Transfer Robustness to Downstream Tasks Through Sampling Adversarial Perturbations

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

Due to the vulnerability of deep neural networks to adversarial attacks, adversarial robustness has grown to be a crucial problem in deep learning. Recent research has demonstrated that even small perturbations to the input data can have a large impact on the model’s output, exposing them susceptible to malicious attacks. In this work, we propose Delta Data Augmentation (DDA), a data augmentation method for enhancing transfer robustness by sampling extracted perturbations from trained models against adversarial attacks. The main idea of our work is to generate adversarial perturbations and to apply them to downstream datasets in a data augmentation fashion. Here we demonstrate, through extensive experimentation the advantages of our data augmentation method over the current State-of-the-Art in Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks for CIFAR10 dataset.

1. Introduction

The research in the field of adversarial robustness for deep learning aims to increase the robustness of models to adversarial attacks [3, 4, 11]. 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, 9]. Applications like autonomous vehicles [38], medical diagnosis [2, 36], and fraud detection [8] 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.

The recent literature has seen a surging interest in this field, resulting in a large number of techniques to improve the robustness of deep learning models to adversarial attacks, such as adversarial training [3, 15, 35], and data augmentation [12, 13]. For example, to enhance the model’s capacity to recognize and fend off hostile attacks, adversarial training entails supplementing the training data with adversarial samples. Conversely, data augmentation creates new data from existing data in order to expand the size of a dataset [28, 33]. This leads to an overall improvement in the robustness of the model by exposing it to a wider range of variations.

Our proposal, Delta Data Augmentation (DDA) (Fig. 1) solves the crucial problem of transfer robustness in deep learning. When there is a lack of labeled data, our transfer robustness approach only requires training a model on one dataset and then applying the obtained knowledge to another dataset. By incorporating perturbations sampled from trained models that are resistant to adversarial attacks, DDA is designed to improve transfer robustness. Given its design, our approach is able to 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, 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 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, in Section 6, we present our conclusions and discuss future work.

2. Related Work and Motivation

Due to deep neural network’s susceptibility to adversarial attacks, adversarial robustness has grown to be a crucial research area in deep learning (DL) \[1,6,10,15\]. It has been demonstrated that DL models are susceptible to adversarial instances, which are intentionally constructed inputs that can lead the model to produce wrong predictions \[21\].

Although several proposals for mitigating adversarial risks for DL models have been investigated \[18,20\], there is still a need for enhanced robustness in many settings. Recently, research on adversarial robustness has focused on creating adversarial attacks to fool a model \[15\]. These attacks can be classified as white-box or black-box attacks. In the former, the attacker has full knowledge of the model weights and architecture and can generate adversarial examples by optimizing a certain loss function. In the latter, the attacker has limited knowledge about the model and can only create adversarial examples by feeding the network with inputs and analyzing the outputs.

Similarly, adversarial defense methods have been proposed in recent years \[7,9\]. Defense methods can be mainly classified into two categories: pre-processing defense, which aims to modify input data before feeding the model. Popular techniques under this category include data augmentation \[12,13\] and input denoising \[25\]. The category, known as post-processing defenses, involves treating the output of a model. Examples of these techniques include adversarial training \[15,21\], defensive distillation \[26\], and model ensembles \[37\]. However, none of these can guarantee full coverage in terms of security and robustness against all adversarial inputs. Thus, robust adversarial defenses are still a high-demanding and high-priority requirement for designing trustable ML solutions.

Additionally, the accuracy under adversarial attacks is the most commonly used metric to evaluate the robustness of a model \[21\]. 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 \[11\]. Moreover, minimum distortion is a metric that assesses the minimum magnitude of adversarial perturbation needed to fool a model \[5\].

The security and dependability of deep learning systems are seriously threatened by adversarial robustness in critical applications. Many defense mechanisms have improved the robustness of models, but considerable work still needs to be done before these models can tolerate various adversarial attacks and still keep their natural accuracy. Therefore, reducing this gap between accuracy and robustness is still a research problem \[17,29,34\].

The main contribution of this work is the data augmentation method based on adversarial attacks and transfer learning to enhance model robustness. DDA is designed by using adversarial perturbations that are effective on a larger and more complex task to transfer them to downstream tasks.

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 \[6\]. 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 $f : X \rightarrow Y$ that maps input data to output labels. Adversarial examples can be generated by adding a small perturbation $\delta$ to the input data $x$, such that $x' = x + \delta$. The perturbation is typically constrained to have a small $\ell_p$-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 Training

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).

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

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

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

$$\delta = \arg \max_{\delta} L(f_{\theta}(x + \delta'), y), \text{s.t.} |\delta|_p \leq \epsilon$$  \hspace{1cm} (2)
One approach to generating adversarial examples is to use an iterative optimization algorithm such as Fast Gradient Descent Method (FGSM) [15] or Projected Gradient Descent (PGD) [21] to compute a perturbation. The resulting adversarial example \( x + \delta \) is then added to the original training data along with 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.

\[
\min_{\theta} \frac{1}{n + m} \sum_{i=1}^{n} L(f_{\theta}(x_i), y_i) + \frac{1}{n + m} \sum_{i=1}^{m} L(f_{\theta}(x_i + \delta_i), y_i),
\]

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

### 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 [34]. In TL, a pre-trained model is used as a starting point for a new model rather than beginning from scratch [17]. For this, \( g_{\phi} : X^\prime \rightarrow Y^\prime \) is a pre-trained model parameterized by \( \phi \), where \( X^\prime \) and \( Y^\prime \) may or may not be the same as \( X \) and \( Y \). The goal of transfer learning is to initialize the parameters of \( f_0 \) using the pre-trained parameters \( \phi \) and then fine-tune \( f_0 \) using a small amount of data from the new task [32]. Then, transfer robustness of \( g_{\phi} \) is defined as the ability of \( f_0 \) to maintain its performance on a new task under adversarial attacks when initialized with the pre-trained parameters \( \phi \) [24,32]. Now, instead of using pre-trained parameters, we look for transfer adversarial perturbations that are effective on a larger and more complex model.

In accordance with 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 [10, 22]. 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. We aim to collect adversarial noise \( \delta \) 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 model that is trained on an upstream task, such as ImageNet Classification, is used to sample adversarial perturbations \( \delta \) given an adversarial attack (Fig. 3). The objective of this process is to obtain a representative sample of perturbations that re-

Figure 2. Accuracy comparison for PGD with \( L_2 \) (2a) and \( L_\infty \) (2c), and FGSM with \( L_2 \) (2b) and \( L_\infty \) (2d), for CIFAR-10 dataset with different methods of data augmentation: Delta Data Augmentation (ours), RandAugment [13], AutoAugment [12], Standard Training with no data augmentation. An epsilon of 0 means natural accuracy.

Figure 3. Example of DDA for CIFAR10. First, is the original image, then, in the second image is the perturbation extracted by DDA. Finally, the third image is the augmented image.
Table 1. Results of Delta Data Augmentation (DDA) on CIFAR10 with ResNet18 on each adversarial attacks: Fast Gradient Sign Method (FGSM) [15], Projected Gradient Descent (PGD) [21], Basic Iterative Attack (BAI), Additive Uniform Noise Attack (AUNA) [30], and Deep Fool Attack (DFA) [23] for different epsilon \( \epsilon \) perturbation intensities, \( \epsilon = 0 \) means natural accuracy (no attack). In bold are the highest robust scores for each \( \epsilon \), and \( \ell_\infty \), \( \ell_2 \) norms respectively.

<table>
<thead>
<tr>
<th>Norm</th>
<th>Attack</th>
<th>0</th>
<th>0.0005</th>
<th>0.001</th>
<th>0.0015</th>
<th>0.002</th>
<th>0.003</th>
<th>0.005</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FGSM</td>
<td>0.843</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>PGD</td>
<td>0.83</td>
<td>0.82</td>
<td>0.807</td>
<td>0.794</td>
<td>0.767</td>
<td>0.725</td>
<td>0.624</td>
<td>0.456</td>
<td>0.331</td>
<td>0.024</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>BIA</td>
<td>0.801</td>
<td>0.762</td>
<td>0.725</td>
<td>0.69</td>
<td>0.615</td>
<td>0.488</td>
<td>0.233</td>
<td>0.037</td>
<td>0.015</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>AUNA</td>
<td>0.843</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>DFA</td>
<td>0.813</td>
<td>0.782</td>
<td>0.761</td>
<td>0.735</td>
<td>0.685</td>
<td>0.579</td>
<td>0.331</td>
<td>0.186</td>
<td>0.139</td>
<td>0.093</td>
<td>0.068</td>
<td>0.054</td>
<td>0.06</td>
<td></td>
</tr>
</tbody>
</table>

Average 0.843 0.8296 0.8171 0.8067 0.796 0.7754 0.7372 0.6627 0.5961 0.5651 0.4809 0.3678 0.2921 0.2169 0.001

5. Discussion

The comparison of accuracy across various data augmentation techniques reveals that DDA performs better than the others in terms of robust accuracy against PGD and FGSM attacks. Particularly, DDA outperforms other approaches in terms of robust accuracy, achieving values of 76.7% and 84% for PGD attack with \( \epsilon = .003 \) for \( \ell_\infty \) and \( \ell_2 \) respectively. Also, for FGSM 65.5% and 83.5% for \( \ell_\infty \) and \( \ell_2 \) respectively with same \( \epsilon \).

The comparison of PGD and FGSM attacks with various \( \epsilon \) values further demonstrates that the resilience of the model is significantly impacted by the choice of attack strength (Table 1). As expected given that greater attacks bring larger perturbations that are more challenging to recover from, our results indicate that stronger attacks result in lower robust accuracies (Fig. 2).

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.

6. Conclusions and Future Work

In this study, we compared a variety of data augmentation strategies, such as DDA, RandAugment, AutoAugment, and Standard Training with no data augmentation, to examine the performance of various adversarial attack methods on the CIFAR10 dataset. Our findings demonstrated that, in terms of adversarial robustness, our approach performed better than or equal to State-of-the-Art approaches. DDA improves the transferability of robustness against adversarial attacks by reducing the gap between natural and robust accuracy. Future studies can examine the use of DDA with additional datasets and see how well it defends against increasingly sophisticated adversarial attacks.
References


