Adversarial Training with Knowledge Distillation Considering Intermediate Representations in CNNs

Hikaru Higuchi[1], Satoshi Suzuki[2], and Hayaru Shouno[1(✉)]

[1] The University of Electro-Communications, Chofu, Tokyo, Japan
{h.higuchi,shouno}@uec.ac.jp
[2] NTT Computer and Data Science Laboratories, Yokosuka, Kanagawa, Japan
satoshi.suzuki.xv@hco.ntt.co.jp

Abstract. A main challenge for training convolutional neural networks (CNNs) is improving the robustness against adversarial examples, which are images with added the artificial perturbations to induce misclassification in a CNNs. This challenge can be solved only by adversarial training, which uses adversarial examples rather than natural images for CNN training. Since its introduction, adversarial training has been continuously refined from various points of view. Some methods focus on constraining CNN outputs between adversarial examples and natural images, resembling knowledge distillation training. Knowledge distillation was originally intended to constrain the outputs of teacher–student CNNs to promote generalization of the student CNN. However, recent methods for knowledge distillation constrain intermediate representations rather than outputs to improve performance for natural images because it directly works well to preserve intraclass cohesiveness. To further investigate adversarial training using recent knowledge distillation methodology (i.e., constraining intermediate representations), we attempted to evaluate this method and compared it with conventional ones. We first visualized intermediate representations and experimentally found that cohesiveness is essential to properly classify not only natural images but also adversarial examples. Then, we devised knowledge distillation using intermediate representations for adversarial training and demonstrated its improved accuracy compared with output constraining for classifying both natural images and adversarial examples.

Keywords: Convolutional neural network · Adversarial training · Knowledge distillation · Intermediate representation · Manifold hypothesis

1 Introduction

Convolutional neural networks (CNNs) play a central role in computer vision for tasks such as an image classification [4,6,11]. However, recent studies have

M. Tanveer et al. (Eds.): ICONIP 2022, CCIS 1791, pp. 683–691, 2023.
https://doi.org/10.1007/978-981-99-1639-9_57

demonstrated that adversarial perturbations, which are artificially made to induce misclassification in a CNN, can cause a drastic decrease in the classification accuracy [16]. In general, humans can naturally and correctly classify adversarial examples, which are images with adversarial perturbations. Since CNNs are originally inspired by human visual systems [4], they should be able to treat adversarial examples in the same way as natural images, like humans. Thus, a main challenge in training CNNs is improving their robustness against adversarial examples, as humans naturally do.

To correctly classify adversarial examples, Mądry *et al.* [13] introduced adversarial training, which uses adversarial examples instead of natural images for CNN training (Fig. 1(a)). Athalye *et al.* [1] found that only adversarial training improves classification robustness for adversarial examples, although diverse methods have been explored. Therefore, subsequent studies have been focused on improving adversarial training [9,10,15,17]. For instance, various methods improve the robustness by constraining the CNN output. For example, ALP [10] and TRADES [17] force the CNN output to be similar for adversarial examples and natural images during adversarial training (Fig. 1(b)). Hence, the corresponding CNNs provide similar outputs regardless of adversarial perturbation. More recent methods such as Smooth Logits [2] or LBGAT [3] employ knowledge distillation, whose constraints bring the outputs of a student (adversarial-trained) CNN closer to those of a teacher (pretrained) CNN (Fig. 1(c)). Knowledge distillation is effective for adversarial training because it enables the student CNN to imitate the decision boundary of the teacher CNN, which is sufficiently generalized after pretraining.

Remarkably, knowledge distillation using intermediate representations rather than outputs in CNNs can further improve the classification performance for usual natural image classification [7,14]. This is because intermediate representations easily determine the decision boundary between classes and preserve intraclass cohesiveness.

As the adversarial training methods in [2,3,10,17] focus only on outputs, the CNNs may not properly reflect intraclass cohesiveness. In contrast, if a CNN with adversarial training can use intermediate representations to similarly classify natural images and adversarial examples, its performance may be improved (Fig. 1(d)). Thus, intermediate representations of CNNs during adversarial training should be further explored. Accordingly, we analyzed CNNs trained with adversarial training by 1) visualizing the intermediate representations and 2) resembling knowledge distillation in intermediate representations to improve the performance of adversarial training.

The contributions of this study can be summarized as follows:

- We confirm phenomena observed in intermediate representations of CNNs trained with adversarial training.
- We visualized intermediate representations and experimentally verify that cohesiveness is essential to correctly classify not only natural images but also adversarial examples.

– We introduce knowledge distillation using intermediate representations and demonstrate that this method is more effective than knowledge distillation at the outputs for improving the classification accuracy of both natural images and adversarial examples.

2 Experimental Analysis in CNNs with Adversarial Training

We begin by formulating adversarial training and investigate phenomena observed in the corresponding CNNs. First, we explain adversarial training and then explain phenomena observed in CNNs based on the ResNet-18 architecture [6] (Fig. 2).

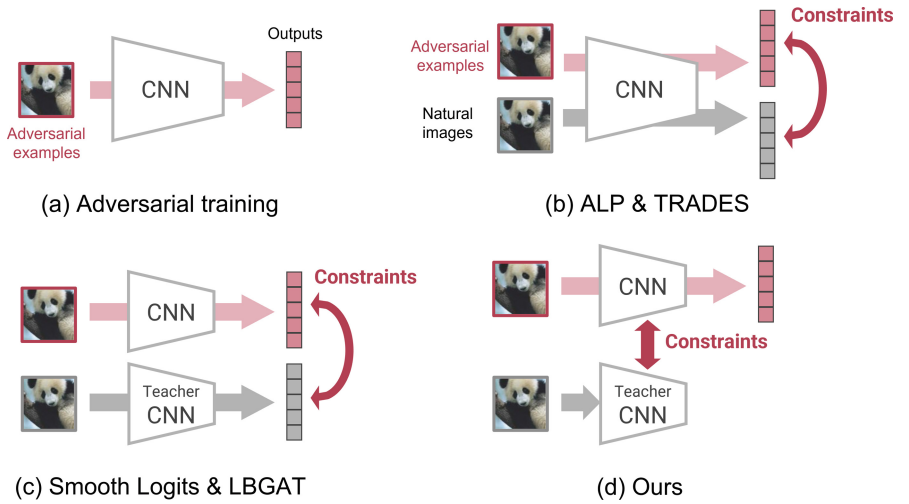

Fig. 1. (a) – (c): Conventional methods for comparison. (d): Our proposed method based on adversarial training.

2.1 Adversarial Training

CNN training using natural images can be formulated as follows:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{p(\boldsymbol{x},y)} \left[\mathcal{L}_{\text{class}}(\boldsymbol{x}, y; \boldsymbol{\theta}) \right], \tag{1}$$

where $(\boldsymbol{x}, y)$ is the pair of input image and its true label and $\mathcal{L}_{\text{class}}(\boldsymbol{x}, y; \boldsymbol{\theta})$ is the classification loss given CNN parameters $\boldsymbol{\theta}$ and data $(\boldsymbol{x}, y)$. Cross-entropy is commonly used as the classification loss. Equation (1) can be interpreted as an optimization problem to search parameters $\boldsymbol{\theta}$ that minimize classification loss $\mathcal{L}_{\text{class}}$.

Meanwhile, adversarial training is formulated as follows:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{p(\boldsymbol{x},y)} \left[\max_{\boldsymbol{\delta}:||\boldsymbol{\delta}||_q \leq \epsilon} (\mathcal{L}_{\text{class}}(\boldsymbol{x}+\boldsymbol{\delta}, y; \boldsymbol{\theta})) \right], \tag{2}$$

where $\boldsymbol{\delta}$ is an adversarial perturbation bounded by an ℓ_q-norm ball. Equation (2) solves the minimization in Eq. (1) after solving the maximization problem for loss function $\mathcal{L}_{\text{class}}$ for $\boldsymbol{x}$ given $\boldsymbol{\theta}$, $\boldsymbol{x}$, and y. Min–max optimization is repeated to obtain parameter $\boldsymbol{\theta}$ that is robust to adversarial examples. As the maximization problem in Eq. (2) cannot be solved explicitly, it is often approximated by applying a strong attack method called projected gradient descent (PGD) [13].

2.2 Visualization of Intermediate Representations in CNNs

We also evaluate intermediate representations between vanilla-CNN trained only with natural images and adv-CNN with conventional adversarial training [13]. Specifically, we visualize and compare intermediate representations of the CNNs by using t-SNE [12] for dimensionality reduction of intermediate representations. We use ResNet-18 [6] (Fig. 2) as the CNN and PGD for adversarial attack [13]. PGD performs strong adversarial attacks by repeatedly generating adversarial perturbations using the fast-gradient sign method [5]. In this study, we used 10 and 20 iterations for the adversarial attack during training and testing, respectively, and the CIFAR-10 as the image classification dataset.

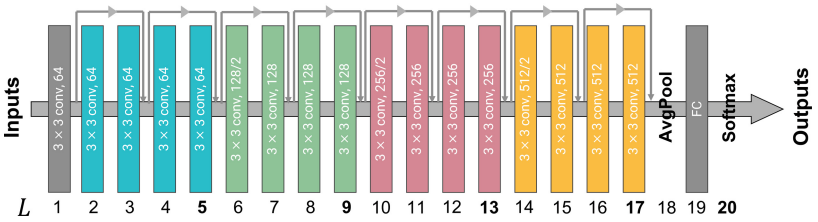

Fig. 2. ResNet-18 architecture (L is described in Eq. (3)).

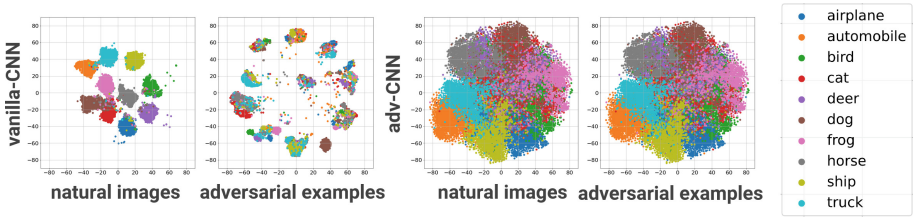

Fig. 3. Visualization of intermediate representations in vanilla-CNN (former two row) and adv-CNN (latter two row) for natural images and adversarial examples.

The former two row of Fig. 3 shows intermediate representations in vanilla-CNN after dimensionality reduction. The upper-left graph of Fig. 3 shows the intermediate representations for natural images, while the upper-right graph shows the representations for adversarial examples. Vanilla-CNN can suitably gather intermediate representations for each class in natural images. The clusters of intermediate representations can contribute to higher classification accuracy for natural images. When the natural images are affected by adversarial perturbations, the clusters are dispersed in the feature space. Hence, adversarial examples degrade intraclass cohesiveness and cause a drastic decrease in the classification accuracy.

The latter two row of Fig. 3 shows intermediate representations in adv-CNN. As shown in the figure, using adv-CNN, similar intermediate representations are obtained for adversarial examples and natural images. However, adv-CNN provides inferior intermediate representations for natural images compared with vanilla-CNN (lower-left graph of Fig. 3) In fact, adv-CNN provides lower accuracy than vanilla-CNN for natural images because it cannot establish clear decision boundaries to classify such images.

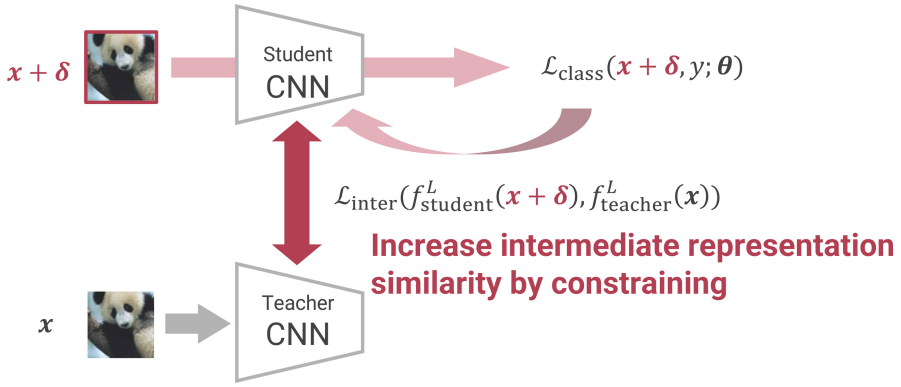

Fig. 4. Diagram of proposed method.

3 Proposed Method: Adversarial Training with Knowledge Distillation

As we mentioned above, vanilla-CNN should have acquired effective representations for classifying natural images. Therefore, in this section, we propose a novel method that adversarially trains the CNN while constraining its representation to preserving the one of vanilla-CNN for natural images.

3.1 Knowledge Distillation

Knowledge distillation [8] shares the representations and constrains the output of a student model from that of a teacher model. Hense, it improve the performance

of the student model (training target). Among a lot of knowledge distillation method, we employed a method using intermediate constraint loss, which aims to bring intermediate representation of the student model closer to those in the teacher model [7,14].

3.2 Adversarial Training with Knowledge Distillation

We propose an adversarial training method with knowledge distillation that employs a CNN trained with natural images as the teacher model. Figure 4 shows a diagram of the proposed method. The student model is the target of adversarial training, and the teacher vanilla-CNN accurately classifies natural images. We aim to make the intermediate representations of the training target similar to those of the teacher vanilla-CNN.

Equation (3) shows the method formulation as an optimization problem.

$$\min_{\theta} \mathbb{E}_{p(\boldsymbol{x},y)} \left[\max_{\boldsymbol{\delta}:||\boldsymbol{\delta}||_q \leq \epsilon} \left(\mathcal{L}_{\text{class}}(\boldsymbol{x}+\boldsymbol{\delta}, y; \boldsymbol{\theta}) + \alpha \cdot \mathcal{L}_{\text{inter}}(f^L_{\text{student}}(\boldsymbol{x}+\boldsymbol{\delta}), f^L_{\text{teacher}}(\boldsymbol{x})) \right) \right] \tag{3}$$

The loss in Eq. (3) consists of two functions, classification loss $\mathcal{L}_{\text{class}}$ and intermediate constraint loss $\mathcal{L}_{\text{inter}}$. In addition, f^L is the intermediate representations of layer L and α is a hyperparameter that determines the contribution of $\mathcal{L}_{\text{inter}}$ to training. Moreover, $\mathcal{L}_{\text{class}}$ improves the classification accuracy for adversarial examples $\boldsymbol{x}+\boldsymbol{\delta}$, and it is the same loss as in conventional adversarial training [13], while loss $\mathcal{L}_{\text{inter}}$ makes the intermediate representations of the student model ($f^L_{\text{target}}(\boldsymbol{x}+\boldsymbol{\delta})$) similar to those of vanilla-CNN ($f^L_{\text{teacher}}(\boldsymbol{x})$).

Table 1. Classification accuracy of evaluated CNNs. The value in boldface indicates the best result, and the underlined value indicates the second best result on each column.

Model	Alpha	Accuracy (natural)	Accuracy (adv)
vanilla-CNN	-	**0.949**	0.0
adv-CNN	-	0.847	0.483
outKD-CNN	0.01	0.849	0.486
outKD-CNN	0.1	0.855	0.500
outKD-CNN	0.5	0.857	0.502
interKD-CNN	1	0.850	0.493
interKD-CNN	50	<u>0.870</u>	**0.522**
interKD-CNN	100	0.866	<u>0.521</u>

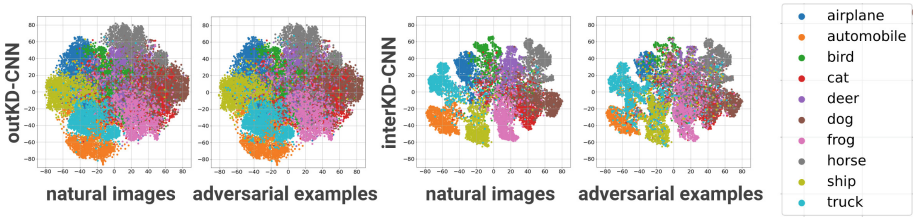

Fig. 5. Visualization of intermediate representations in outKD-CNN (former two row) and interKD-CNN (latter two row) for natural images and adversarial examples.

4 Experimental Evaluation

We compared the proposed method with output constraining [2,3] and evaluated the constraint effectiveness.

4.1 Experimental Setup

We conducted experiments under the same conditions as in the experimental analysis reported in Sect. 2. We used the mean squared error as intermediate constraint loss $\mathcal{L}_{\text{inter}}$. Let us denote the CNN trained using the proposed method (Eq. (3), Fig. 4) for constrained layer $L = 20$ (i.e., using outputs as constraints) as outKD-CNN and the CNN with constrained layer $L < 20$ (i.e., using intermediate representations as constraints) as interKD-CNN.

4.2 Classification Accuracy

Table 1 lists the classification accuracy of vanilla-CNN trained only with natural images, adv-CNN trained with conventional adversarial training [13], outKD-CNN [2,3], and interKD-CNN. We evaluated weight $\alpha \in \{1, 50, 100\}$ for outKD-CNN and $\alpha \in \{0.01, 0.1, 0.5\}$ for interKD-CNN. The CNNs with adversarial training and knowledge distillation (outKD-CNN and interKD-CNN) tend to achieve higher accuracy than adv-CNN for natural images and adversarial examples. InterKD-CNN ($\alpha = 50, L = 17$) exhibits the highest accuracy for adversarial examples and the second highest accuracy for natural images among the evaluated CNNs, even outperforming outKD-CNN. Thus, constraining intermediate representations seems more effective for improving the classification accuracy than constraining outputs.

4.3 Visualization of Intermediate Representations

To evaluate the representations obtained from training with the proposed method, we evaluated the CNN trained using proposed method in terms of intermediate representations, as in Sect. 2. Figure 5 (former two row) and Fig. 5

(latter two row) show intermediate representations obtained from interKD-CNN ($\alpha = 50$, $L = 17$) and outKD-CNN ($\alpha = 0.5$, $L = 20$), respectively. As shown in Fig. 5, interKD-CNN obviously has cohesive intermediate representations compared with outKD-CNN, as we expected. Hence, knowledge distillation in interKD-CNN effectively worked as an anchor to preserve the representations of each class for natural images provided by vanilla-CNN and promotes the classification accuracy.

5 Conclusions

After evaluating intermediate representations in CNNs, we found that training using only natural images provides effective intermediate representations in terms of classifying natural images, while conventional adversarial training does not. This indicates that intraclass cohesiveness is important to correctly classify natural images. Accordingly, we propose a method involving knowledge distillation using intermediate representations from a teacher CNN trained only using natural images to a student CNN with adversarial training. This method aims to preserve representations for natural images of the teacher, achieving a higher accuracy than CNNs with conventional adversarial training.

As future works, we will further explore an effective training method in preserving representation for adversarial examples and achieving higher classification performance. Also, in this study, we used the mean squared error as the intermediate constraint loss to achieve similar intermediate representations for natural images and adversarial examples, but this loss may be inappropriate. In future work, we will explore more appropriate loss functions for constraining by considering the characteristics of intermediate representations (e.g., manifolds).

Acknowledgement. This study was partly supported by MEXT KAKENHI, Grant-in-Aid for Scientific Research on Innovative Areas 19H04982 and Grant-in-Aid for Scientific Research (A) 18H04106.

References

1. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In: International Conference on Machine Learning (ICML) (2018)
2. Chen, T., Zhang, Z., Liu, S., Chang, S., Wang, Z.: Robust overfitting may be mitigated by properly learned smoothening. In: International Conference on Learning Representations (ICLR) (2020)
3. Cui, J., Liu, S., Wang, L., Jia, J.: Learnable boundary guided adversarial training. In: IEEE International Conference on Computer Vision (ICCV) (2021)
4. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. **36**(4), 193–202 (1980)
5. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)

6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
7. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: IEEE International Conference on Computer Vision (ICCV) (2019)
8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. Stat (2015)
9. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
10. Kannan, H., Kurakin, A., Goodfellow, I.: Adversarial logit pairing. arXiv preprint arXiv:1803.06373 (2018)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2012)
12. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. (JMLR) (2008)
13. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR) (2018)
14. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
15. Stutz, D., Hein, M., Schiele, B.: Disentangling adversarial robustness and generalization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
16. Szegedy, C., et al.: Intriguing properties of neural networks. In: International Conference on Learning Representations (ICLR) (2014)
17. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning (ICML) (2019)