Research Article GU J Sci, Part A, 11(2): 274-288 (2024) 10.54287/gujsa.1458880

Gazi University
Journal of Science
PART A: ENGINEERING AND INNOVATION
http://dergipark.org.tr/gujsa

Perturbation Augmentation for Adversarial Training with Diverse Attacks

Duygu SERBES[1] İnci M. BAYTAŞ[1*]

[1] Boğaziçi University, Department of Computer Engineering, İstanbul, Türkiye

Keywords

Adversarial Attacks
Adversarial Training
Adversarial Robustness
Deep Neural Networks

Abstract

Adversarial Training (AT) aims to alleviate the vulnerability of deep neural networks to adversarial perturbations. However, the AT techniques struggle to maintain the performance on natural samples while improving the deep model's robustness. The absence of perturbation diversity in generated during the adversarial training degrades the generalizability of the robust models, causing overfitting to particular perturbations and a decrease in natural performance. This study proposes an adversarial training framework that augments adversarial directions from a single-step attack to address the trade-off between robustness and generalization. Inspired by feature scattering adversarial training, the proposed framework computes a principal adversarial direction with a single-step attack that finds a perturbation disrupting the inter-sample relationships in the mini-batch during adversarial training. The principal direction obtained at each iteration is augmented by sampling new adversarial directions within a region spanning 45 degrees from the principal adversarial direction. The proposed adversarial training approach does not require extra backpropagation steps in adversarial direction augmentation. Therefore, generalization of the robust model is improved without posing an additional burden on the feature scattering adversarial training. Experiments on CIFAR-10, CIFAR-100, SVHN, Tiny-ImageNet, and The German Traffic Sign Recognition Benchmark consistently improve the accuracy on adversarial with an almost pristine natural performance.

Cite

Serbes, D., & Baytas, I. M (2024). Perturbation Augmentation for Adversarial Training with Diverse Attacks. *GU J Sci, Part A, 11*(2), 274-288. doi:10.54287/gujsa.1458880

Author ID (ORCID Number)

0000-0003-1067-866X Duygu SERBES
0000-0003-4765-2615 İnci M. BAYTAS

Article Process

Submission Date 26.03.2024
Revision Date 29.04.2024
Accepted Date 21.05.2024
Published Date 04.06.2024

1. INTRODUCTION

Deep Neural Networks (DNNs) establish the best performances in various fields with challenging problems, including natural language processing (Alzantot, 2018), and image (Goodfellow et al., 2014; Moosavi-Dezfooli et al., 2016) and speech (Carlini et al., 2016) recognition. Although the DNNs are versatile and successful, they are susceptible to imperceptible perturbations named adversarial attacks (Szegedy et al., 2014). Adversarial attacks can force a network into misclassifying an input that was correctly classified before the attack. Such attacks are commonly encountered as additive perturbations that the human eye cannot catch in vision applications (Goodfellow et al., 2014). The vulnerability of deep models against adversarial attacks has raised concerns about their reliability and robustness, specifically in safety-critical applications in the fields, e.g., autonomous driving (Wang et al. 2021), medical diagnostics (Finlayson et al., 2019), and finance (Fursov et al., 2021). As the vulnerability against various adversarial attacks emerges, developing defense mechanisms to fortify deep models against such attacks has become of great interest to the applications employing DNNs.

Adversarial Training (AT) is one of the commonly proposed approaches to defend against adversarial attacks. AT (Madry et al., 2018; Tramer et al., 2018) essentially trains a DNN with a training set augmented or replaced by the adversarial counterparts of training samples. The adversarial samples used in the AT should be generated

*Corresponding Author, e-mail: inci.baytas@bogazici.edu.tr

at each training iteration. Consequently, AT can be posed as a min-max optimization problem where the inner maximization generates the adversarial perturbation, and the model is updated via the minimization step. Although it is an effective defense, AT has specific challenges and limitations. When adversarial attacks are generated by attacking a supervised loss function with labels, correlations between the perturbations and the ground-truth labels might emerge during the AT leading to label leakage (Kurakin et al., 2017), degrading the model's generalization. Furthermore, the insufficient diversity in adversarial perturbations generated during training causes catastrophic overfitting (Wong et al., 2020; Kim et al., 2021).

Adversarial defense techniques often aim to alleviate the gap between robustness and generalization. Previous studies show that generating more complex (Madry et al., 2018; Schmidt, 2018) and diverse adversarial samples during AT may improve the robustness up to some extent (Jang, 2019). However, strong attacks may hurt the model's generalization to unseen natural samples (Zhang H. et al., 2019). We stress that the goal of adversarial robustness should not be improving the robust accuracy while sacrificing the model's accuracy on the natural test samples. To alleviate this trade-off, mixup (Zhang et al., 2018) and feature scattering-based techniques (Zhang H. et al., 2019; Baytaş & Deb, 2023) are employed in the literature. More recently, augmenting the training set with millions of images synthesized by generative models has also been employed to improve both robust and natural performance (Wang et al., 2023). However, generating millions of images to train a robust model would not be sufficient with limited resources.

This study proposes an adversarial training framework inspired by Zhang & Wang (2019), where adversarial perturbations are generated by disrupting the inter-sample relations in the mini-batch. On the other hand, the proposed approach augments the adversarial perturbations during training to improve the generalization. The primary motivation behind this study is to preserve the natural accuracy and avoid exacerbating the training complexity while enhancing the adversarial robustness. Training with strong but not diverse attacks hurts the model's generalization to different adversarial perturbations and natural samples. On the other hand, robust training with relatively weaker attacks to alleviate overfitting to specific adversarial perturbations improves generalization to natural samples but degrades the adversarial robustness. Therefore, this study proposes Perturbation Direction Augmentation for Adversarial Training (PDA-AT) to increase the attack diversity during training and enhance the robustness and generalization of the DNN. Contributions of the study are outlined below.

- Differing from the standard feature scattering AT (Zhang & Wang, 2019), we generate perturbations that increase the optimal transport distance between different samples of the mini-batch instead of between the natural mini-batch and its randomly perturbed version. We empirically show that we can obtain more diverse adversarial directions at each iteration.
- Adversarial perturbations are augmented by sampling new adversarial directions within a region spanning 45 degrees from the principal adversarial direction obtained from the gradient of the distance between the mini-batch of data points measured by optimal transport.
- Perturbation augmentation can enhance the adversarial robustness compared with the baseline AT methods without synthetic images. Notably, the gap between the natural and adversarial accuracies is substantially reduced.
- The proposed approach addresses the trade-off between robustness and generalization. The augmented perturbations provide robustness without sacrificing the accuracy for the natural test samples.
- Experimental results on CIFAR-10, CIFAR-100, SVHN, Tiny-ImageNet, and The German Traffic Sign Recognition Benchmark (GTSRB) datasets demonstrate consistent performance improvement for PDA-AT with augmented perturbations compared with the baseline AT techniques in the literature.

The following Literature Review presents related studies from the adversarial robustness literature. The Material and Method section explains the proposed framework and deep model architecture in detail. The Experimental Results section presents the performance of PDA-AT and a comparison with the baselines. The experimental results are interpreted in the Discussion section. The Conclusion section summarizes the proposed approach and key findings.

2. LITERATURE REVIEW

Adversarial attacks (Szegedy et al., 2014; Goodfellow et al., 2014) and robustness literature have been growing since it became evident that DNNs are sensitive against imperceptible but adversarially crafted perturbations. AT has become an effective and notable method to defend DNNs against adversarial attacks and improve robustness with the help of various adversarial attack algorithms. Goodfellow et al. (2014) were among the first studies to introduce an AT approach with Fast Gradient Sign Method (FGSM) attack. Their AT scheme augments the training samples in a mini-batch with adversarial samples obtained with FGSM. In other words, they optimize two cross-entropy terms computed with natural and adversarial samples separately. On the other hand, Madry et al. (2018) observed that incorporating natural samples in the AT weakens the robustness. Therefore, they proposed optimizing the model with only adversarial samples, given in Equation 1, generated by an iterative gradient-based adversarial attack named Projected Gradient Descent (PGD).

$$\mathbf{x}^{t+1} = \prod_{\mathbf{x}+S} \left(\mathbf{x}^t + \eta \text{sign}\left(\nabla_{\mathbf{x}^t} \mathcal{L}(\mathbf{x}, y; \theta) \right) \right) \tag{1}$$

where t denotes the attack iteration, S is the allowed perturbations set, η is the attack step size, $\prod_{\mathbf{x}+S}(\cdot)$ is the projection operator of the L-infinity ball, and θ is the parameters of a deep model.

PGD can explore stronger attacks than a single-step adversarial attack such as FGSM. However, robustness literature often discusses that the trade-off between robustness and generalization inevitably grows when the PGD attack is used in AT (Baytaş & Deb, 2023). For this reason, various modifications and improvements are proposed to address the lack of generalization of PGD adversarial training (Wong & Kolter, 2018; Wang et al., 2019; Zhang & Wang, 2019). Nevertheless, there are challenges, such as label leaking (Kurakin et al., 2017) and gradient masking (Athalye et al., 2018), that leads to a misleading robustness. Although training a deep model with powerful adversarial perturbations, such as PGD, boosts robustness more than employing FGSM attacks, training with PGD attacks is impractical for large-scale problems due to its increased number of backpropagation steps.

To alleviate AT's time complexity, researchers attempted to attain state-of-the-art robustness with single-step attacks (Zhang D. et al., 2019; Shafahi et al., 2019; Wong et al., 2020) by adopting accumulative perturbation and perturbation initialization. Shafahi et al. (2019) employed a single backpropagation step to update the model weights and generate the adversarial perturbations. On the other hand, Wong et al. (2020) identified the reasons behind the phenomenon where robust accuracy drops to 0 % during AT with FGSM attacks. Authors claimed that random initializing of the FGSM attack mitigates catastrophic overfitting and attains the desired level of robustness (Wong et al., 2020). On the other hand, Andriushchenko & Flammarion (2020) discussed the correlation between the catastrophic overfitting and local non-linearities. Consequently, the authors proposed a regularizer, GradAlign, to produce stronger adversarial examples by explicitly maximizing the gradient alignment within the attack. Furthermore, Kim et al. (2021) explained that distortions in decision boundaries and a highly curved loss surface due to characteristics of FGSM-based AT cause overfitting, which they address with an appropriate step size. However, the experimental results of these studies show that FGSM-based AT approaches suffer from poor generalization to natural samples while their robustness does not significantly exceed the PGD AT's performance by Madry et al. (2018).

Mixup-based approaches have interested the adversarial robustness domain due to their potential to enhance the generalization of AT. Mixup introduces linear behavior in the data manifold by interpolating inputs and their labels (Zhang et al., 2018). It becomes evident in the literature that the AT with mixup might outperform standard AT while improving the generalization performance (Lee et al., 2020). In the literature, various mixup studies are proposed to support adversarial robustness. For instance, Manifold Mixup (Verma et al., 2019) aims to smooth decision boundaries for multiple levels of hidden representations of deep models that provide robustness against single-step adversarial attacks. Furthermore, Lee et al. (2020) introduced Adversarial Vertex Mixup, a soft-labeled data augmentation approach that interpolates the virtual adversarial vector and the natural input.

This proposed PDA-AT is inspired by the feature scattering AT method of Zhang & Wang (2019). Feature scattering is based on a single-step adversarial attack that concerns the inter-sample relationships between adversarial and natural samples defined by the optimal transport (OT) distance. An unsupervised perturbation is generated by maximizing the OT of the perturbed and natural images. In addition, the authors also designed bilateral adversarial training (Wang & Zhang, 2019), where both images and labels are attacked during training with a single targeted perturbation where the target is the most mistaken class. Although the aforementioned AT frameworks try to address several issues of AT, there is still room for improvement.

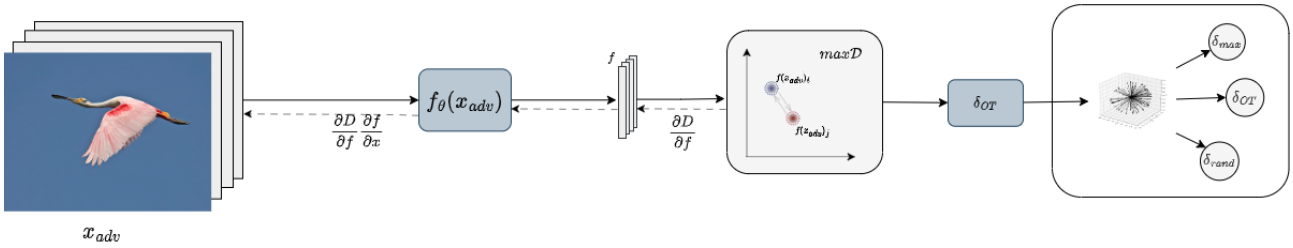

Figure 1. *Proposed PDA-AT Adversarial Training Framework*

3. MATERIAL AND METHOD

One of the widely discussed causes of the model's susceptibility against adversarial perturbations is explained with overfitting. Since deep models are prone to overfitting to the training data distribution, the model's predictions could be altered by the adversarial samples that can be considered out-of-distribution instances. Therefore, some AT approaches try to augment the training data with possible adversarial samples so that the model can have experience in handling adversarial perturbations. Thus, AT can be posed as the following two-step optimization problem:

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},y)\sim D}\left[\max_{\delta\in S} \mathcal{L}(f(\mathbf{x}+\boldsymbol{\delta};\theta),y)\right] \tag{2}$$

where $f(.;\theta)$ is a neural network parametrized by θ, $\mathbf{x}$ is the natural sample, $\boldsymbol{\delta}$ is an adversarial attack, the set S contains the possible perturbations bounded by an L-infinity norm ball with an ϵ maximum amount of perturbation, y is the ground truth, and $\mathcal{L}$ is the loss function, which is cross-entropy for classification tasks.

In the optimization problem above, $\boldsymbol{\delta}$, is the perturbation that maximizes the loss function, where adversarial directions in a single step or multiple steps are obtained. Although AT is expected to contribute to the generalization of the model for adversarial samples, we commonly observe several phenomena, such as catastrophic overfitting (Wong et al., 2020), label leaking (Kurakin et al., 2017), and gradient masking (Athalye et al., 2018; Ilyas et al., 2019) that result in overfitting to certain perturbations and hampering the training. One of the contributing factors to these challenges in AT is the adversarial sample generation based on the gradient of the cross-entropy loss (Baytaş & Deb, 2023). Therefore, the adversarial direction obtained via increasing the cross-entropy loss is insufficient to generate stronger and diverse attacks (Etmann et al., 2019).

One of the most critical weaknesses of the traditional AT approaches is the generalization to natural samples. As the model's adversarial robustness improves, test accuracy on the natural samples drops below what the literature reports, which is unacceptable. We stress that the primary goal of a robust training framework should be preserving the natural performances while enhancing the robustness. Secondly, an AT training framework should be scalable such that the adversarial sample generation procedure should not require multiple costly backpropagation steps.

This study proposes an AT approach, PDA-AT, illustrated in Figure 1, where the adversarial directions are generated using Optimal Transport (OT) distance, and they are augmented to reinforce the generalization. Thus, the goal of PDA-AT is a more generalizable AT regarding robust and natural accuracies. OT distance, also known as Wasserstein distance (Xie et al., 2020) or earth mover's distance, measures the distance between

two distributions based on the minimum cost of transforming one probability distribution into another. OT distance (Villani, 2009; Cuturi, 2013) between two distributions can be defined as:

$$\mathcal{D}(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \mathbb{E}_{(x,y)\sim\gamma} c(x,y) \tag{3}$$

where $\Pi(\mu,\nu)$ represents the joint distributions $\gamma(x,y)$ with marginals of $\mu(x)$ and $\nu(y)$. The cost function, $c(x,y)$ denotes $K \times K$ transport cost matrix C, where $C_{ij} = c(x_i, y_j)$ indicates how much it costs to transfer the i-th data vector in mini-batch **X** to the j-th data vector in mini-batch **Y**.

Feature Scattering based AT, first introduced by Zhang & Wang (2019), replaces PGD attack on the cross-entropy loss in the standard AT (Madry et al., 2018) with an FGSM attack on OT distance between the natural and the adversarial mini-batches. To be more specific, before the OT loss attack is generated, we need to compute the loss. Since a single-step attack is considered, the mini-batch of adversarial samples is obtained by adding a uniform random matrix sampled within the ϵ-ball to the natural samples to compute the OT loss before the attack. In other words, in the standard feature scattering AT, the adversarial direction is generated by altering the inter-sample relationships between a natural mini-batch and its randomly shifted version to deepen the discrepancy between natural and adversarial distributions.

The standard feature scattering AT obtains a perturbation by computing the gradient of the OT distance between the natural image and its randomly shifted counterpart. Thus, the perturbation aims to maximize the gap between the distributions of the mini-batch of natural and randomly shifted samples at each training iteration. Although training with feature scattering perturbation improves the robustness to an extent, the attack diversity, discussed in Section 4.4, gradually decreases during training, which is detrimental to the model's generalization. Therefore, we argue that maximizing the OT distance between natural and randomly shifted samples may limit the attack diversity since the difference between the two sets in question is the amount and the direction of the random shift at each iteration. Consequently, the OT distance between the same mini-batch given in Equation 4 is preferred.

$$\boldsymbol{\delta}_{\mathrm{OT}} = \arg\max_{\mathbf{x}} \mathcal{D}(\boldsymbol{\nu}_{\mathrm{adv}}, \boldsymbol{\nu}_{\mathrm{adv}}) \tag{4}$$

$$\mathcal{D}(\boldsymbol{\nu}_{\mathrm{adv}}, \boldsymbol{\nu}_{\mathrm{adv}}) = \min_{\mathbf{T}\in\Pi(\boldsymbol{\nu}_{\mathrm{adv}}, \boldsymbol{\nu}_{\mathrm{adv}})} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{T}_{ij} \cdot c\left(f(\mathbf{x}_{\mathrm{adv}})_i, f(\mathbf{x}_{\mathrm{adv}})_j\right)$$

where $c\left(f(\mathbf{x}_{\mathrm{adv}})_i, f(\mathbf{x}_{\mathrm{adv}})_j\right)$, given below, is the element of the transport cost matrix which is zero for the same samples.

$$c\left(f(\mathbf{x}_{\mathrm{adv}})_i, f(\mathbf{x}_{\mathrm{adv}})_j\right) = \left\| f(\mathbf{x}_{\mathrm{adv}})_i - f(\mathbf{x}_{\mathrm{adv}})_j \right\|_2^2 \tag{5}$$

The proposed modification aims to facilitate exploring adversarial directions that change the natural sample distribution more freely since we do not force the adversarial direction to specifically deepen the discrepancy between the natural samples and their randomly shifted versions. The proposed perturbation is intended to diverge the adversarial sample distribution from the distribution of the natural samples. Therefore, we claim that the perturbation direction should change at each iteration since the representation distribution will differ after each weight update. Thus, Equation 5 aims to diversify the adversarial direction at each iteration compared with constantly leading the adversarial direction toward increasing the disparity between the natural and randomly perturbed samples. In experiments, we empirically show that the proposed modification increases attack diversity and improves the robustness.

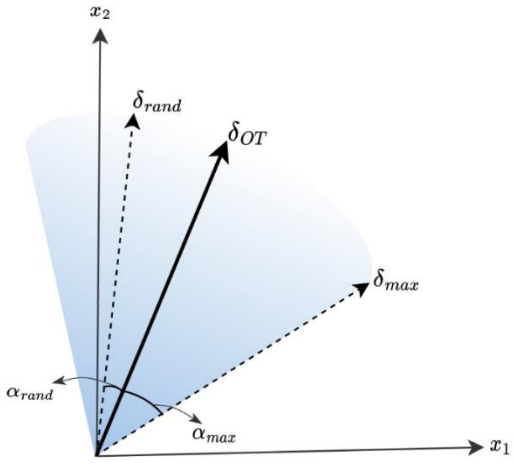

Figure 2. *Adversarial Direction Augmentation in 2-Dimensional Space*

3.1. Adversarial Direction Augmentation

It is evident in adversarial robustness literature that presenting a wide range of perturbations to model training alleviates overfitting to certain perturbations. This study introduces an intuitive and effective adversarial perturbation augmentation approach. The perturbation $\boldsymbol{\delta}_{\mathrm{OT}}$ in Equation 4 denotes the principal adversarial direction that increases the OT loss. This direction contains the most distinctive information about how to increase the gap between the distributions of adversarial and natural data representations. Although relying only on the principal adversarial direction enhances the robustness of the model, the attack diversity could be further improved. However, increasing the attack diversity during training should not create an additional computational overhead since the complexity of AT is already higher than the standard training. Therefore, we propose to benefit from the principal adversarial direction carried in $\boldsymbol{\delta}_{\mathrm{OT}}$ to obtain new adversarial directions without requiring extra backpropagation steps during AT iterations. Particularly, we propose randomly sampling a new adversarial direction within a 45-degree angle from the principal direction as given below:

$$\boldsymbol{\delta}_{\mathrm{rand}} = \cos(\alpha)\boldsymbol{\delta}_{\mathrm{OT}} + \sin(\alpha)\,\boldsymbol{\delta}_{\perp} \tag{6}$$

where α is a random angle between 0 and 45 degrees, and $\boldsymbol{\delta}_{\perp}$ is a randomly generated perturbation vector perpendicular to $\boldsymbol{\delta}_{\mathrm{OT}}$. In addition to $\boldsymbol{\delta}_{\mathrm{rand}}$, we generate one more adversarial direction denoted by $\boldsymbol{\delta}_{45}$ when α is set to 45 degrees. Thus, three adversarial directions, shown in Figure 2, can be obtained to increase the variety of adversarial attacks at each iteration without significantly increasing the computational cost. The fundamental reason behind 45 degrees is to ensure that the new sampled direction maintains its adversarial characteristic. As we move away from the principal adversarial direction that stems directly from the gradient of the OT loss, the perturbation's strength might degrade since the new direction might be toward a decrease in loss. Therefore, we constrain the sampling region at 45-degree from the principal direction so that the sampled perturbations can still decrease the loss up to an extent. We also hypothesize that there should be more than one adversarial direction for each natural sample in the input space. Therefore, we propose to explore more adversarial directions with the sampling approach in Equation 6. it is intuitive to investigate the new potential adversarial attacks obtained from similar directions to the gradient of the loss.

Finally, the mini-batch is augmented with perturbed images to update the model as follows:

$$\min_{\theta} \frac{1}{N}\left[\sum_{i=1}^{N/3} \mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{\mathrm{OT}};\theta), y) + \sum_{i=1}^{N/3} \mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{\mathrm{rand}};\theta), y) + \sum_{i=1}^{N/3} \mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{45};\theta), y)\right] \tag{7}$$

where the mini-batch size is denoted by N. As seen in Equation 7, the model is updated with samples perturbed with the principal and augmented adversarial directions. Clipping is applied to perturbations to stay inside the

ϵ-ball and to the perturbed images to avoid stepping out of the input domain. The proposed PDA-AT framework is given in Algorithm 1.

Algorithm 1 PDA-AT Framework. Robustness for T epochs, D dataset, θ network parameters, ϵ perturbation budget, τ learning rate, n mini-batch size, $\alpha_{\max}$ maximum deviation angle

Require: $\mathcal{D}$ optimal transport (OT) distance

for t $= 1, \cdots,$ T **do**

  for mini-batch $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$~D **do**

    $\delta_{\mathrm{OT}} \leftarrow \arg\max_{\mathbf{x}} \mathcal{D}\left(\boldsymbol{\nu}_{\mathrm{adv}}, \boldsymbol{\nu}_{\mathrm{adv}}\right)$

    $\alpha \sim \mathcal{U}(0, \alpha_{\max})$

    $\boldsymbol{\delta}_{\mathrm{rand}} \leftarrow \cos(\alpha)\boldsymbol{\delta}_{\mathrm{OT}} + \sin(\alpha)\boldsymbol{\delta}_{\perp}$

    $\boldsymbol{\delta}_{\max} \leftarrow \cos(\alpha_{\max})\boldsymbol{\delta}_{\mathrm{OT}} + \sin(\alpha_{\max})\boldsymbol{\delta}_{\perp}$

    $\mathbf{x}' \leftarrow [\mathbf{x} + \boldsymbol{\delta}_{\mathrm{OT}}, \mathbf{x} + \boldsymbol{\delta}_{\mathrm{rand}}, \mathbf{x} + \boldsymbol{\delta}_{\max}]$

    model update:

    $\theta \leftarrow \theta - \tau \frac{1}{n} \sum_{i=1}^{n} \nabla_{\theta} \mathcal{L}(\mathbf{x}', y; \theta)$

  end for

end for

4. EXPERIMENTAL RESULTS

The PDA-AT's robustness is evaluated in both white-box and black-box settings. Extensive experiments are conducted on five commonly used benchmark datasets, including CIFAR-10, CIFAR-100 (Krizhevsky & Hinton, 2009), The Street View House Numbers (SVHN) (Yuval et al., 2011), German Traffic Sign Recognition Benchmark (GTSRB) (Houben et al., 2013), and Tiny-Imagenet (Le & Yang, 2015). The selected datasets are widely used benchmarks in the adversarial robustness literature. Furthermore, the datasets have varying characteristics such as the number of categories, background, and scale. Therefore, the datasets of interest are suitable for evaluating the adversarial robustness of the proposed method and compare it with literature.

4.1. Baselines

The PDA-AT's performance is compared with the common AT baseline methods that are based on gradient-based attacks. The proposed method does not utilize synthetically generated data during training, nor do the baselines. Performances of the following baselines are reported:

Natural: Standard training with natural images.

Madry: AT with PGD attack proposed by Madry et al. (2018), accepted as one of the most effective defense methods.

Bilateral: Training with both image and label adversarial perturbations (Wang & Zhang, 2019).

Feature Scattering: Maximizes the OT distance between natural and perturbed images (Zhang & Wang, 2019).

Adv-Interp: Generating of adversarial samples by adversarial interpolation (Zhang & Xu, 2020).

AV-Mixup: Adversarial training of data augmentation-based soft labeling approach (Lee et al., 2020).

4.2. Datasets

Comprehensive experiments are conducted across five publicly available datasets with various configurations. The CIFAR-10 dataset (Krizhevsky & Hinton, 2009) is widely recognized as the benchmark for adversarial training and comprises 32×32 RGB images of ten different object classes. The training set of CIFAR-10 contains a total of 50K images, 5K images per class, whereas the test set has 10K images. Similarly, the CIFAR-100 dataset (Krizhevsky & Hinton, 2009) comprises the same number of images and dimensions as CIFAR-10 but includes 100 object classes, each containing 500 images in a total of 50K images in the training set. CIFAR-10 and 100 datasets are constructed as a subset of 80 million tiny images dataset. The categories in the dataset do not have overlapping samples. The CIFAR-10 and CIFAR-100 are well-known object recognition datasets used for benchmarking in wide-variety of problems. The SVHN dataset (Yuval et al., 2011) is another widely used dataset in computer vision and comprises approximately 100K labeled digit images with varying sizes and orientations collected from Google Street View house numbers. The SVHN dataset is divided into $73{,}257$ training images and $26{,}032$ test images. Although the SVHN dataset also has 10 categories of same sized images as CIFAR-10, the SVHN images has digits in the wild instead of objects. Therefore, the SVHN data characteristics are substantially different from CIFAR-10 and 100 datasets. Tiny-ImageNet (Le & Yang, 2015) consists of 110 K images of 200 classes, each with 500 training images and 50 validation images. The Tiny-ImageNet dataset poses a challenge due to the high number of categories with less number of samples from each category. Thus, we can investigate how the number of classes impact the robustness of AT. The GTSRB (Houben et al., 2013) comprises a collection of 43 different traffic sign classes, featuring a total of $39{,}209$ training images and $12{,}630$ test images. The images in the dataset exhibit diverse lighting conditions and complex backgrounds, complicating the traffic sign recognition models. The GTSRB is also selected since its a large-scale multi-class image dataset of completely different pattern than the above-mentioned datasets.

4.3. Training Scheme

In all experiments, the WideResnet28-10 model (Zagoruyko & Komodakis, 2016) is used for the object recognition task. The number of epochs is set to 200 with a batch size of 60 compatible with the configurations in Feature Scattering AT study (Zhang & Wang, 2019). The optimizer is chosen as the Stochastic Gradient Descent with a weight decay of 2×10^{-4}, a momentum of 0.9, and a learning rate of 0.01 and 0.1 for SVHN and other datasets, respectively. The learning rate decays with 0.1 at the epochs 30 and 60.

PyTorch is used in the experiments. The codebase provided by Zhang & Wang (2019) is modified to implement PDA-AT. In the experiments, the maximum perturbation amount is set to $\epsilon = 8/255$, which is the commonly accepted maximum perturbation amount by the literature for the benchmark datasets considered in this study. Data augmentations, random crops and flips, are used to improve the model generalization for CIFAR and Tiny-ImageNet datasets. Label smoothing is applied with a factor of 0.5. For all datasets except Tiny-ImageNet, the images are randomly cropped into 32×32 with a padding size of 4. Meanwhile, images from Tiny-ImageNet are cropped to 64×64. To calculate the optimal transport (OT) distance, we adopted the Sinkhorn algorithm with a regularization parameter of 0.01 in a one-step attack configuration (Zhang & Wang, 2019).

4.4. Attack Diversity

One of the essential motivations of this study is to enhance the attack diversity during AT. The proposed approach is based on the inter-sample relationship between adversarial samples within the mini-batch to augment the attacks. We investigate the attack diversity by comparing the proposed method with Feature Scattering AT (Zhang & Wang, 2019). In that regard, we first create adversarial directions of CIFAR-10 training images at several epochs. Next, we sample a random direction as a reference point to compute cosine similarities with the generated adversarial directions. The standard deviations of the cosine similarities at various epochs are plotted in Figure 3. Similarly, the change in the standard deviations of the element-wise sum of the gradient direction tensor's elements for each sample can be seen in Figure 3. In both plots, we expect the proposed attack against the OT distance between the inter-samples of the mini-batch given in Equation 4 to generate a higher standard deviation than the standard Feature Scattering attack (Zhang & Wang, 2019).

Figure 3 provides empirical evidence that the proposed approach can generate more diverse attack directions than the standard single-step OT distance attack.

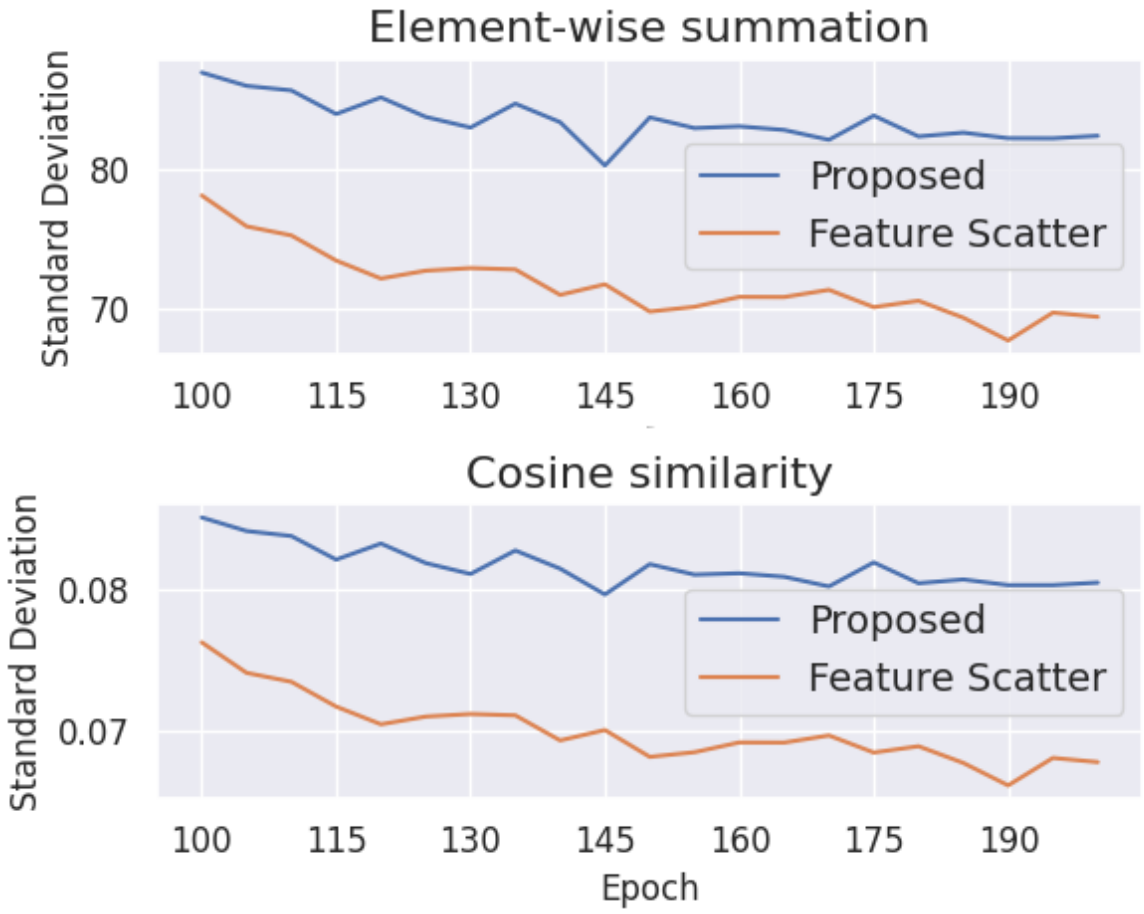

Figure 3. *Diversity of adversarial directions in training*

4.5. Variations of Perturbation Augmentations

This section investigates the optimal combination of the proposed perturbation augmentations. In this experiment, the OT distance between natural and adversarial samples is considered to obtain the primary adversarial direction, $\boldsymbol{\delta}_{\mathrm{FS}}$, as in Feature Scattering AT (Zhang & Wang, 2019). Thus, we can observe the independent effect of the perturbation augmentation given in Table 1 below. The table demonstrates natural and robust test accuracies of proposed model trained with augmented adversarial perturbations. Natural column presents the classification accuracy of the robust model on natural test images. FGSM column demonstrates the performance on test images with FGSM attack, which is a single-step attack. On the other hand, PGD20 and CW20 columns are results on test images with PGD and Carlini and Wagner (CW) (Carlini & Wagner, 2017) attacks of 20 steps, respectively. Thus, the Table 1 investigates the contributions of different adversarial direction augmentations to the classification performance.

Table 1. *Performance of the proposed PDA-AT framework with various perturbation augmentations. Results denote classification accuracy in percentages*

Perturbation Augmentation	Natural	FGSM	PGD20	CW20
$\boldsymbol{\delta}_{FS}, \boldsymbol{\delta}_{\mathrm{rand}}$	92.98	92.54	62.83	58.60
$\boldsymbol{\delta}_{FS}, \boldsymbol{\delta}_{\mathrm{rand}}, \boldsymbol{\delta}_{\mathrm{rand}}$	92.44	89.27	59.74	57.40
$\boldsymbol{\delta}_{FS}, \boldsymbol{\delta}_{\mathrm{rand}}, \boldsymbol{\delta}_{45}$	93.68	94.33	68.08	61.88

Once the primary direction $\boldsymbol{\delta}_{\mathrm{FS}}$ is identified, we augment it within a cosine similarity region of 45 degrees. Through various experiments, we determined that the most effective attack perturbation involved a combination of the primary perturbation direction and augmented directions generated by both random and maximum allowed perturbations. Table 1 shows that adding two random perturbation directions, $\boldsymbol{\delta}_{\mathrm{rand}}$, could hurt the generalization, while augmentation with the direction within 45 degrees of the principal adversarial direction boosts the adversarial accuracies and brings the natural performance close to state-of-the-art.

4.6. Robustness Against White-Box Attacks

We evaluate the PDA-AT framework's test accuracies for the adversarial samples and baseline methods against the following white-box attacks: FGSM (Goodfellow et al., 2014), PGD (Madry et al., 2018), and Carlini and Wagner (CW) attack (Carlini & Wagner, 2017), as well as black-box variants. We report the accuracy of the natural test images described as natural. Iterative attacks are denoted by PGD*T* and CW*T*, where T is the attack iteration. The attack step size in iterative attacks is set to $2/255$ and $8/255$ for single-step attacks. All the attacks are randomly initialized. The adversarial budget is set to $\epsilon = 8/255$. We summarize the natural and robust accuracies for the CIFAR-10, CIFAR-100, and SVHN datasets in Table 2. In the table, the proposed approaches are denoted by $\boldsymbol{\delta}_{\mathrm{OT}}$ and $\boldsymbol{\delta}_{\mathrm{OT,rand,45}}$ represent AT with the single-step attack given in Equation 4, and AT with perturbation augmentation obtained by Equation 6, respectively.

Performances in the cells with hyphen symbols in Table 2 are not available. The baseline results without a start sign are the reported results in their corresponding studies. The selected baseline studies do not report results for each test presented in Table 2. Some of the missing results were obtained by training the baseline approach from scratch. However, we prefer to compare the PDA-AT's performance with the original results reported by the studies to reduce the changes in baseline performances due to issues such as random initialization. As a result, unavailable results are denoted by a hyphen in Table 2.

In Table 2, when Feature Scattering (Zhang & Wang, 2019) line is compared with PDA-AT with $\boldsymbol{\delta}_{\mathrm{OT}}$, denoted by PDA-AT-$\boldsymbol{\delta}_{\mathrm{OT}}$**,** we can conclude that using perturbations altering the distribution of the mini-batch rather than increasing the gap between mini-batch distributions of the natural and adversarial samples can boost the robustness in CIFAR-10 and SVHN datasets. On the other hand, PDA-AT-$\boldsymbol{\delta}_{\mathrm{OT}}$ without perturbation augmentation does not contribute to the robustness of CIFAR-100. We discuss that the change in this behavior might be due to the higher number of classes. When the number of classes increases from 10 to 100, perturbing different mini-batches at each iteration might not sufficiently generalize. Moreover, it is evident in Table 2 that the proposed perturbation augmentation significantly improves adversarial and natural accuracies compared with the baselines. Specifically, when comparing the proposed method with the best baseline performance, we can observe an increment of 13.14%, 10.28%, and 5.82% in CIFAR-10, CIFAR-100, and SVHN accuracies against the FGSM attack, respectively.

Performances for Tiny-ImageNet and GTSRB datasets are provided in Table 3. Since performances for the two datasets are not reported by all the baselines, we present their results separately by comparing them with the Feature Scattering model trained from scratch using the implementation in its GitHub repository (HaicHao, 2019). All models used in this comparison were trained with the same training and attack configurations as the other datasets. Since the Tiny-ImageNet dataset is quite challenging, the discrepancy between the robustness provided by Feature Scattering and the proposed method is not significant. However, the improvement in the robustness of GTSRB can be seen in Table 3.

4.7. Results Against Black-Box Attacks

The PDA-AT's performance in transfer-based black-box attack settings is also evaluated (Papernot et al., 2017). For this purpose, the models trained with only natural and adversarial samples with PGD attacks (Madry et al., 2018) are used to generate the gradient-based black-box attacks. The AT model has trained with PGD7. Feature Scattering and the proposed model, PDA-AT, are evaluated with a set of PGD and CW attacks created in the black-box setting. Feature Scattering (Zhang & Wang, 2019) models are trained from scratch using the code provided in (Zhang, 2019). As seen in Table 4 and Table 5, the robustness of the proposed method can generalize against black-box attacks and outperforms the Feature Scattering AT.

Table 2. *Adversarial accuracy comparison of baselines and PDA-AT under white-box attacks. The star, *, symbol indicates result of training from scratch. The dagger, †, symbol denotes results taken from Baytaş and Deb (2023). Other values are the best results reported by the baselines*

CIFAR-10								
Defenses	Natural	FGSM	PGD10	PGD20	PGD100	CW10	CW20	CW100
Natural*	96.08	34.7	0.01	0.00	0.00	0.01	0.00	0.00
Madry et al. (2018) †	87.25	62.64	47.33	45.91	45.29	-	46.99	46.54
Bilateral	91.00	70.70	63.00	57.80	55.20	-	56.20	53.80
Feature Scattering (Zhang & Wang, 2019)	90.0	78.40	70.90*	70.50	68.60	62.60*	62.40	60.60
Adv-Interp (Zhang & Xu, 2020)	90.30	78.00	-	73.50	73.00	-	69.70	68.70
AV-Mixup (Lee et al., 2020)	93.24	78.25	62.67	58.23	-	53.63	-	-
PDA-AT-$\boldsymbol{\delta}_{\mathrm{OT}}$	90.24	78.11	73.97	72.29	71.10	62.11	59.66	57.52
PDA-AT-$\boldsymbol{\delta}_{\mathbf{OT,rand,45}}$	92.83	91.54	74.16	71.83	66.74	64.35	60.60	53.85
CIFAR-100								
Natural*	79.78	5.57	0.00	0.00	0.00	0.00	0.00	0.00
Madry et al. (2018) †	59.78	32.70	23.49	22.78	22.44	-	23.05	22.87
Bilateral	66.20	31.30	-	-	22.40	-	-	20.00
Feature Scattering (Zhang & Wang, 2019)	73.90	61.00	47.60*	47.20	46.20	30.76*	34.60	30.60
Adv-Interp (Zhang & Xu, 2020)	73.60	58.30	-	41.00	40.20	-	32.40	31.2
AV-Mixup (Lee et al., 2020)	74.81	62.76	-	38.49	-	-	-	-
PDA-AT-$\boldsymbol{\delta}_{\mathrm{OT}}$	71.80	49.31	41.50	40.83	25.08	23.18	22.48	24.46
PDA-AT-$\boldsymbol{\delta}_{\mathbf{OT,rand,45}}$	75.73	73.04	49.19	48.83	48.77	35.16	34.64	33.86
SVHN								
Natural*	96.33	45.29	1.59	0.62	0.39	1.04	0.63	0.48
Madry et al. (2018) †	90.74	64.50	44.23	41.38	40.37	-	42.46	41.60
Bilateral	94.10	69.80	-	53.90	50.30	-	-	48.90
Feature Scattering (Zhang & Wang, 2019)	96.20	83.50	55.40*	62.90	52.00	58.18*	61.30	50.80
Adv-Interp (Zhang & Xu, 2020)	94.10	81.83	-	-	-	-	-	-
AV-Mixup (Lee et al., 2020)	95.59	91.51	37.97	-	-	-	-	-
PDA-AT-$\boldsymbol{\delta}_{\mathrm{OT}}$	96.31	95.78	71.74	66.38	55.43	67.35	60.75	48.43
PDA-AT-$\boldsymbol{\delta}_{\mathbf{OT,rand,45}}$	97.15	97.33	72.48	65.02	50.61	65.13	56.58	42.45

Table 3. *The performance comparison for Tiny-ImageNet and GTRSB datasets.*

Tiny-ImageNet								
Defenses	Natural	FGSM	PGD10	PGD20	PGD100	CW10	CW20	CW100
Natural	69.04	2.28	0.02	0.01	0.00	0.00	0.00	0.00
Feature Scattering (Zhang & Wang, 2019)	57.42	24.42	12.54	11.25	10.33	9.11	7.97	7.55
PDA-AT-$\boldsymbol{\delta_{\mathbf{OT,rand,45}}}$	57.34	25.31	12.93	11.73	11.24	11.14	10.31	9.97
GTSRB								
Natural	98.62	63.34	19.94	14.31	11.29	17.29	13.22	11.36
Feature Scattering (Zhang & Wang, 2019)	96.5	91.81	87.82	84.56	78.52	76.01	72.02	67.42
PDA-AT-$\boldsymbol{\delta_{\mathbf{OT,rand,45}}}$	97.01	94.86	90.46	88.27	82.66	78.56	74.76	69.35

Table 4. *Results against black-box attacks transferred from the naturally trained model on CIFAR-10.*

		Defenses	
		Feature Scattering	**PDA-AT**
Attacks	**PGD20**	89.42	91.94
	PGD100	89.33	92.14
	CW20	89.41	91.87
	CW100	89.34	91.88

Table 5. *Results against black-box attacks transferred from standard AT on CIFAR-10.*

		Defenses	
		Feature Scattering	**PDA-AT**
Attacks	**PGD20**	76.60	78.40
	PGD100	76.46	77.94
	CW20	77.71	78.86
	CW100	77.44	78.76

5. DISCUSSION

The performance of the proposed PDA-AT is investigated from qualitative and quantitative perspectives. Figure 3 indicates that the adversarial direction augmentation improves the attack diversity during training. We also observe that the attack diversity slightly decreases as the training epochs proceed. In Tables 2 and 3, we can see that the augmented attack diversity results in improved robustness and generalization. The proposed PDA-AT does not sacrifice the natural accuracy compared with the baseline robust training approaches. The PDA-AT substantially improves the robustness against the FGSM attack. Adversarial accuracy might decrease for CW attacks compared to the robustness against PGD attacks for some datasets. This result shows that there is still room to alleviate overfitting certain types of attacks. On the other hand, Tiny-ImageNet and GTSRB results in Table 3 show that the PDA-AT can maintain the adversarial performance against the CW attacks better than the Feature Scattering AT. Finally, the adversarial defense techniques are notoriously prone to gradient masking (Papernot, et al., 2017). Robustness against black-box attacks in Tables 4 and 5 is reported particularly to investigate the gradient masking effect. When gradient masking occurs, adversarially trained models perform well against white-box attacks but cannot stand a black-box attack (Papernot et al., 2017). Tables 4 and 5 show that the PDA-AT can maintain its robustness against black-box attacks better than the model trained with Feature scattering AT for the CIFAR10 dataset. Thus, the proposed AT approach does not demonstrate a gradient masking effect (Papernot, et al., 2017).

6. CONCLUSION

This study proposes an adversarial training approach, PDA-AT, with adversarial direction augmentation during training. The PDA-AT employs alternative adversarial directions within a region spanning a 45-degree angle from the principal adversarial direction determined by the gradient of the optimal transport distance between the mini-batch samples. The principal adversarial direction generation differs from the Feature Scattering adversarial training, where the perturbation aims to increase OT distance between a mini-batch of adversarial samples and randomly perturbed natural samples. We empirically show that conditioning the adversarial direction to perturb the mini-batch distribution to differ from random perturbation hampers the attack diversity during training. Therefore, the proposed approach obtains adversarial directions that alter the natural sample distribution with random initializations. The extensive experimental results of the benchmark datasets, CIFAR-10, CIFAR-100, SHVN, GTSRB, and Tiny-ImageNet, against well-known white-box and black-box attacks provide empirical evidence that the PDA-AT improves adversarial robustness and augmenting the adversarial directions further boosts the adversarial accuracy.

AUTHOR CONTRIBUTIONS

D. Serbes and İ. M. Baytaş wrote the manuscript, İ. M. Baytaş developed the approach, D. Serbes conducted the experiments, and D. Serbes and İ. M. Baytaş interpreted the results.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Alzantot, M., Sharma, Y., Elgohary, A., Ho, B., Srivastava, M., & Chang, K. (2018). *Generating natural language adversarial examples*. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, (pp. 2890–2896).

Andriushchenko, M., & Flammarion, N. (2020). *Understanding and improving fast adversarial training.* In: Proceedings of Advances in Neural Information Processing Systems, *33*, (pp. 16048-16059).

Athalye, A., Carlini, N., & Wagner, D. (2018, July). *Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples.* In: International Conference on Machine Learning (pp. 274-283).

Baytaş, İ. M., & Deb, D. (2023). Robustness-via-synthesis: Robust training with generative adversarial perturbations. *Neurocomputing*, *516*, 49-60. https://doi.org/10.1016/j.neucom.2022.10.034

Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., ... & Zhou, W. (2016). *Hidden voice commands.* In: 25th USENIX security symposium (USENIX security 16), (pp. 513-530).

Carlini, N., & Wagner, D. (2017, May). *Towards evaluating the robustness of neural networks.* In: Proceedings of the IEEE Symposium on Security and Privacy. (pp. 39-57).

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, *26*.

Etmann, C., Lunz, S., Maass, P., & Schönlieb, C. B. (2019). *On the connection between adversarial robustness and saliency map interpretability.* In: Proceedings of the 36th International Conference on Machine Learning, 97, (pp. 1823-1832).

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, *363*(6433), 1287-1289. https://doi.org/10.1126%2Fscience.aaw4399

Fursov, I., Morozov, M., Kaploukhaya, N., Kovtun, E., Rivera-Castro, R., Gusev, G., ... & Burnaev, E. (2021). *Adversarial attacks on deep models for financial transaction records.* In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, (pp. 2868-2878).

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples.* In: Proceedings of the 3th International Conference on Learning Representations. https://arxiv.org/abs/1412.6572

Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., & Igel, C. (2013, August). *Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark.* In: Proceedings of the 2013 International Joint Conference on Neural Networks, (pp. 1-8).

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). *Adversarial examples are not bugs, they are features*. In: Proceedings of Advances in Neural Information Processing Systems, 32.

Jang, Y., Zhao, T., Hong, S., & Lee, H. (2019). *Adversarial defense via learning to generate diverse attacks.* In: Proceedings of the IEEE/CVF International Conference on Computer Vision, (pp. 2740-2749).

Kim, H., Lee, W., & Lee, J. (2021). *Understanding catastrophic overfitting in single-step adversarial training*. In: Proceedings of the AAAI Conference on Artificial Intelligence, (pp. 8119-8127).

Krizhevsky, A., & Hinton, G. (2009). *Learning multiple layers of features from tiny images*. University of Toronto.

Kurakin, A., Goodfellow, I. J. & Bengio, S. (2017). *Adversarial machine learning at scale*. In: Proceedings of the 5th International Conference on Learning Representations. https://arxiv.org/abs/1611.01236

Le, Y., & Yang, X. (2015). *Tiny imagenet visual recognition challenge*. CS 231N, 7(7), 3.

Lee, S., Lee, H., & Yoon, S. (2020). *Adversarial vertex mixup: Toward better adversarially robust generalization.* In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (pp. 272-281).

Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. (2018). *Towards deep learning models resistant to adversarial attacks.* In: Proceedings of the International Conference on Learning Representations. https://arxiv.org/abs/1706.06083

Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). *Deepfool: A Simple and accurate method to fool deep neural networks.* In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (pp. 2574-2582).

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017, April). *Practical black-box attacks against machine learning.* In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. (pp. 506-519).

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., & Madry, A. (2018). *Adversarially robust generalization requires more data.* In: Proceedings of Advances in Neural Information Processing Systems, (pp. 5019-5031).

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., ... & Goldstein, T. (2019). *Adversarial training for free!*. In: Proceedings of Advances in Neural Information Processing Systems, 32.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing properties of neural networks.* In: Proceedings of International Conference on Learning Representations. http://arxiv.org/abs/1312.6199

Tramer, F., Kurakin, A., Papernot, N., Goodfellow, I. J., Boneh, D. & McDaniel, P. D. (2018). *Ensemble adversarial training: Attacks and defenses.* In: Proceedings of the 6th International Conference on Learning Representations. https://arxiv.org/abs/1705.07204

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., & Bengio, Y. (2019, May). *Manifold mixup: Better representations by interpolating hidden states.* In: International Conference on Machine Learning, (pp. 6438-6447).

Villani, C. (2009). *Optimal transport: old and new* (Vol. 338, p. 23). Berlin: Springer.

Wang, J., & Zhang, H. (2019). *Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks.* In: Proceedings of the IEEE/CVF International Conference on Computer Vision, (pp. 6629-6638).

Wang, K., Li, F., Chen, C. M., Hassan, M. M., Long, J., & Kumar, N. (2021). Interpreting adversarial examples and robustness for deep learning-based auto-driving systems. *IEEE Transactions on Intelligent Transportation Systems*, *23*(7), 9755-9764. https://doi.org/10.1109/TITS.2021.3108520

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., & Gu, Q. (2019, September). *Improving adversarial robustness requires revisiting misclassified examples.* In: Proceedings of International Conference on Learning Representations. https://openreview.net/forum?id=rklOg6EFwS

Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., & Yan, S. (2023). *Better diffusion models further improve adversarial training.* In: Proceedings of the 40th International Conference on Machine Learning. (202:36246-36263) https://proceedings.mlr.press/v202/wang23ad.html

Wong, E., & Kolter, Z. (2018, July). *Provable defenses against adversarial examples via the convex outer adversarial polytope.* In: Proceeding of International Conference on Machine Learning, (pp. 5286-5295).

Wong, E., Rice, L., Kolter, J. Z. (2020). *Fast is better than free: Revisiting adversarial training.* In: Proceedings of the 8th International Conference on Learning Representations. https://arxiv.org/abs/2001.03994

Xie, Y., Wang, X., Wang, R., & Zha, H. (2020, August). *A fast proximal point method for computing exact Wasserstein distance*. In: Proceedings of Uncertainty in Artificial Intelligence (pp. 433-453).

Yuval, N., Wang, T., Coates, A., Bissacco, A., Wu, B., & Ng, A. Y. (2011). *Reading digits in natural images with unsupervised feature learning.* In: Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning.

Zagoruyko, S., & Komodakis, N. (2016) *Wide residual networks.* In: Proceedings of the British Machine Vision Conference. (pp. 1-12).

Zhang, D., Zhang, T., Lu, Y., Zhu, Z., & Dong, B. (2019). *You only propagate once: Accelerating adversarial training via maximal principle.* In: Proceedings of Advances in Neural Information Processing Systems, 32.

Zhang, H., Cisse, M., Dauphin, Y. N., Lopez-Paz, D. (2018). *Mixup: Beyond empirical risk minimization.* In: Proceedings of the 6th International Conference on Learning Representations. https://arxiv.org/abs/1710.09412

Zhang, H., & Xu, W. (2020). Adversarial interpolation training: A simple approach for improving model robustness. https://openreview.net/forum

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., & Jordan, M. (2019). *Theoretically principled trade-off between robustness and accuracy*. In: Proceedings of the International Conference on Machine Learning, (pp. 7472-7482).

Zhang, H., & Wang, J. (2019). *Defense against adversarial attacks using feature scattering-based adversarial training*. In: Proceedings of the Advances in Neural Information Processing Systems, 32.

Zhang, H. (2019). Feature Scattering Adversarial Training (NeurIPS 2019) (Accessed: 24/03/2024) https://github.com/Haichao-Zhang/FeatureScatter