Enhancing Image Classification Robustness through Adversarial Sampling with Delta Data Augmentation (DDA)

Ivan Reyes-Amezcua
CINVESTAV
Guadalajara, Mexico.
ivan.reyes@cinvestav.mx

Gilberto Ochoa-Ruiz
Tecnologico de Monterrey
Guadalajara, Mexico
gilberto.ochoa@tec.mx

Andres Mendez-Vazquez
CINVESTAV
Guadalajara, Mexico
andres.mendez@cinvestav.mx

Abstract

Deep learning models are susceptible to adversarial attacks, highlighting the critical need for enhanced adversarial robustness. Recent studies have shown that minor alterations to the input can significantly affect the model's prediction accuracy, making it prone to such attacks. In our study, we present the Delta Data Augmentation (DDA) technique, a novel approach to improving transfer adversarial robustness by using perturbations derived from models trained to resist adversarial threats. Unlike conventional methods that attack the model directly, our approach sources adversarial perturbations from higher-level tasks and integrates them into the training of new tasks. This strategy aims to increase both the robustness and the adversarial diversity of the datasets. Through extensive empirical testing, we showcase the superiority of our data augmentation strategy over existing leading methods in enhancing adversarial robustness. This is particularly evident in our evaluations using Projected Gradient Descent (PGD) attacks with l_2 and l_∞ norms on datasets such as CIFAR10, CIFAR100, SVHN, MNIST, and FashionMNIST.

1. Introduction

The research in the field of adversarial robustness for deep learning aims to increase the robustness of models to adversarial attacks [3, 6, 10]. These attacks are deliberate attempts to trick a model by purposefully introducing undetectable perturbations to the input data, leading the algorithm to misclassify or make inaccurate predictions [1, 8]. Applications like autonomous vehicles[34], medical diagnosis[2, 33], and fraud detection [7] are all susceptible to adversarial attacks, which can have major repercussions. Thus, it has become crucial to conduct research on increasing the adversarial robustness of deep learning models in order to make such systems safe and useful in real settings.

Due to deep neural network's susceptibility to adversarial attacks, adversarial robustness has grown to be a crucial research area in deep learning (DL) [1, 4, 9, 15]. As matter of fact, it has been demonstrated again and again that DL models are susceptible to adversarial instances, which are intentionally constructed inputs that can lead the model to produce wrong predictions[24]. Although several proposals for mitigating adversarial risks for DL models have been investigated [20, 23], there is still a need for enhanced robustness in many settings.

Delta Data Augmentation (DDA) (Fig. 1), our proposal, aims to handle the crucial problem of transfer robustness in deep learning and improve adversarial robustness. When there is a lack of labeled data, our approach to transfer robustness requires only a pre-trained model from one dataset, which we then apply, along with the acquired knowledge, to another model and dataset. By incorporating perturbations sampled from trained models that are resistant to adversarial attacks, DDA is designed to improve transfer robustness. The proposed DDA method focuses on collecting adversarial perturbations from upstream tasks rather than attacking the model directly. By using these perturbations in data augmentation for downstream tasks, the approach aims to enhance the adversarial diversity and robustness of the training datasets.

In a well-studied domain with existing datasets, researchers may have developed specialized techniques to enhance robustness, such as adversarial training [3, 15]. However, when dealing with brand-new datasets and there is a constraint in the training budget, these established methods may not directly apply, or they may require significant adaptation and fine-tuning. Adding these constraints usually involves the following: **i) Generating adversarial examples,** which entails creating input data that looks normal but is designed to fool the model; **ii) Adversarial training,** the model on both regular and adversarial examples, which requires additional computational power and time; and finally, **iii) Hyperparameter tuning,** which means fine-tuning the model's parameters to balance between natural accuracy (performance on clean data) and robust accuracy

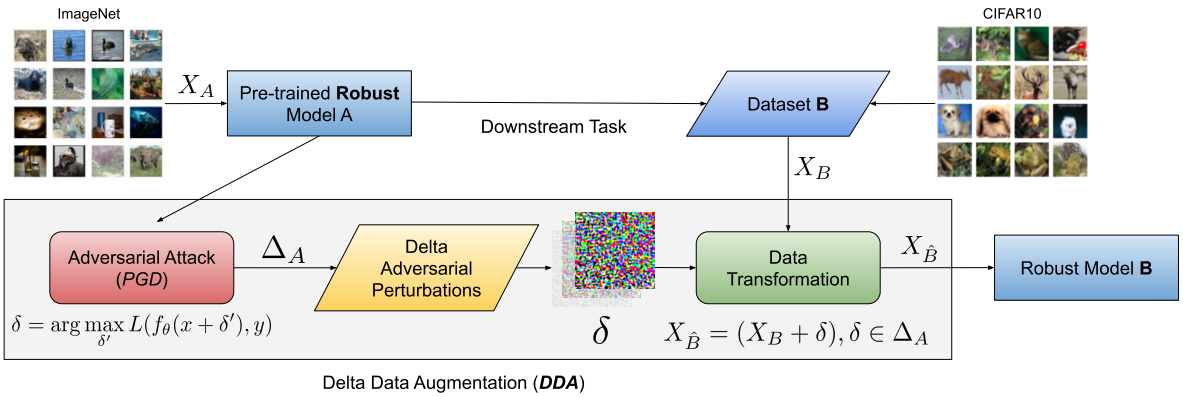


Figure 1. Overview of Delta Data Augmentation (DDA). A method for data augmentation to enhance adversarial robustness through sampling adversarial perturbations $\delta \in \Delta$ from upstream trained robust models (e.g. ImageNet) and applying them to downstream training datasets (e.g. CIFAR10). Our method benefits from extracting adversarial perturbations and applying them to training datasets to enhance robustness.

(performance on adversarial data).

The traditional adversarial training approach might fail with new datasets for several reasons, including a lack of prior knowledge of robust methods tailored to the specific data [18, 22], making effective training strategy development potentially prohibitively expensive. Additionally, striking the right balance between maintaining high accuracy on clean data and ensuring robustness against adversarial attacks can be particularly challenging when faced with unknown data characteristics and limited resources [32]. Therefore, reducing this gap between accuracy and robustness is still an open research problem [19, 29, 31]. On the other hand, DDA boosts adversarial robustness by using perturbations from models trained against adversarial attacks. Instead of direct model attacks, it gathers adversarial perturbations from complex tasks, integrating them into subsequent training and reducing the complexity of adversarial training for a specific dataset.

Given its design, our approach can incorporate samples that have been generated by the addition of perturbations of previous datasets, leading to more diverse training examples that can better reflect the heterogeneity of the target dataset. Compared to other approaches in the literature (Table 1), such as RandAugment[13], AutoAugment[12] and AugMix[17], DDA does not require additional labeled data or knowledge of the target dataset. Instead, it makes use of the robust model's acquired knowledge to produce perturbations that are pertinent to the target domain.

The contributions of this work are:

- A novel data augmentation method based on adversarial attacks and transfer learning to enhance model robustness downstream tasks.
- The implementation of a pipeline for adversarial defense method to include adversarial examples in training without performing adversarial training

Using perturbations from previous datasets, the DDA approach increases training variety. It outperforms RandAugment, AutoAugment, and AugMix by using the knowledge of a robust model without requiring more data. DDA's primary contributions combine adversarial attacks with transfer learning and create an efficient adversarial defensive system.

The rest of the paper is organized as follows: In Section 2, we describe the related work on the adversarial robustness problem. In Section 3, we describe the proposed method. We first explain the adversarial training procedure and then explain the generation of adversarial perturbations for transfer robustness. Next, we discuss the design choices we made for DDA for sampling adversarial perturbations. In Section 4, we detail the experimental setup used for implementing the models. In Section 5, we discuss the performance obtained by using DDA. Finally, Section 6 presents our conclusions and discusses future work.

2. Related Work

The creation of reliable and accurate models is crucial in the fast-moving field of computer vision. However, the requirement for large, varied, and representative datasets is one of the main difficulties faced by researchers and practitioners when training such models. The idea of data augmentation comes into play here. Data augmentation is a potent approach that, without the need for further data collection, artificially increases the size and diversity of a dataset. In addition to increasing the amount of training data, this exposes the model to a greater variety of variances, which aids in its ability to generalize.

Finding the ideal augmentation strategy is still an open research subject. The choice of augmentation strategies, their parameters, and their application can considerably affect how

well a model can perform. For instance, the work in [12] introduces a method known as AutoAugment. This method automatically seeks better data augmentation strategies tailored to image classification tasks. A search algorithm is used in this approach to identify the best policy, which is composed of sub-policies that perform image processing operations such as translation, rotation, and shearing. RandAugment is another recently proposed DA augmentation approach [13]; this method introduces an automated data augmentation strategy that can be applied across different datasets and tasks, without additional search steps. Similarly, AugMix [17] is designed to improve the robustness and uncertainty of image classifiers. This technique makes use of a wide range of augmentation operations, such as translation, rotation, and shearing, that are drawn from AutoAugment's large library. AugMix stands out due to its "mixing" technique, which combines several augmentation chains in a weighted way to provide augmented images that are more thorough and varied than a single training set.

These methods of advanced data augmentation have shown promise in enhancing model robustness against unpredictable data shifts. However, despite these advancements, the challenge of achieving adversarial robustness with data augmentation techniques remains. Therefore, data augmentation still plays a pivotal role in the research for truly robust and reliable models.

3. Methodology

In deep learning, the term *adversarial robustness* describes a model's capacity to continue operating effectively even in the minimum engineered changes to the input data intended to trick the model [4]. Let X be the set of possible input data, and let Y be the set of possible output labels. A supervised learning model can be represented as a function $f : X \to Y$ that maps input data to output labels. Adversarial examples can be generated by adding a small perturbation δ to the input data x, such that $x' = x + \delta$. The perturbation is typically constrained to have a small ℓ_p-norm, where p is a positive integer (e.g., $p = 2$ corresponds to the Euclidean distance).

The utilization of adversarial examples as a means of data augmentation during the training phase constitutes a technique referred to as *Adversarial Training*. This technique aims to enhance the robustness of deep learning models against adversarial examples.

3.1. Adversarial Examples

Let $x \in X$ be an input data vector and $y \in Y$ be its corresponding label. The loss function is typically defined as the cross-entropy loss between the predicted output of the model and the true label (Eq. 1).

$$L(f_\theta(x), y) = -\sum_{i=1}^{|Y|} y_i \log f_\theta(x)_i, \tag{1}$$

where $f\theta(x)_i$ is the i-th output of the model for input x.

To generate an adversarial perturbation δ for input x, the first step is to compute the δ that maximizes the loss function (Eq. 2), subject to a constraint on the ℓ_p-norm of the perturbation, such that $|\delta|_p \leq \epsilon$. In adversarial attacks, ϵ is a parameter used to define a constraint on the magnitude of the perturbation that can be applied to the input data x.

$$\delta = \arg\max_{\delta'} L(f_\theta(x + \delta'), y), \text{s.t.} |\delta|_p \leq \epsilon \tag{2}$$

One approach to generating adversarial examples is to use an iterative optimization algorithm such as the Fast Gradient Descent Method (FGSM) [15] or Projected Gradient Descent (PGD) [24] to compute a perturbation.

The resulting adversarial example $x + \delta$ is then added to the original training data and its corresponding label y, creating an augmented training dataset. Then, the empirical risk over the augmented training data is used as the final objective function for adversarial training, as shown in Equation 3.

$$\min_\theta \frac{1}{m} \sum_{i=1}^{m} L(f_\theta(x_i + \delta_i), y_i), \tag{3}$$

where m is the size of the training data, and $(x_i + \delta_i, y_i)$, are pairs of adversarial training examples.

Additionally, the accuracy under adversarial attacks (Robust Accuracy) is the most commonly used metric to evaluate the robustness of a model [24]. This metric determines the percentage of correctly classified examples under a particular attack. Similarly, the robustness radius is a metric that measures the maximum magnitude of adversarial perturbation that a model can withstand [10]. Moreover, minimum distortion is a metric that assesses the minimum magnitude of adversarial perturbation needed to fool a model [5].

3.2. Delta Data Augmentation (DDA)

Transfer Learning (TL) is a technique used in deep learning to transfer knowledge learned from one model to another [31]. In TL, a pre-trained model is used as a starting point for a new model rather than beginning from scratch [19]. For this, $g_\phi : X' \to Y'$ is a pre-trained model parameterized by ϕ, where X' and Y' may or may not be the same as X and Y. The goal of transfer learning is to initialize the parameters of f_θ using the pre-trained parameters ϕ and then fine-tune f_θ using a small amount of data from the new task [30]. Then, *transfer robustness* of g_ϕ is defined as the ability of f_θ to maintain its performance on a new task under adversarial attacks when initialized with the pre-trained parameters ϕ

Algorithm 1 Delta Data Augmentation

Require: Pre-trained Robust Model M_A, Dataset D, Length of adversarial samples k
Ensure: Augmented Dataset $\hat{D}$
1: Attack model M_A with PGD
2: Select the images that fooled model M_A
3: Sample k effective adversarial images
4: Extract in $\Delta \leftarrow k$ the effective perturbations
5: Resize Δ for D image size
6: **for** $x_i \in D_{train}$ **do**
7: Randomly select a perturbation $\delta \sim \Delta$
8: Apply perturbation on original image $\hat{x_i} \leftarrow x_i + \delta$
9: **end for**

[26, 30]. Now, instead of using pre-trained parameters, we look for transfer adversarial perturbations that are effective on a larger and more complex model.

Following this notion, a model may enhance its performance by incorporating a greater variety of data into the training phase [28]. The augmentation of training data via adversarial examples can result in an improvement in model generalization. Nevertheless, adversarial training is a computationally demanding and time-consuming undertaking. One approach to address the challenges of adversarial training is the use of universal adversarial perturbations [9, 25].

These perturbations can be generated once and applied to any image, which makes the process more efficient compared to generating adversarial examples for each image individually. Incorporating such perturbations into the training data can enhance model robustness and improve its generalization performance [28]. However, generating these perturbations can also be a computationally demanding task.

Instead of attacking a model to create a set of adversarial examples, we propose to gather adversarial perturbations by attacking upstream model tasks (e.g. ImageNet [14]). This approach will yield sample adversarial noise that is effective across other models. In our proposed training pipeline (Alg. 1), we aim to collect adversarial noise δ and apply it to downstream tasks in a data augmentation fashion. We call this method *Delta Data Augmentation* (DDA) (Fig. 1). In DDA, a pre-trained robust model that is trained on an upstream task, such as ImageNet Classification, is used to sample adversarial perturbations δ given an adversarial attack. The objective of this process is to obtain a representative sample of perturbations that reflects the same underlying structure, which can be used to make downstream training datasets more adversarially diverse and thus more robust.

We evaluated our Delta Data Augmentation (DDA) method's efficacy using a shortened evaluation procedure. This involved creating adversarial perturbations δ from a robust upstream model, using these on downstream datasets, and contrasting baseline models' robust and natural accuracies before and after DDA. We also compared DDA's effectiveness with conventional augmentation methods and tested the improved DDA models against a range of adversarial approaches. With the least amount of typical trade-offs associated with adversarial training, this thorough examination attempted to demonstrate how DDA might improve model robustness against a variety of adversarial perturbations intensities.

Thus, DDA can collect and create more complex transformations on data rather than traditional techniques, (rotation, scaling, flipping, etc.). Furthermore, the training time used for data augmentation of model B is reduced due to the absence of adversarial training, and the gap between natural accuracy and robust accuracy is minimized as we avoid learning explicit adversarial perturbations, preventing overfitting to specific types of adversarial attacks.

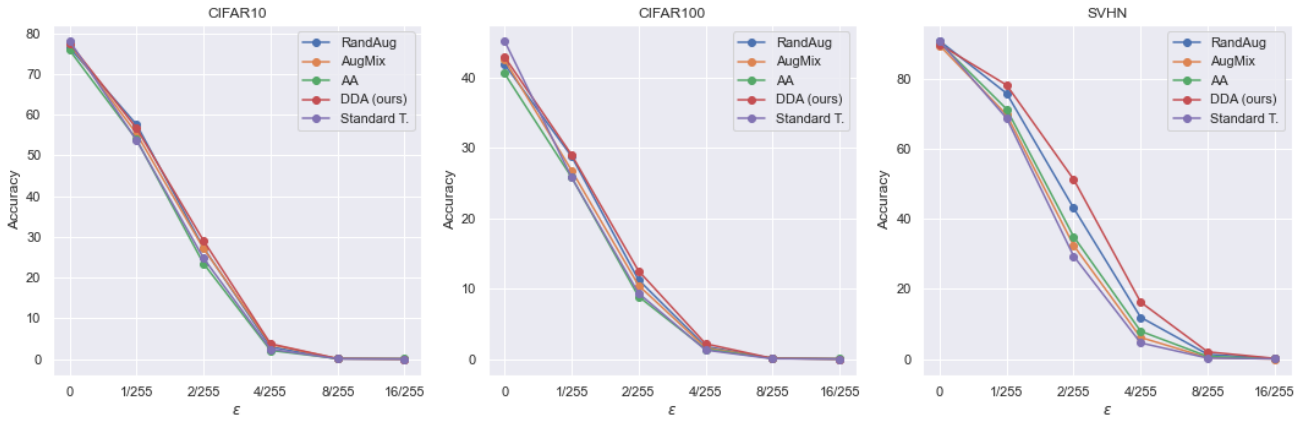
4. Model Configuration

We set up the models in our work to evaluate how well our Delta Data Augmentation (DDA) technique improves adversarial robustness. We used an architecture known as ResNet18, which is well-known for its effectiveness and performance on a range of image recognition applications. Using both L_2 and L_∞ norms with a variety of epsilon values: 0.01, 0.1, 0.2, 0.3, 0.5 for L_2 and 1/255, 2/55, 4/255, 8/255, 16/255 for L_∞, adversarial perturbations were applied to this model using Projected Gradient Descent (PGD) adversarial attacks.

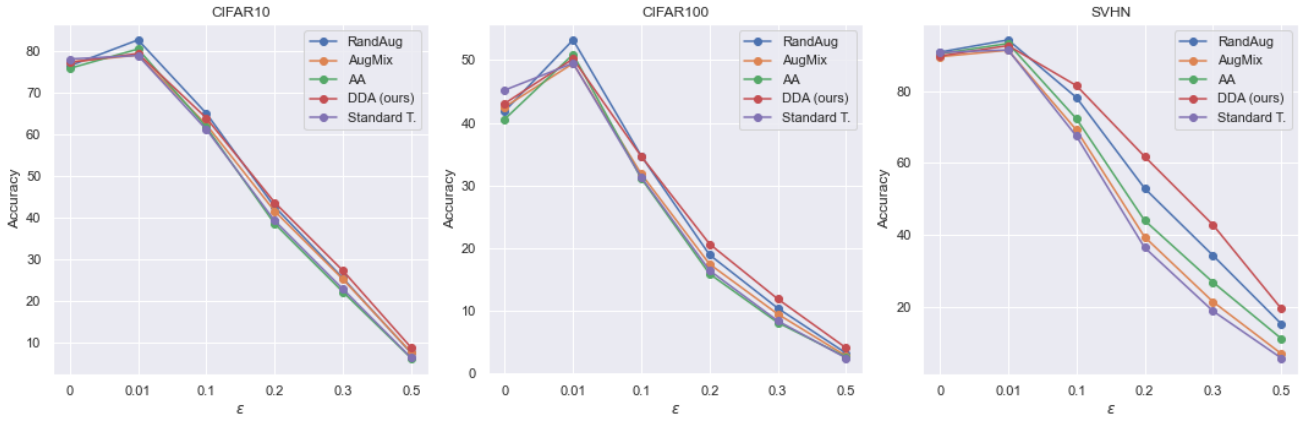
Then, to replicate a range of adversarial scenarios, these perturbations were applied to the training datasets of downstream tasks, such as CIFAR10, CIFAR100, SVHN for RGB datasets, and MNIST and FashionMNIST for single-channel datasets. Under the same training conditions, our DDA method was tested against popular data augmentation methods including RandAugment, AutoAugment, and AugMix, as well as a baseline scenario without any data augmentation. The goal of this comparison investigation was to show how much better DDA is at preparing models for fighting against adversarial attacks while preserving or improving their performance on common benchmark datasets.

5. Experimental Setup

In our Delta Data Augmentation (DDA) approach, we leverage the capabilities of PyTorch [27] for training *Model B*, as illustrated in Fig. 1. Specifically, we utilize a ResNet18 [16] architecture pre-trained on the ImageNet dataset [14] for this purpose. To enhance the robustness of *Model B*, we introduce perturbations extracted from an adversarially robust *Model A*. This model, proposed in [30] and implemented in the RobustBench python package [11], has been shown to be effective in defending against adversarial attacks, particularly the PGD attack, despite its slightly lower accuracy



(a) PGD l_∞ ResNet18



(b) PGD l_2 ResNet18

Figure 2. Natural accuracy ($\epsilon = 0$), and Robust accuracy ($\epsilon > 0$) results for PGD with l_∞ (2a) and l_2 (2b) on CIFAR10, CIFAR100 and SVHN datasets, compared with RandAugment, AutoAugment, AugMix, No Data Augmentation, and DDA (ours), trained with ResNet18.

on ImageNet. The authors of [30] have demonstrated that models trained to be robust against adversarial attacks often surpass their standard-trained counterparts in transfer learning tasks.

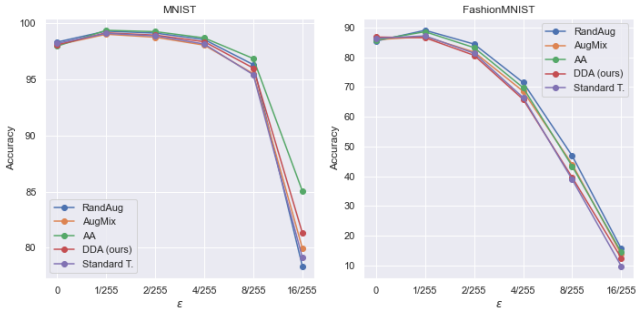
To integrate this robustness into *Model B*, we apply the methodology described by Eq. 4, where X_B denotes the training images for *Model B*, and δ_A represents the perturbations extracted from *Model A* under PGD attack conditions:

$$X'_B = X_B + \delta_A \tag{4}$$

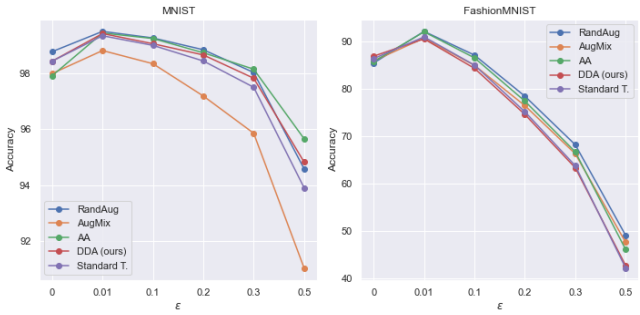
Here, X'_B denotes the augmented training images for *Model B*, which are obtained by combining the initial training images X_B with the adversarially determined perturbations δ_A. Through this approach, *Model B* is exposed to more data during its training phase and also gains a significant boost in resistance against adversarial attacks. As a result, its performance and robustness are improved in a way that is informed by *Model A*'s adversarial resilience.

We compare our method against three common and popular data augmentation techniques: RandAugment [13], AutoAugment [12] and AugMix [17]. Also, we developed a baseline on ResNet18 by standard training with no data augmentation at all. All models were trained with cross-entropy loss, and Adam optimizer with 0.001 as a starting learning rate and 30 epochs (Fig. 2).

Following Algorithm 1, in DDA, we determine the value of k by conducting a series of preliminary experiments aimed at optimizing the balance between the diversity of perturbations and computational efficiency. Through these experiments, we observed that a value of $k = 50$ perturbations strikes an optimal balance, providing a sufficiently diverse set of perturbations to enhance model robustness without imposing excessive computational demands. These k perturbations (k = 50 by default) are randomly sampled from δ_A from an upstream pre-trained model A trained on dataset X_A. Subsequently, we apply these perturbations to the downstream training dataset X_B for model B. This approach ensures that the downstream model B benefits from the ro-



(a) PGD l_∞ ResNet18 - Grayscale



(b) PGD l_2 ResNet18 - Grayscale

Figure 3. Experiments on single-channel datasets (i.e. grayscale) for MNIST and FashionMNIST. Accuracy results for PGD with l_∞ (3a) and l_2 (3b) compared with RandAugment, AutoAugment, AugMix, No Data Augmentation, and DDA (ours), trained with ResNet18.

bustness characteristics of the upstream model A, thereby improving its generalization capability on dataset X_B.

DDA is designed to extract perturbations from RGB datasets, specifically from ImageNet dataset. To test DDA on single-channel datasets (i.e. grayscale images), we preprocess the extracted perturbations set Δ to only have one channel by getting the mean of the three channels in RGB space. From (Eq. 4) we take the mean of the RGB channels for each $\delta \in \Delta$ as follows:

$$X_B^{grayscale'} = X_B^{grayscale} + mean_{RGB}(\delta_A) \tag{5}$$

Then we resize each of the $delta$ perturbations for each dataset (MNIST and FashionMNIST). In (Fig. 3) we present extensive experimentation of DDA and state-of-the-art data augmentation methods, with the same configuration as RGB datasets.

6. Results

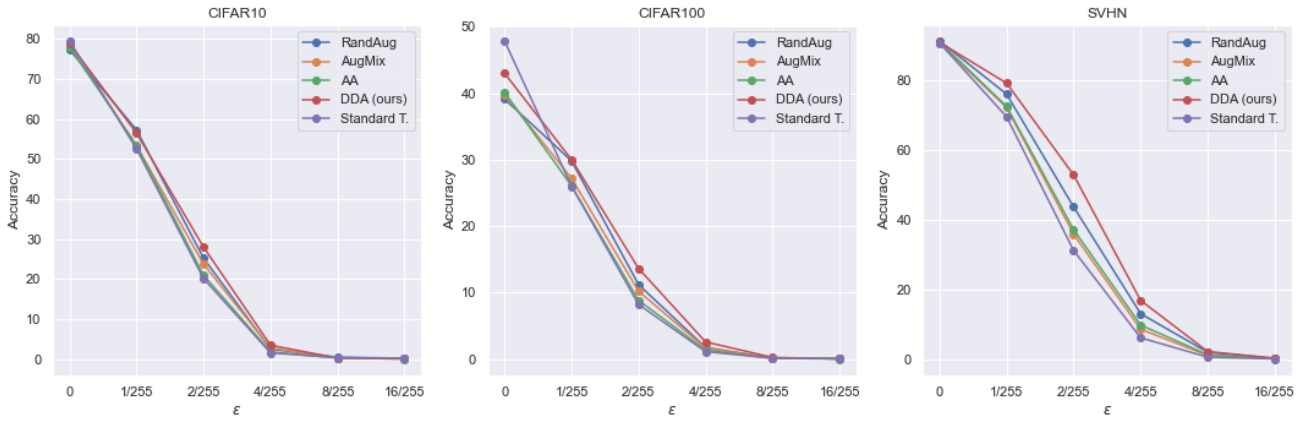
We test our method (Table 1) on the CIFAR10, CIFAR100, and SVHN datasets with natural accuracy and robust accuracy using Projected Gradient Descent (PGD) with 40 iterations. For all results, we report the robust accuracy average and standard deviation on three runs of experiments. These attacks are performed on ℓ_2-norm and ℓ_∞-norm using the implementations of *torchattacks* [21].

The comparison of accuracy across various data augmentation techniques reveals that DDA performs better than the others in terms of robust accuracy against PGD with several ϵ for l_∞ and l_2 Particularly, DDA outperforms other approaches in terms of robust accuracy, achieving values on ResNet18 of 8.94% and 3.73% for PGD attack with $\epsilon = 0.5$ for l_2 and $\epsilon = 4/255$ for l_∞ respectively. In particular, for the SVHN dataset, the results are remarkable, for ResNet18 and PGD with l_2 our method surpasses the state-of-the-art. DDA achieves an improvement with best-related method of 3.30%, 8.83%, 8.50% and 4.35% for $\epsilon \in \{0.1, 0.2, 0.3, 0.5\}$ respectively (Fig. 2b). Also, for SVHN and PGD with l_∞ and $\epsilon \in \{1/255, 2/255, 4/255, 8/255, 16/255\}$, our method beats the best-related method by 2.37%, 8.19%, 4.32%, 0.70%, and 0.05% respectively (Fig. 2a).

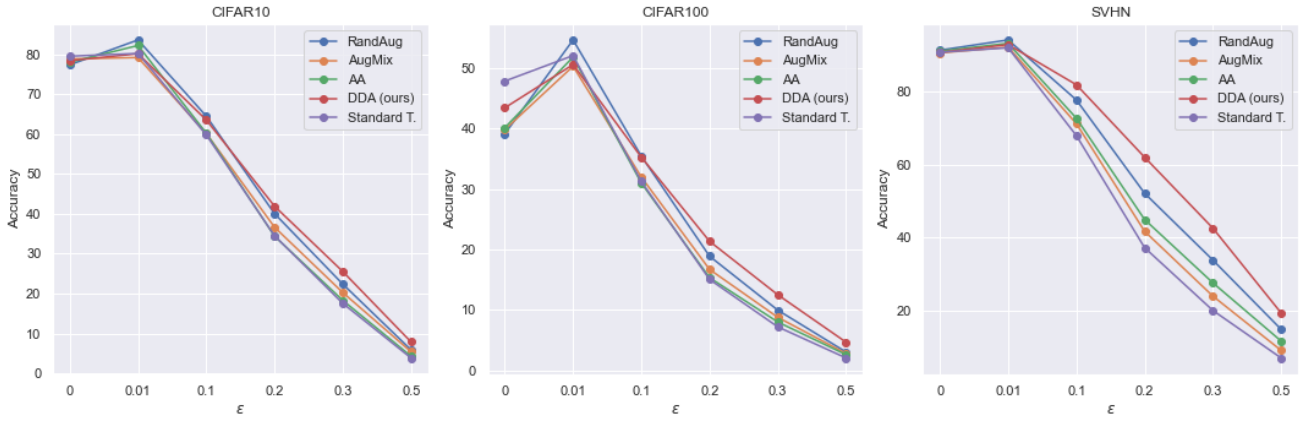
On the other hand, due to the design choices of DDA, it is not so successful when applied to non-RGB images.

PGD Norm	Dataset	No DA	RandAug	AutoAug	AugMix	DDA (ours)
$l_2 = 0.5$	CIFAR10	6.48±0.63%	7.64±0.54%	6.42±0.54%	7.79±0.97%	**8.94 ± 0.69%**
	CIFAR100	2.46±0.24%	3.23±0.25%	2.61±0.16%	2.84±0.34%	**4.06 ± 0.35%**
	SVHN	5.77±1.03%	3.23±0.25%	11.21±0.80%	6.99±0.67%	**19.55 ± 2.37%**
$l_\infty = 4/255$	CIFAR10	2.39±0.34%	2.95±0.32%	2.11±0.11%	3.55±0.74%	**3.73 ± 0.57%**
	CIFAR100	1.27±0.21%	1.78±0.11%	1.39±0.20%	1.60±0.26%	**2.17 ± 0.33%**
	SVHN	4.60±0.35%	11.91±1.23%	7.93±0.89%	6.16±0.27%	**16.23 ± 2.88%**

Table 1. Robust Accuracy (performance on adversarial data) results for ResNet18 under PGD adversarial attack with 40 iterations on $l_2 = 0.5$ and $l_\infty = 4/255$ for CIFAR10, CIFAR100 and SVHN datasets. We compare our method (DDA) with others in the state-of-the-art, such as RandAugment, AutoAugment, and AugMix. Also, "No DA" stands for a training method with no data augmentation at all. In bold are the best results, the standard deviation is also reported for 3 rounds of experiments.



(a) PGD l_∞ ResNet50



(b) PGD l_2 ResNet50

Figure 4. Ablation study for model size with ResNet50. Same datasets, methods and conditions used for ResNet18 training and evaluations (Fig. 3)

For MNIST and FashionMNIST datasets, DDA achieves competitive results although is not designed for this kind of single-channel datasets.

6.1. Ablation Studies

To prove DDA works and is independent of model size, we test our method on different configurations (Table 2). The main goal is to determine whether DDA's effectiveness is limited to a particular network size. To achieve this, we test our approach using the same condition settings shown in (Fig. 2), but with ResNet50 as the network architecture instead of ResNet18. Here, we show the comparison of DDA with a larger model (ResNet50). It can be seen the performance is similar across all datasets, showing that our method does not depend on the model size.

Similarly, we test our method on single-channel datasets with ResNet50 (Fig. 5). Although DDA is not significantly better than state-of-the-art, it has competitive results. We argue our method gets the most advantage on RGB datasets, due to the extraction of perturbations from RGB datasets (e.g. ImageNet). In other words, the transferring of adversarial robustness is not capturing well enough from the three-channel domains to single-channel domains. The extensive testing on various scenarios showed that DDA's performance holds across diverse network architectures.

Additionally, we used ResNet50 to evaluate the effectiveness of our technique on single-channel datasets. Although the findings weren't noticeably better than those of cutting-edge techniques, DDA nevertheless produced decent outcomes.

6.2. Discussion

The comparison of ResNet18 and ResNet50 and PGD attacks with various ϵ attack intensities demonstrates that the robustness of the model is significantly impacted by the choice of attack strength (Fig. 2, 4). As expected given that greater attacks bring larger perturbations that are more challenging to recover from, our results indicate that stronger

Model	Channels	Datasets	Image Size	PGD Norm
ResNet18	RGB	CIFAR10, CIFAR100, SVHN	32×32	l_2
ResNet18	RGB	CIFAR10, CIFAR100, SVHN	32×32	l_∞
ResNet50	Grayscale	MNIST, FashionMNIST	28×28	l_2
ResNet50	Grayscale	MNIST, FashionMNIST	28×28	l_∞

Table 2. Table showing the experimental configurations for ablation studies, detailing the model size, image channel, datasets used, image dimensions, and PGD norm settings.

attacks result in lower robust accuracies. The results, however, also indicate some space for improvement, particularly in the transfer of adversarial robustness from three-channel domains to single-channel domains, which did not seem to be effectively recorded. As a result, there is now a promising path for further study to improve the resilience and versatility of DDA. Overall, the findings imply that DDA is a successful technique for boosting the robustness of deep neural networks against adversarial attacks. The suggested method can be used to increase the robustness of models in a variety of applications and is simple to integrate into current training pipelines.

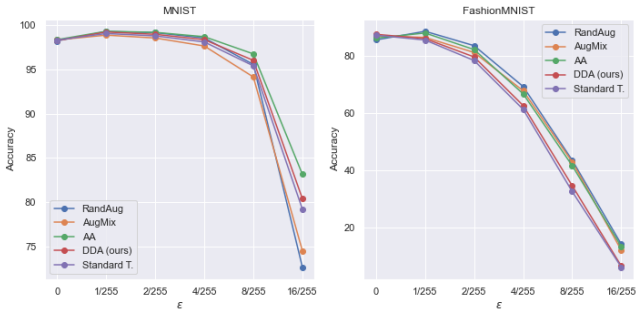
7. Conclusions and Future Work

In our research, we introduce Delta Data Augmentation (DDA), a technique for improving transfer robustness through the utilization of perturbations gathered from models trained to withstand adversarial attacks. Rather than attacking the model directly, this method collects adversarial perturbations from more complex tasks. By weaving these perturbations into the training process of later tasks, we aim to enhance both the robustness and adversarial diversity within the datasets.

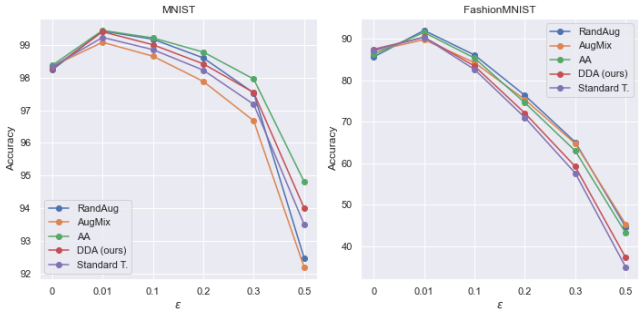
We compared a variety of data augmentation strategies, such as DDA, RandAugment, AutoAugment, AugMix, and Standard Training with No Data Augmentation, to examine the performance of various adversarial attack methods on the CIFAR10, CIFAR100, SVHN, MNIST, and FashionMNIST datasets. Our findings demonstrated that our approach performed better than or equal to state-of-the-art approaches in terms of adversarial robustness. DDA improves the transferability of robustness against adversarial attacks by reducing the gap between natural and robust accuracy.

Interestingly, with its unique approach and potential, DDA opens up avenues for further investigations of the extraction of adversarial perturbations for improving training datasets. Future studies can further enhance our findings by employing DDA on a broader spectrum of robust pre-trained models for extracting perturbations. This will provide a more comprehensive understanding of its performance capabilities against a diverse array of adversarial attacks. Additionally, it will be instructive to observe how DDA fared against increasingly sophisticated hostile attacks. This investigation may pave the way for improving the defense mechanisms of machine learning models, strengthening their resistance to novel adversarial dangers.

As a result of our research, it is possible to significantly improve the adversarial robustness of machine learning systems by including adversarial perturbations in training datasets, such as DDA.



(a) PGD l_∞ ResNet50 - Grayscale



(b) PGD l_2 ResNet50 - Grayscale

Figure 5. Ablation study on single-channel datasets (i.e. grayscale) for MNIST and FashionMNIST for model size with ResNet50. Same datasets, methods and conditions used for ResNet18 (Fig. 2)

Acknowledgements

This work has been supported by Azure Sponsorship credits granted by Microsoft's AI for Good Research Lab through the AI for Health program. The project was also supported by the French-Mexican ANUIES CONAHCYT Ecos Nord grant (MX 322537/FR M022M01).

References

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

[2] Kyriakos D Apostolidis and George A Papakostas. A survey on adversarial deep learning robustness in medical image analysis. *Electronics*, 10(17):2132, 2021.

[3] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021.

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[5] Nicholas Carlini, Guy Katz, Clark Barrett, and David L Dill. Provably minimally-distorted adversarial examples. *arXiv preprint arXiv:1709.10207*, 2017.

[6] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[7] Francesco Cartella, Orlando Anunciacao, Yuki Funabiki, Daisuke Yamaguchi, Toru Akishita, and Olivier Elshocht. Adversarial attacks for tabular data: Application to fraud detection and imbalanced data. *arXiv preprint arXiv:2101.08030*, 2021.

[8] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

[9] Ashutosh Chaubey, Nikhil Agrawal, Kavya Barnwal, Keerat K Guliani, and Pramod Mehta. Universal adversarial perturbations: A survey. *arXiv preprint arXiv:2005.08087*, 2020.

[10] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[11] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

[12] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[13] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[18] Matthew Holland. Robustness and scalability under heavy tails, without strong convexity. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 865–873. PMLR, 2021.

[19] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

[20] Yunseok Jang, Tianchen Zhao, Seunghoon Hong, and Honglak Lee. Adversarial defense via learning to generate diverse attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2740–2749, 2019.

[21] Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.

[22] Xiao Li, Ziqi Wang, Bo Zhang, Fuchun Sun, and Xiaolin Hu. Recognizing object by components with human prior knowledge enhances adversarial robustness of deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[23] Hongshuo Liang, Erlu He, Yangyang Zhao, Zhe Jia, and Hao Li. Adversarial attack and defense: A survey. *Electronics*, 11 (8):1283, 2022.

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[25] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[26] Awais Muhammad and Sung-Ho Bae. A survey on efficient methods for adversarial robustness. *IEEE Access*, 10:118815–118830, 2022.

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[28] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[29] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*, 2020.

[30] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet

models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

[31] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 270–279. Springer, 2018.

[32] Haotao Wang, Tianlong Chen, Shupeng Gui, TingKuei Hu, Ji Liu, and Zhangyang Wang. Once-for-all adversarial training: In-situ tradeoff between robustness and accuracy for free. *Advances in Neural Information Processing Systems*, 33:7449–7461, 2020.

[33] Mengting Xu, Tao Zhang, Zhongnian Li, Mingxia Liu, and Daoqiang Zhang. Towards evaluating the robustness of deep diagnostic models by adversarial attack. *Medical Image Analysis*, 69:101977, 2021.

[34] Qingzhao Zhang, Shengtuo Hu, Jiachen Sun, Qi Alfred Chen, and Z Morley Mao. On adversarial robustness of trajectory prediction for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15159–15168, 2022.