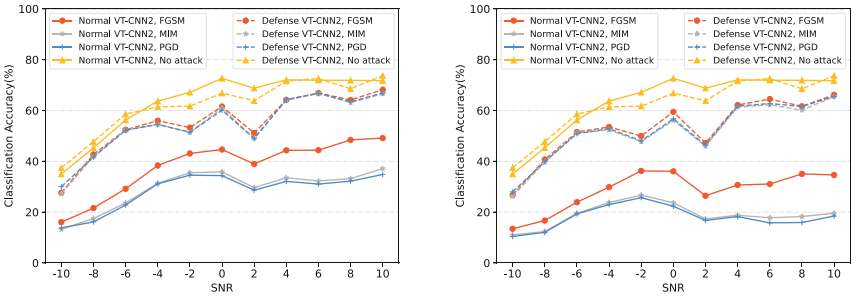
Under Different SNRs. Figure 6 shows the classification accuracy of Normal VT-CNN2 and Defense VT-CNN2 under three white-box attacks at different SNRs, where $\varepsilon = 0.001$ for subplot Fig. 6a and $\varepsilon = 0.0015$ for subplot Fig. 6b. As shown in Fig. 6a and 6b, when there is no attack, the classification accuracy of Normal VT-CNN2 exhibits an increasing trend first and then almost remains constant ($\approx$72%) with the increase of SNR. Specifically, the classification accuracy of Normal VT-CNN2 increases rapidly from 38% to 72% as the SNR increases from -10 dB to 0 dB. But under adversarial attacks, the classification accuracy of Normal VT-CNN2 drops by almost one half at all SNRs. Obviously, the strength of iterative-step attacks is stronger than that of one-step attack. We find that our defense method is very effective against these adversarial attacks, especially against the iterative-step attacks. In detail, the classification accuracy of Defense VT-CNN2 under the three white-box attacks decreases by



(a) No Defense, No Attack (b) No Defense, MIM Attack

(c) With Defense, No Attack (d) With Defense, MIM Attack

Fig. 5. Confusion Matrix of VT-CNN2 in different scenarios. Note the SNR = 10 dB and perturbations magnitude $\varepsilon = 0.0015$.

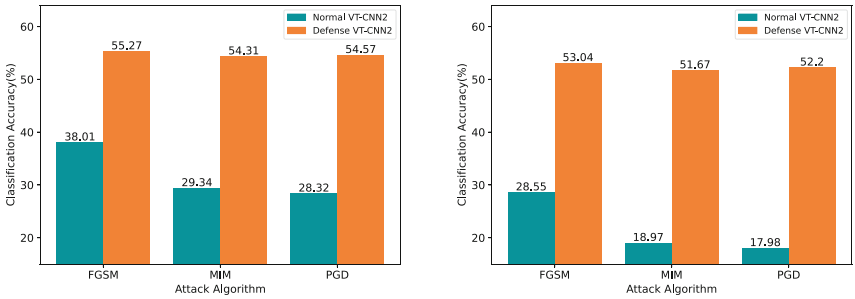


(a) $\varepsilon = 0.001$ (b) $\varepsilon = 0.0015$

Fig. 6. Classification accuracy of VT-CNN2 at different SNRs.

no more than 10% at almost all SNRs compared to the case where there is no attacks.

To further quantify the defense effect of the proposed approach against the adversarial attack algorithms, we calculated the average classification accuracy of the Normal VT-CNN2 and the Defense VT-CNN2 on adversarial examples under



(a) $\varepsilon = 0.001$ (b) $\varepsilon = 0.0015$

Fig. 7. Average accuracy of VT-CNN2 under all SNRs.

all SNRs. The results are shown in Fig. 7. It can be seen clearly that our Defense VT-CNN2 is extremely robust to FGSM, MIM, and PGD attacks. Compare to the Normal VT-CNN2, the classification accuracy of the Defense VT-CNN2 is improved by about 17% for FGSM and by nearly 25% for MIM and PGD when perturbations magnitude $\varepsilon = 0.001$. With the increase of perturbations magnitude ε, this defense effect is more significant. As shown in Fig. 7b, when perturbations magnitude $\varepsilon = 0.0015$, the classification accuracy of the Defense VT-CNN2 is improved by about 34% under the iterative-step attacks (i.e. MIM and PGD).

This results show that the Normal VT-CNN2 is more sensitive to the perturbations magnitude ε of the attack algorithms, since it's average classification accuracy decreased by almost 10% when the perturbations magnitude ε increased from 0.001 to 0.0015 while the classification accuracy of Defense VT-CNN2 nearly remains unchanged. It is worth noting that, although the iterative-step attacks are stronger, our Defense VT-CNN2 achieves almost the same accuracy under one-step attack (FGSM) and iterative-step attacks (MIM and PGD), which is about 52%.

5 Conclusion

In this work, we evaluated the security problems on automatic modulation recognition based on deep learning and demonstrated the efficacy of using an appropriate high temperature within the softmax layer as a defense method against adversarial attacks. We analyzed the reason why this defense method works and determined an "optimal" temperature for the VT-CNN2 model through experiment. Our empirical findings showed that using an appropriate high temperature during training can significantly reduce the successfulness of adversarial attacks against automatic modulation recognition classifiers. Moreover, it nearly maintains the accuracy rates of VT-CNN2 on benign examples. Lastly, this defense technique is extremely easy to implement and introduces no overhead for training after the "optimal" temperature is chosen.

Acknowledgment. This work was supported by National Natural Science Foundation of China (Nos. 62171334, 62001359), Fundamental Research Funds for the Central Universities (No. XJS211502) and Guangxi Key Laboratory of Cryptography and Information Security (No. GCIS201716).

References

1. Yin, S.L., Zhang, X.L., Zuo, L.Y.: Defending against adversarial attacks using spherical sampling-based variational auto-encoder. Neurocomputing **478**, 1–10 (2022)
2. N.P., et al.: Deepsig dataset: RadioML 2016.10a (2016). https://www.deepsig.io/datasets
3. Allen-Zhu, Z., Li, Y.: Feature purification: how adversarial training performs robust deep learning. In: 62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, 7–10 February 2022, pp. 977–988. IEEE (2021)
4. Carlini, N., Wagner, D.A.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, 22–26 May 2017, pp. 39–57. IEEE Computer Society (2017)
5. Warde-Farley, D., Goodfellow, I., Hazan, T., Papandreou, G., Tarlow., D.: Adversarial perturbations of deep neural networks, pp. 311–342 (2017)
6. Dong, J., Zhou, P.: Detecting adversarial examples utilizing pixel value diversity. In: Asian Hardware Oriented Security and Trust Symposium, AsianHOST 2021, Shanghai, China, 16–18 December 2021. pp. 1–6. IEEE (2021)
7. Dong, Y., et al.: Boosting adversarial attacks with momentum. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 9185–9193. Computer Vision Foundation/IEEE Computer Society (2018)
8. FCC: Notice of proposed rule making and order. ET, Docket No 03-222 (2003)
9. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015)
10. Gui, G., Huang, H., Song, Y., Sari, H.: Deep learning for an effective nonorthogonal multiple access scheme. IEEE Trans. Veh. Technol. **67**(9), 8440–8450 (2018)
11. Guo, C., Rana, M., Cissé, M., van der Maaten, L.: Countering adversarial images using input transformations. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018, Conference Track Proceedings. OpenReview.net (2018)
12. Hashemi, A.S., Bär, A., Mozaffari, S., Fingscheidt, T.: Transferable universal adversarial perturbations using generative models. CoRR abs/2010.14919 (2020)
13. Jdid, B., Hassan, K., Dayoub, I., Lim, W.H., Mokayef, M.: Machine learning based automatic modulation recognition for wireless communications: a comprehensive survey. IEEE Access **9**, 57851–57873 (2021)
14. Li, Y., Tang, T., Hsieh, C., Lee, T.C.M.: Detecting adversarial examples with Bayesian neural network. CoRR abs/2105.08620 (2021)
15. Liang, B., Li, H., Su, M., Li, X., Shi, W., Wang, X.: Detecting adversarial image examples in deep neural networks with adaptive noise reduction. IEEE Trans. Dependable Secur. Comput. **18**(1), 72–85 (2021)

16. Lin, Y., Zhao, H., Ma, X., Tu, Y., Wang, M.: Adversarial attacks in modulation recognition with convolutional neural networks. IEEE Trans. Reliab. **70**(1), 389–401 (2021)
17. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018, Conference Track Proceedings. OpenReview.net (2018)
18. Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. CoRR abs/1610.08401 (2016)
19. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.: DeepFool: a simple and accurate method to fool deep neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 2574–2582. IEEE Computer Society (2016)
20. O'Shea, T.J., West, N.: Radio machine learning dataset generation with gnu radio (2016)
21. O'Shea, T.J., Corgan, J., Clancy, T.C.: Convolutional radio modulation recognition networks. In: Jayne, C., Iliadis, L. (eds.) EANN 2016. CCIS, vol. 629, pp. 213–226. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44188-7_16
22. Papernot, N., Faghri, F., Carlini, N., et al.: Technical report on the CleverHans v2.1.0 adversarial examples library. arXiv preprint arXiv:1610.00768 (2018)
23. Papernot, N., McDaniel, P.D., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, 22–26 May 2016, pp. 582–597. IEEE Computer Society (2016)
24. Poursaeed, O., Katsman, I., Gao, B., Belongie, S.J.: Generative adversarial perturbations. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 4422–4431. Computer Vision Foundation/IEEE Computer Society (2018)
25. Samangouei, P., Kabkab, M., Chellappa, R.: Defense-GAN: protecting classifiers against adversarial attacks using generative models. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018, Conference Track Proceedings. OpenReview.net (2018)
26. Szegedy, C., et al.: Intriguing properties of neural networks. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014, Conference Track Proceedings (2014)
27. Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., McDaniel, P.D.: Ensemble adversarial training: attacks and defenses. CoRR abs/1705.07204 (2017)
28. Wang, X., He, X., Wang, J., He, K.: Admix: enhancing the transferability of adversarial attacks. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 10–17 October 2021, pp. 16138–16147. IEEE (2021)
29. Wu, W., Su, Y., Lyu, M.R., King, I.: Improving the transferability of adversarial samples with adversarial transformations. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, 19–25 June 2021, pp. 9024–9033. Computer Vision Foundation/IEEE (2021)
30. Zhong, Y., Deng, W.: Towards transferable adversarial attack against deep face recognition. IEEE Trans. Inf. Forensics Secur. **16**, 1452–1466 (2021)