A Robust Signal Modulation Recognition Method Against Black-Box Detection Attack

Zhihui An, Peihan Qi(✉), Xiaoyu Zhou, and Yongchao Meng

State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China
phqi@xidian.edu.cn

Abstract. Deep learning (DL) models have been widely used in the recognition of modulation types with outstanding recognition effects. With the improvement of modulation recognition, the perturbation from the attacker has also changed from adding physical interference to the original signal to an adversarial attack based on the neural network. The adversarial attack adding subtle perturbation which is imperceptible to the human eye, makes the neural network produce false recognition results with high confidence. This kind of perturbation is hard to be reflected in spectrogram or constellation diagram, so it is seriously destructive to the modulation recognition algorithm based on neural networks. In response to adversarial attacks, we propose a modulation recognition method against black-box detection attacks. In this paper knowledge distillation is used to defend against the attack that comes from the attacker's black-box detection. The experimental results demonstrated that the defense method constructed in this paper can improve the ability to defend adversarial samples and keep the recognition accuracy of the recognition network. This article aims at improving the robustness of the network and constructing a robust modulation recognition network.

Keywords: Modulation recognition · Adversarial attack · Knowledge distillation · Robust network

1 Introduction

With the continuous development of modern communication systems, the signal environment is becoming more and more complex. In such a complex scenario, human comprehension ability can neither complete the perception of large amounts of data, nor can its comprehension speed match the update speed of information, which can easily cause a series of key goals and situation perception errors or untimely perception problem.

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62171334.

Y. Chenggang et al. (Eds.): MobiMedia 2022, LNICST 451, pp. 329–339, 2022.
https://doi.org/10.1007/978-3-031-23902-1_25

With the rapid development of artificial intelligence in recent years, deep learning as one of its core technologies has been widely used in many fields and has already demonstrated breakthrough progress in signal recognition technology [1]. O'Shea et al. [1] proposed to use the convolutional neural network (CNN) structure for automatic modulation classification of 11 modulation signals and to automatically learn and classify the originally collected signals. Peng et al. [2] proposed a constellation diagram (CD) to transform a time-series IQ signal into an image for adopting AlexNet and GoogLeNet which have a better performance on image classification. N. E. West and T. O'Shea [3] studied the effects of convolution kernel size, network depth, and the number of neurons in each layer of convolution neural network on modulation classification performance, and compared convolution neural network, residual network and convolution-based long short-term memory network (LSTM) for modulation classification. Zhang et al. [4] proposed a CNN-LSTM structure based on a dual-stream structure that can explore the feature interaction and spatial-temporal properties of radio signals.

However, due to the interpretability of neural networks and a large number of non-linear high-dimensional operations, neural networks often make false judgments with high confidence when faced with some subtle perturbation that is hardly detectable by the human eye. In response to this phenomenon, Christian Szegedy et al. [6] proposed the concept of adversarial samples. They used the L-BFGS method to generate adversarial examples to solve general target problems. Goodfellow et al. [7] proposed the Fast Gradient Sign Method (FGSM) which is used to fool CNN-based classification networks. FGSM generates adversarial examples by computing the gradient of the neural network loss function knowledge distillation. Kurakin et al. [8] applied adversarial examples to the physical world. They extended the FGSM by running finer optimizations (smaller changes) for multiple iterations. Papernot et al. [9] proposed Jacobian-based saliency map attack (JSMA). This approach produces adversarial perturbations by forwarding derivatives. Dezfooli et al. [13] proposed the DeepFool algorithm to find the closest distance from the original input to the adversarial sample decision boundary.

Research on defending adversarial samples has achieved results in some fields, such as image classification, speech recognition, target detection, etc. Using adversarial examples for training is one of the strategies to improve the robustness of neural networks. Goodfellow et al. use adversarial examples during training [7]. They will generate adversarial examples at every step of the training and send them into the training set. From the results, adversarial training can improve the robustness of the network, but an expanded training set cannot defend against black-box attacks. Metzen et al. [14] created a detector for adversarial examples as an auxiliary network to the original neural network.

In order to reduce the influence of adversarial samples on the signal modulation recognition network and improve the robustness of the network, this paper proposes a method based on knowledge distillation to update the parameters of the signal modulation recognition network through the teacher model and the

student model to achieve the effect of against black-box detection attack. The research data in this paper is drawn from the public data set RML2016.10a. The student network uses the VTCNN2, and the teacher network uses a 16-layer ResNet network. The results showed that the teacher network has a high recognition accuracy for the adversarial samples. After knowledge distillation and parameter transfer, the recognition accuracy of the signal modulation recognition network in adversarial samples has increased from 0.37 to 0.69 (SNR = 10 dB). It can be seen that the method proposed in this paper has excellent defensive capabilities in defending black-box detection attacks. In addition, the methods mentioned in this paper also have a generalization ability and show good defensive capabilities against adversarial samples generated in different ways.

2 Methodology

2.1 Black-Box Detection Attack

Adversarial attacks can be divided into black-box attacks and white-box attacks. The difference between the two is the degree of knowledge acquisition of the target model. In practical applications, attackers can't directly obtain all the knowledge of the target model but obtain the relevant parameters of the target model by accessing the target model a limited number of times. In this paper, we set up a black-box detection attack, assuming that the attacker has detected the recognition network structure of the target model through previous access, and trained the attack samples on the same recognition network structure. In other words, the attacker knows the network structure of the modulation recognition network and trains the modulation recognition network. The adversarial sample is generated according to the trained modulation recognition network to attack the target model.

In the experiment, the same network structure and training set as the modulation recognition network is used to train the attacker's network, and the FGSM method is used to generate adversarial samples on the attacker's network to simulate the black-box detection attack.

2.2 Defensive Distillation

Knowledge distillation methods are widely used in network pruning and network structure compression. Taking advantage of its property of transferring knowledge from complex networks to simple networks, knowledge distillation is applied to defend against black-box detection attacks. The black-box detection attack considers that the attacker knows the structure of the modulation recognition network, and the attacker achieves the purpose of simulating the network performance by training the attack network with the same structure. On the premise of keeping the structure of the modulation recognition network unchanged, defensive distillation transfers the knowledge of the complex network

to the modulation recognition network through distillation, so that it can achieve the performance of the complex network under the condition of having a simple structure. Thereby preventing black-box detection attacks from attackers.

The teacher network and the student network are also constructed in the defense distillation. The structure of the teacher network is complex and the performance is better, but due to the long training time and the huge structure, it is not easy to deploy. The student network has a simpler structure and is convenient for training. In this paper, the student network adopts the same network structure as the modulation recognition network.

$$q_i = \left[\frac{\exp(Z_i/T)}{\sum_{j=0}^{N-1} \exp(Z_j/T)} \right]_{i \in 0,\cdots,N-1} \tag{1}$$

In the formula, q_i is the current category classification probability, N represents N categories, and Z_i is the logits output of the current category into the activation function. The process of defensive distillation can be described as using the training set and the hard labels to train the teacher network, and after the training is completed, the teacher network is used to label the corresponding soft labels for the training data. The specific method for generating soft labels is to introduce a temperature function T into the softmax layer of the teacher network. When T = 1, it is the original softmax function. The larger T is, the softer the distribution between various classes output by the function is. Compared with hard labels, soft labels can not only classify the data correctly but also reflect the similarity between categories.

2.3 Framework of Teacher Network and Student Network

The structure of the student network and teacher network constructed in the process of knowledge distillation in this paper is shown in Fig. 1 and Fig. 2. We designed a teacher network for modulation recognition, and the student network is VTCNN2.

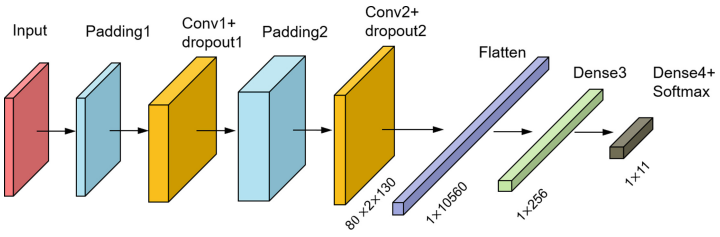


Fig. 1. The structure of the student network (VTCNN2)

The structure of teacher network is designed as follows:

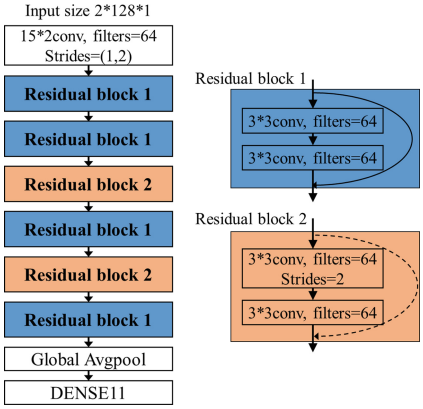


Fig. 2. The structure of the teacher network

The first convolutional layer performs simple preliminary feature extraction on the signal data input to the network and feeds the feature map into the ResNet blocks.

We add ResNet blocks to further extract the features of the signal. ResNet introduces jump connections between different convolutional layers, which can better learn data features than simple convolutional layers.

The function of the fully connected (FC) layer is to output the final recognition result according to the feature map extracted by the ResNet blocks. Here we rewrite the activation function softmax, and introduce the temperature parameter T into it, so as to achieve the effect of generating soft labels by the teacher network.

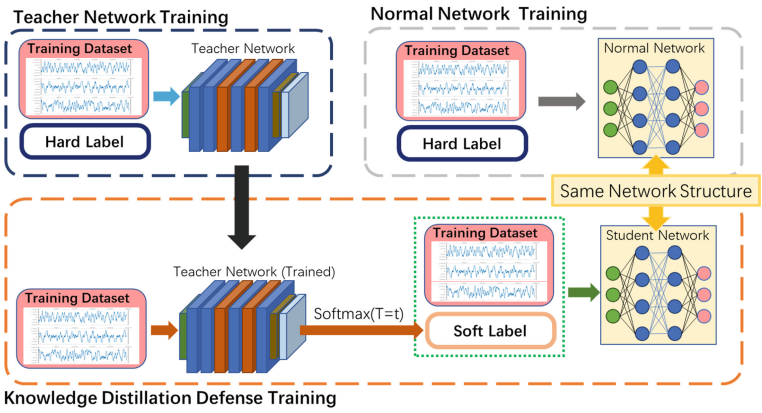


Fig. 3. Knowledge distillation defense training process and normal network training process

The knowledge distillation defense training process and the normal network training process are shown in Fig. 3. The knowledge distillation defense training uses datasets and hard labels to train a teacher network with a complex structure. After the teacher network training is completed, the teacher network is used to generate soft labels. The soft labels are combined with the dataset to retrain the student network. The normal training process is to directly use the data set and hard labels to train the recognition network.

2.4 Datasets

We chose the open-source simulation dataset RADIOML 2016.10A designed by DeepSiG. This dataset was selected because it is publicly available and is based on CNN. RADIOML 2016.10A consists of modulation signals at different SNRs, including eight kinds of digital signals: 8PSK, QPSK, BPSK, GFSK, CPFSK, PAM4, QAM16, and QAM64, and three kinds of analog signals: wide band frequency modulation (WBFM), amplitude modulation-double side band (AM-DSB), and amplitude modulation-single side band (AM-SSB). The dataset generates a total of 220 000 data samples with 20 kinds of SNRs, from 18 to–20 dB in steps of 2 dB, which means 2000 samples for each signal category. We used 80% of the samples as the training set and the rest samples as the test set. Each signal vector consists of an in-phase component and an orthogonal component, and each component has a length of 128.

3 Experiments

To analyze the performance of the proposed defense method, we conduct a series of comparative experiments. In this section, the complex network in the distillation step is called the teacher network, and the simple network model is called the student network. In the comparative experiments, the recognition network without defense training is called a normal network (with the same structure as the student network). We call the samples without adversarial attacks clean samples and the samples with adversarial attacks adversarial samples (FGSM-Adversarial Samples and BIM-Adversarial Samples). The main performance improvement is between the student network trained by the defense method and the normal network. It should be noted that although the teacher network has better performance, due to its complex structure and a large number of parameters, it is not easy to train and deploy. We only care about the improvement of network performance by defense methods under a simple structure.

Before conducting defense training, we need to select a teacher network with excellent performance for knowledge distillation. Using the teacher network model and data set proposed above for training, good recognition accuracy is obtained. Use the temperature parameter adjustment in the teacher network to generate soft labels, and use the soft labels and the original data set to train the student network. Figure 4 shows the recognition effect of the teacher network on

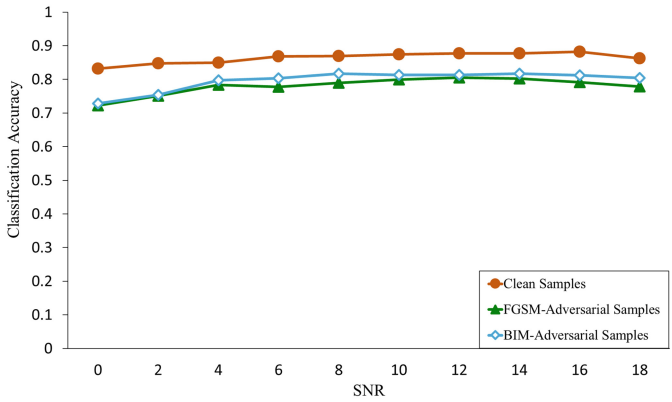


Fig. 4. Recognition accuracy of teacher network on clean samples and adversarial samples

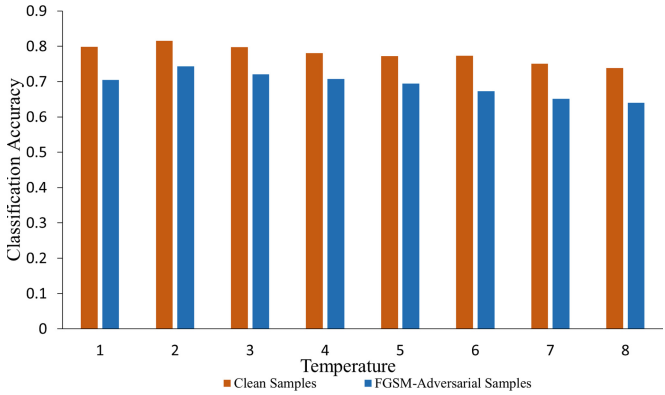


Fig. 5. The recognition accuracy of the student network on clean samples and adversarial samples (SNR = 10 dB) at different temperatures

clean samples and adversarial samples. It can be seen that the teacher network has a good recognition effect on both clean data and adversarial samples.

Next, we determine how changes in the temperature parameter affect the knowledge distillation defense effect. We use 10 dB data to conduct experiments, adjust the parameter T in the teacher network, generate different soft labels to train the student network, and compare the recognition performance of the student network under the soft label training generated by different temperature parameters. It can be seen from Fig. 5 that when the temperature T = 2, the performance of the student network is slightly improved, and as the temperature increases, the performance begins to decline.

So we decided to perform knowledge distillation defense when the temperature parameter T = 2. We used FGSM and BIM to generate adversarial samples,

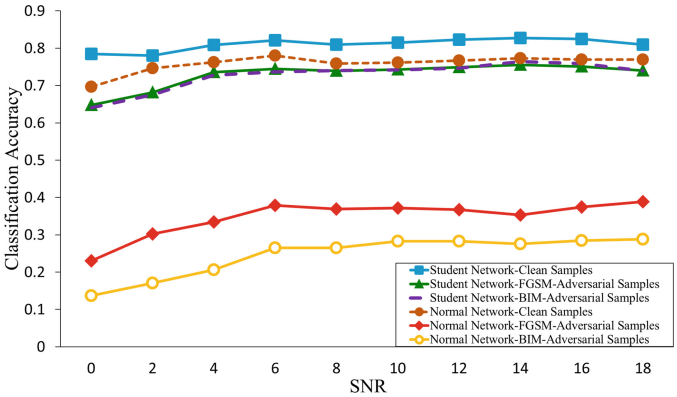
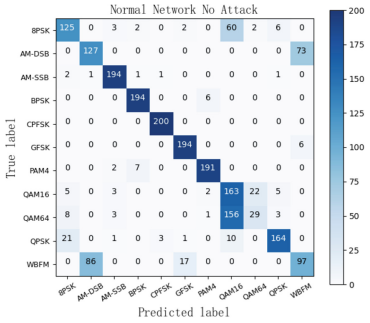


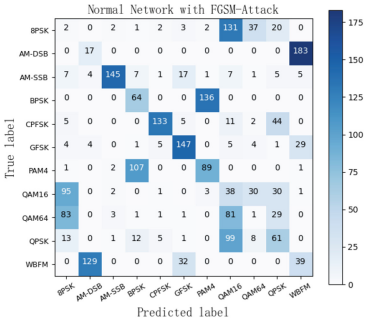
Fig. 6. The performance of student network and normal network on clean samples and adversarial samples

and the perturbation parameter of the adversarial samples was set to 0.001. We compare the performance of the student network trained with knowledge distillation defense and the normal network between 0 dB and 18 dB, mainly through the recognition accuracy rate on clean samples, FGSM adversarial samples, and BIM adversarial samples to show the performance improvement, as shown in Fig. 6, when the student network is attacked by adversarial samples, the recognition performance drops sharply. When the SNR = 4 dB, the recognition accuracy drops from 0.76 to 0.33 (FGSM-Attack) and 0.20 (BIM-Attack). It drops from 0.77 to 0.40 (FGSM-Attack) and 0.29 (BIM-Attack) at 18 dB. After the knowledge distillation defense training, the student network with the same structure as the normal network shows performance improvement on the same test samples. The student network has demonstrated good defense capabilities. When the SNR = 4 dB, the recognition accuracy is improved to 0.73 (FGSM-Attack) and 0.72 (BIM-Attack), and when the SNR = 18 dB, the recognition accuracy is 0.74 (FGSM-Attack) and 0.74 (BIM-Attack). Not only in the defense of adversarial samples, but also on clean samples, the recognition accuracy of the student network is also improved.

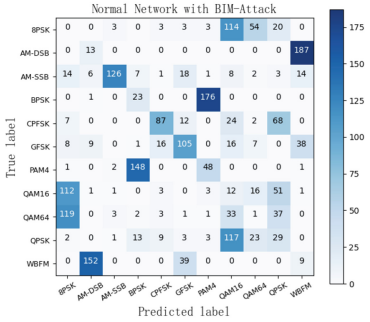
In order to further understand the positive impact of the knowledge distillation defense method on the recognition results, we plot the confusion matrix of 11 recognition results of the student network and the normal network when SNR = 4 dB. As shown in Fig. 7. From Fig. 7a, it can be seen that when not attacked by adversarial examples, the normal network produces serious confusion between analog signals 8PSK and QPSK, AM-DSB and WBFM, and between digital signals QAM16 and QAM64. After the normal network is attacked by adversarial samples, as shown in Figs. 7b and 7c, the confusion matrix is very chaotic, and the classification accuracy of the normal network is seriously affected. And the damage caused by BIM is even more serious. Figure 7d shows the case where the student network is not attacked by adversarial examples, and Fig. 7e and 7f are the confusion matrices of the student network under the attack of adversarial



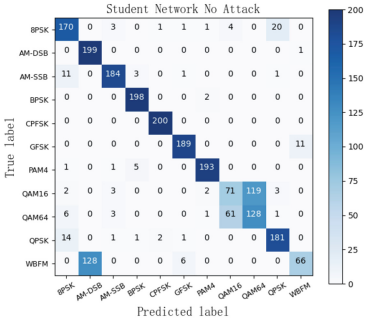
(a) Normal network No Attack



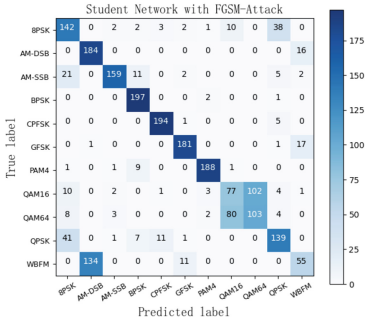
(b) Normal network FGSM-Attack



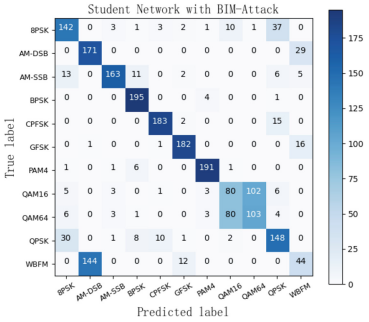
(c) Normal network BIM-Attack



(d) Student network No Attack



(e) Student network FGSM-Attack



(f) Student network BIM-Attack

Fig. 7. Confusion matrix of student network and normal network

examples. It can be seen that when there is no attack, the student network guarantees the classification accuracy with a slight improvement. Except for QAM16 and QAM64, the student network shows a dramatic improvement in classification accuracy when attacked by adversarial examples. Figure 7 confirms that the

knowledge distillation defense method can exhibit a relatively prominent defense capability in the field of electromagnetic signal modulation recognition.

4 Conclusion

In this work, we investigate the defense method of knowledge distillation against adversarial attacks in the field of electromagnetic signal modulation recognition. The effectiveness of the knowledge distillation defense method is verified through experiments, and the optimal temperature is obtained by adjusting the temperature parameters, it is not found that the higher the temperature, the softer the label, the better the student's network learning effect. Using the knowledge distillation method can resist the black-box detection attack of the attacker on the premise of ensuring that our recognition network structure remains unchanged. Specifically, in the face of FGSM-Attack and BIM-Attack, based on the network structure of VTCNN2, the distribution of accuracy increases by 40% (FGSM-Attack) and 52% (BIM-Attack) when SNR = 4 dB. It also slightly improves the performance of the network on clean samples. We believe that our work will help improve the reliability of deep learning algorithms in the field of electromagnetic signal recognition and build a robust electromagnetic signal recognition system.

References

1. 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
2. Peng, S., et al.: Modulation classification based on signal constellation diagrams and deep learning. IEEE Trans. Neural Netw. Learn. Syst. **30**(3), 718–727 (2019)
3. West, N.E., O'Shea, T.: Deep architectures for modulation recognition. In: Proceedings of the IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), Piscataway, NJ, USA, pp. 1–6 (2017)
4. Zhang, Z., Luo, H., Wang, C., Gan, C., Xiang, Y.: Automatic modulation classification using CNN-LSTM based dual-stream structure. IEEE Trans. Veh. Technol. **69**(11), 13 521–13 531 (2020)
5. Rajendran, S., Meert, W., Giustiniano, D., Lenders, V., Pollin, S.: Deep learning models for wireless signal classification with distributed low-cost spectrum sensors. IEEE Trans. Cogn. Commun. Netw. **4**(3), 433–445 (2018)
6. Szegedy, C., et al.: Intriguing properties of neural networks (2013). https://arxiv.org/abs/1312.6199
7. Goodfellow, I., Shlens, J., Szegedyn, C.: Explaining and harnessing adversarial examples. In: Proceedinngs of the International Conference on Learning Representations, pp. 189–199 (2015)
8. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world (2016). https://arxiv.org/abs/1607.02533
9. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. Proceedings of the IEEE European Symposium on Security and Privacy, vol. 1, no. 1, pp. 372–387 (2016)

10. Zhou, R., Liu, F., Gravelle, C.W.: Deep learning for modulation recognition: a survey with a demonstration. Behav. Ecol. Sociobiol. **8**, 67366–67376 (2020)
11. Chen, P., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.-J.: ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models (2017). https://arxiv.org/abs/1708.03999
12. Su, J., Vargas, D.V., Kouichi, S.: One pixel attack for fooling deep neural networks (2017). https://arxiv.org/abs/1710.08864
13. Dezfooli, S.M., Fawzi, A., Frossard, P.: DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, no. 1, pp. 2574–2582 (2016)
14. Metzen, J.H., Genewein, T., Fischer, V., Bischoff, B.: On detecting adversarial perturbations. In: Proceedings of the ICLR (2017). https://openreview.net/pdf?id=SJzCSf9xg
15. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In: Proceedings of the International Conference on Learning Represention, pp. 1–13 (2017)
16. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7130–7138 (2017)