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

Ivan Reyes-Amezcua, Gilberto, Ochoa-Ruiz, Andres Mendez-Vazquez

Introduction

Adversarial robustness is a machine learning model's **ability to resist** subtle, intentionally manipulated inputs designed **to cause incorrect predictions.**

A **data augmentation method** that enhances transfer robustness by **sampling perturbations** extracted from models that have been robustly trained against **adversarial attacks.**

Motivation

Traditional **adversarial training** can falter with unfamiliar datasets due to the absence of tailored robust methods

Balancing **high accuracy** on clean data with robustness against adversarial attacks becomes challenging **with unknown dataset traits** and **constrained resources.**

Method

- **Adversarial perturbations** are collected by attacking **upstream pre-trained** model tasks (e.g. *ImageNet*)
- The resulting **adversarial perturbations** (Fig. 1) are added to the original training data along with its corresponding label y, creating an **augmented training dataset**

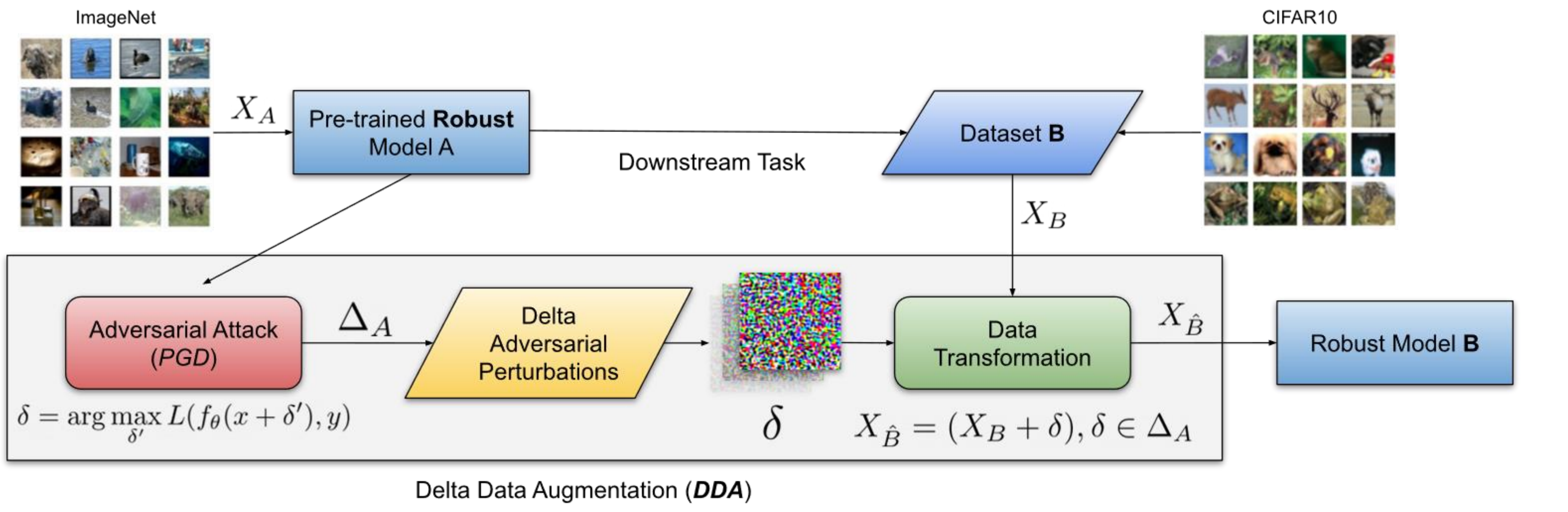

Figure 1: Example of PGD attack with l_∞ on a ResNet18 pre-trained with ImageNet. Each column is the resulting adversarial image and perturbation with different ε attack intensities (e.g. $0/255, 2/255, \ldots, 16/255$). The left-most image is the original example with $\varepsilon = 0$, meaning no adversarial attack.

Results

Through a comparison with other data augmentation strategies (Fig. 2) on datasets like **CIFAR10**, **CIFAR100**, and **SVHN**, we found that **DDA** either outperformed or matched the performance of leading techniques, bridging the **gap between natural and robust accuracy.**

This indicates the potential of incorporating **adversarial perturbations,** like DDA, into **training to enhance adversarial robustness significantly.**

Conclusion

We introduce ***Delta Data Augmentation (DDA)***, a unique data augmentation technique that utilizes **adversarial attacks** and **transfer learning** to boost model **robustness** in downstream tasks.

This is the **pioneering approach** that incorporates adversarial examples into training **without engaging in adversarial training for defense.**

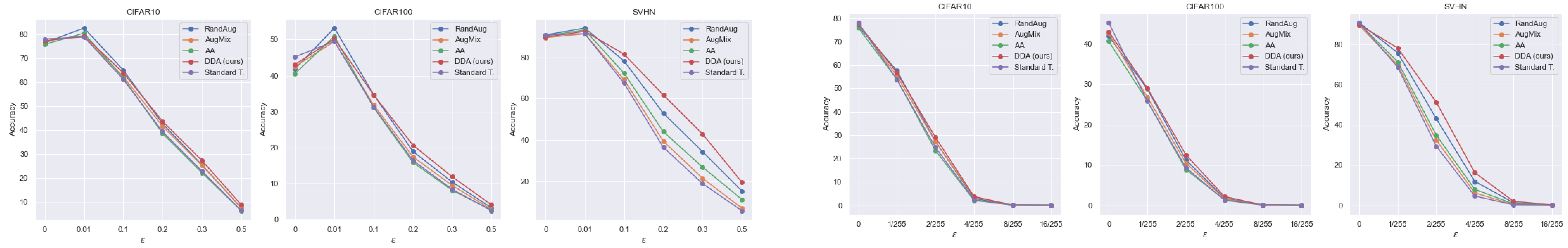

Figure 2: Accuracy results for PGD with l_∞ (left) and l_2 (right) on CIFAR10, CIFAR100 and SVHN datasets, compared with RandAugment, AutoAugment, AugMix, No Data Augmentation, and DDA (ours), trained with ResNet18.