Evaluate and Extend: Taxonomy Driven Fast Adversarial Training

Sai Manikandan
School of Computing
gmanikandan@binghamton.edu

Shyam Kannan
School of Computing
skannan@binghamton.edu

***Abstract*—Adversarial Training (AT) is one of the most effective defenses against gradient-based adversarial attacks. However, when applied using single-step methods like FGSM, AT often suffers from catastrophic overfitting—where models become highly vulnerable to multi-step attacks such as PGD. Taxonomy-Driven Adversarial Training (TDAT) [ref] addresses this issue by categorizing examples based on their robustness failure modes and then adapting the training strategy according to the issue. In this work, we evaluate the generalizability of TDAT across multiple architectures and domains, including ResNet-18 and DeiT-small on CIFAR-100, and DistilBERT on the SST-2 NLP benchmark. Our results show that while TDAT improves robustness in vision models, direct application to NLP models is less effective without architectural and training adaptations. This study highlights both the potential and the limitations of taxonomy-guided adversarial training across modalities.**

I. Introduction

Deep neural networks have shown remarkable performance across computer vision and natural language processing tasks. However, their susceptibility to adversarial examples, i.e. inputs perturbed with noise poses a critical challenge, especially in security driven tasks. To mitigate this vulnerability, adversarial training (AT) has emerged as one of the most effective defenses. By augmenting training data with adversarial examples, AT forces the model to learn robust decision boundaries.

Inspite of its success, standard adversarial training methods suffer from a phenomenon known as *catastrophic overfitting*. This occurs in single-step AT strategies like FGSM, where models quickly overfit to weak perturbations and remain highly vulnerable to stronger multi-step attacks like PGD. This exposes a fundamental weakness in the way robustness is enforced and has motivated researchers to explore more nuanced training strategies.

Taxonomy-Driven Adversarial Training (TDAT) [1] introduces a novel framework to mitigate catastrophic overfitting in single-step adversarial training. The authors first identify and categorize adversarial examples into five failure cases, offering a new perspective on the root causes of training collapse. Based on this insight, TDAT incorporates three key modifications: batch momentum initialization, dynamic label relaxation, and a taxonomy-driven loss function. Together, these modifications help improve robustness while preserving training efficiency. The approach demonstrates significant improvements across various CNN architectures, including ResNet-18.

However, TDAT's evaluation is restricted to convolutional networks. It remains unclear whether the taxonomy-guided training approach generalizes to newer architectures such as vision transformers or to other domains like text classification. The applicability of TDAT beyond CNNs, as well as the relative importance of its individual components, has not been studied.

In this work, we aim to bridge this gap by first reproducing TDAT results on ResNet-18 and then extending the approach to DeiT-small (Data-Efficient Image Transformers) on CIFAR-100, and DistilBERT on the SST-2 sentiment classification dataset. We re-implement TDAT's taxonomy-based perturbation logic, label relaxation, and robust loss mechanism across these architectures. We then analyze the results from the experiments to analyze their effectiveness and limitations revealing that TDAT improves robustness in vision models, whereas its adaptation in the NLP domain requires better adaptation to function effectively.

Code Repositories. Our implementation builds upon the official TDAT codebase[1] and BERT adversarial training frameworks[2]. The source code for our experiments is available at https://github.com/musical-shyam/Evaluate-and-Extend-TDAT.git.

II. Background

A. Taxonomy of Adversarial Examples

Figure 1 below (reproduced from the TDAT paper) visualizes the relationships between these cases, showing how adversarial perturbations impact the decision boundary and class predictions.

TDAT introduces a novel taxonomy to better understand the dynamics of adversarial examples (AEs) in single-step adversarial training (AT), particularly in the context of catastrophic overfitting (CO). Each training example (x, y) is categorized into one of five mutually exclusive cases based on its clean and adversarial predictions:

- **Case 1 (Clean Correct, Adv Incorrect)**: The clean input is correctly classified, but its adversarial version

[1]https://github.com/bookman233/TDAT.git
[2]https://github.com/VijayKalmath/AdversarialTraining_in_Distilation_Of_BERT.git

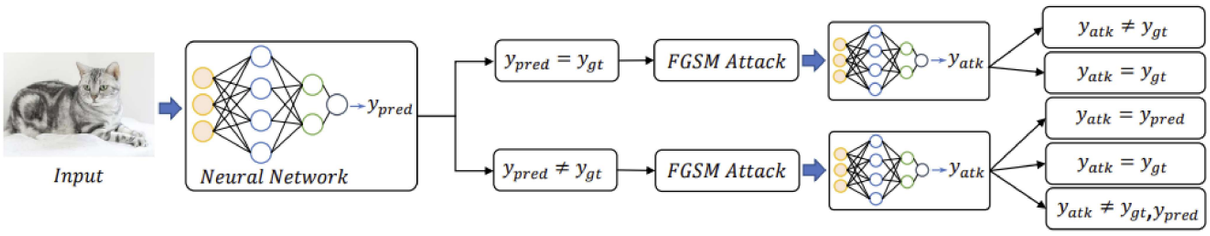

Fig. 1. TDAT's taxonomy of adversarial examples. Case 4 and Case 5 are considered the most problematic as they drive incorrect training signals.

is misclassified. This is the classic goal of adversarial training—creating hard AEs from correct inputs.

- **Case 2 (Clean Correct, Adv Correct)**: Both clean and adversarial inputs are correctly classified. This is the ideal outcome, indicating true robustness.
- **Case 3 (Clean Incorrect, Adv Incorrect - Same Class)**: The clean input is misclassified as class j and the adversarial version is also classified as j. The adversarial example preserves the (wrong) clean prediction.
- **Case 4 (Clean Incorrect, Adv Correct)**: The clean input is misclassified, but the adversarial version flips to the correct label. This might look like a success, but helps in the model to overfit the wrong clean prediction to the correct adversarial prediction. This case is also called **label flipping**.
- **Case 5 (Clean Incorrect, Adv Incorrect - Different Class)**: The clean input is misclassified as class j, and the adversarial version is misclassified as yet another class k (where $k \neq i, j$).

Figure 1 (from the TDAT paper) visualizes the relationships between these cases, showing how adversarial perturbations impact the decision boundary and class predictions.

B. Taxonomy-Driven Insights into Catastrophic Overfitting

A key insight from TDAT is that catastrophic overfitting (CO) in single-step adversarial training arises from a growing proportion of Case 4 examples during training. In Case 4, clean examples that were originally misclassified are perturbed such that their adversarial counterparts are incorrectly predicted as the ground-truth label. This creates a misleading gradient signal that reinforces incorrect behavior — a phenomenon referred to as label flipping.

TDAT illustrates this behavior in **Figure 2**, showing how adversarial examples in Case 4 lead the model to learn incorrect mappings. Furthermore, the paper tracks the distribution of taxonomy cases across epochs (Figure 3) and observes a sharp spike in Case 4 just before the collapse of robust accuracy, highlighting its central role in CO. These observations motivate the taxonomy-driven loss and training strategy proposed in TDAT.

III. Methodology

To counter the influence of Case 4 examples, TDAT proposes a three changes to the adversarial training solution:

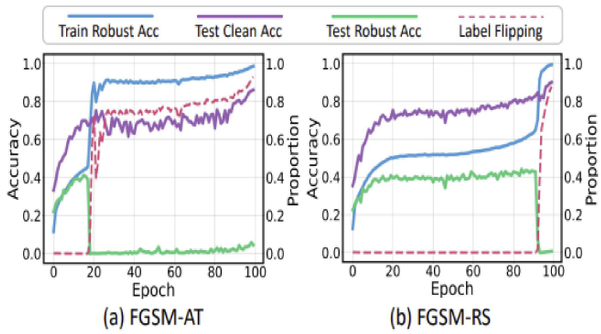

Fig. 2. CO and Label Flipping. As CO occurs, the proportion of case 4 (red line) occurs. Against PGD-10 attack.

1. Batch Momentum Initialization

To increase the strength of adversarial perturbations while preventing CO, TDAT introduces a momentum-based initialization strategy. Instead of initializing perturbations randomly each time, the method builds on perturbations from the previous batch to guide the next:

$$\delta_m = \alpha \cdot \delta_{m-1} + (1 - \alpha) \cdot \delta \tag{1}$$

Here, δ is the current batch's perturbation, and δ_{m-1} is the momentum from the previous batch. This approach adds a strength to perturbation, rather than it being random and thus the model easily overfitting on the perturbations .

2. Dynamic Label Relaxation

TDAT addresses the issue of overconfident weights mainly for adversarial examples from misclassified clean inputs (Case 3-5). It is done by relaxing the ground-truth label dynamically over training:

$$\hat{y} = y \cdot \gamma + (1 - y) \cdot \frac{\gamma - 1}{L - 1} \tag{2}$$

Here, y is the original one-hot label, L is the number of classes, and γ is the label relaxation factor.

This allows confident labels early in training, gradually transitioning to softer targets.

3. Taxonomy-Driven Loss

In order to ensure stable training and discourage harmful learning signals from misclassified examples, TDAT modifies the standard loss function by introducing a regularization term that pushes for similarity between the clean and adversarial outputs:

$$\mathcal{L}_{TD} = \mathcal{L}_{CE} + \lambda \cdot \|f(x+\delta) - f(x)\|_2^2 \cdot \tanh(1-p) \quad (3)$$

Here, $f(x)$ and $f(x+\delta)$ denote the model's logits for clean and adversarial inputs respectively, and p is the softmax confidence score of the correct class.

IV. Our Contribution

This project explores the generalizability of Taxonomy-Driven Adversarial Training (TDAT) beyond its original scope in CNN architecturs. We pursue three main goals:

1) Reproduce the original TDAT results using ResNet-18 on the CIFAR-100 dataset.
2) Extend the TDAT framework to transformer-based vision models by applying it to DeiT-Small.
3) Investigate the applicability of TDAT to NLP by adapting it to a sentiment classification task using DistilBERT on the SST-2 dataset.

Our hypothesis is that TDAT's core principles, i.e. batch momentum initialization, dynamic label relaxation, and taxonomy-driven loss should provide improved adversarial robustness in image classification. Also, a part of our pursuit is to increase robust accuracy in language models with TDAT adapted, taking into consideration the challenge of NLP models due to its discrete input nature.

V. Experimental Setup

We evaluate the effectiveness of the TDAT framework across both vision and language domains. Specifically, we conduct experiments on the following architectures:

- **ResNet-18** and **DeiT-Small** for image classification on CIFAR-100.
- **DistilBERT** for sentiment analysis on SST-2.

Vision Models. All models are evaluated under an ℓ_∞ threat model with perturbation budget $\epsilon = 8/255$, using the following attacks:

- **FGSM**
- **PGD-10**
- **AutoAttack**
- **ResNet-18** is trained from scratch using only the TDAT training procedure. This setup is used to replicate and verify the results from the original TDAT paper on CIFAR-100.
- **DeiT-Small** is fine-tuned from a publicly available pre-trained checkpoint. We experiment with both standard FGSM-based adversarial training and the full TDAT pipeline to compare robustness outcomes.

NLP Model. DistilBERT is fine-tuned on SST-2 using adversarial training via data augmentation. The attack is done using **Textattack**, to follow TDAT's philosophy of attacking AT trained models using better attacks. Also, the training is done using the following augmenters:

- **WordNetAugmenter**
- **SynonymInsertionAugmenter**
- **EmbeddingAugmenter**

Hardware and Parallelism. Training was conducted using a mix of local and cluster resources:

- **4× NVIDIA A40 GPUs (spiedie) —** using Distributed Data Parallel (DDP)
- **1× RTX 4070 (local)**

VI. Results

A. Reproducing TDAT on ResNet-18

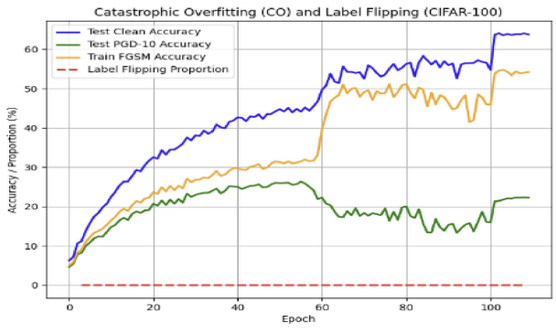

Fig. 3. Our TDAT results on ResNet18.

To verify the effectiveness of TDAT, we reproduced its results on ResNet-18 using the CIFAR-100 dataset. No baseline FGSM adversarial training (AT) was applied in this experiment; we focus solely on the TDAT framework as introduced in the original paper.

Training Stability. Figure 2 (from the TDAT paper) shows the robust accuracy of PDG-10 attacking a FGSM based AT model. A clear spike in label flipping around epoch 80 is observed, along with a collapse in test robust accuracy, thus catastrophic overfitting. In contrast, our TDAT training dynamics (Figure 3) show stable behavior throughout all epochs. The label flipping proportion remains consistently zero, and robust accuracy is preserved without collapse. This validates TDAT's core principal of protecting the robust accuracy of two step attacks against single-step AT models.

TABLE I
Robust accuracy (%) of various methods on CIFAR-100 under different attacks for ResNet18

Method	Clean	PGD-10	FGSM	C&W	APGD	Square
FGSM-RS	51.67	22.61	31.02	20.92	22.61	18.72
Original TDAT	57.32	**33.56**	40.29	**28.47**	**33.15**	**31.06**
Ours (TDAT)	**63.73**	22.26	**54.16**	18.37	25.34	21.03

Quantitative Results. Table II compares the performance of FGSM-RS, original TDAT (reported in the paper), and our

reproduced TDAT. Our implementation achieves the highest clean accuracy (63.73%) and FGSM robustness (54.16%), suggesting that our model is more specialized toward the FGSM training setup. However, robustness under stronger attacks like PGD, C&W, APGD, and Square is slightly lower, indicating reduced generalization.

B. DeiT-Small: Vision Transformer Results on CIFAR-100

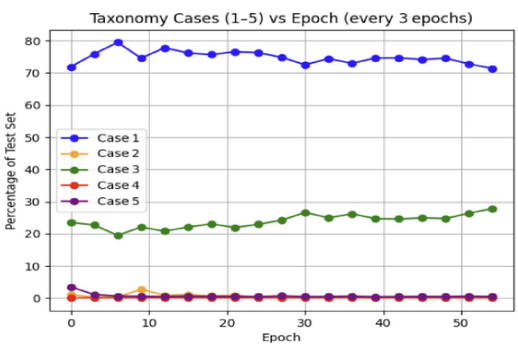

Fig. 4. Taxonomy case distribution for DeiT-Small under FGSM-AT.

Failure of FGSM-AT. Adversarial training using FGSM fails to produce robustness on DeiT-Small, despite avoiding label flipping. As shown in Figure 4, nearly all adversarial examples belong to **Case 1** (clean correct, adv incorrect), while **Case 2** (clean correct, adv correct) remains close to zero throughout. This imbalance in training severely limits generalization. Figure 5 confirms this failure as although clean and FGSM accuracy are moderately high, **PGD-10 accuracy collapses to 0%**, indicating catastrophic overfitting.

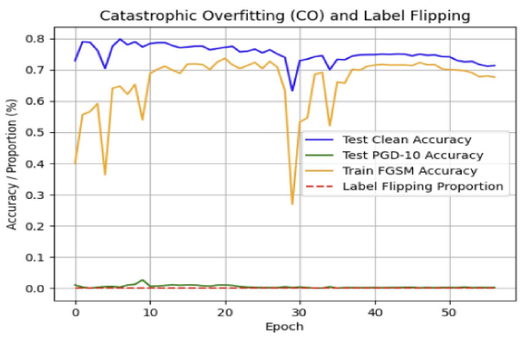

Fig. 5. Training curves for DeiT-Small under FGSM-AT.

TDAT Improves Taxonomy Distribution. With TDAT, we observe a meaningful shift in taxonomy distribution. As shown in Figure 6, the proportion of **Case 2** examples increases during training. This is a sign that the model learns to correctly classify adversarial examples. The dominance of **Case 1** reduces, improving the model's generalization to stronger attacks. This is further confirmed by Figure 7, where PGD-10 accuracy improves substantially.

TABLE II
ROBUST ACCURACY (%) OF AT AND TDAT ON DEIT

Method	Clean	PGD-10	FGSM	C&W
DeiT-AT	**71.33**	0.24	**67.62**	0.16
DeiT-TDAT	36.74	**21.28**	25.07	**18.12**

Quantitative Comparison. Table II summarizes the performance of FGSM-AT and TDAT. While TDAT slightly reduces clean and FGSM accuracy compared to FGSM-AT, it delivers a significant improvement under stronger attacks like PGD-10 and C&W, demonstrating its superior generalization.

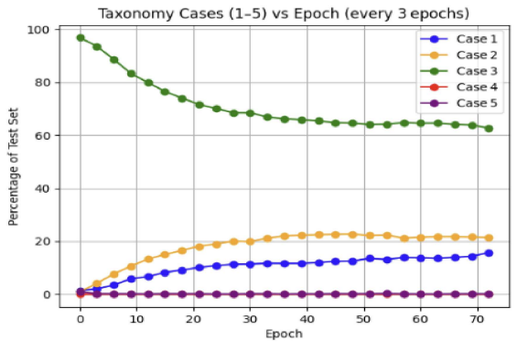

Fig. 6. Taxonomy case distribution for TDAT on DeiT-Small.

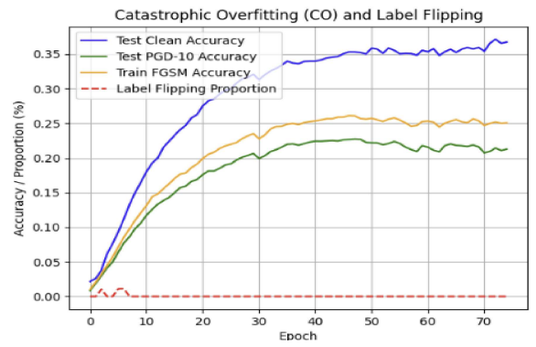

Fig. 7. Training curves for TDAT on DeiT-Small.

C. DistilBERT on SST-2

We attempted to extend the TDAT framework to the language domain using DistilBERT on the SST-2 sentiment classification task. Due to the lack of well-defined input perturbation methods in text, we only implemented the label relaxation component of TDAT, applying it alongside adversarial data augmentation.

Standard AT Behavior. Adversarial training using weak textual augmentations (WordNet, SynonymInsertion, and Embedding-based substitutions) preserved reasonable clean

validation accuracy (around 85%). However, the model remained fragile under stronger perturbations such as those generated by TextFooler, showing negligible adversarial accuracy.

TDAT Label Relaxation Failure. When applying label relaxation, validation accuracy collapsed by the 7^{th} epoch. Despite retaining the same training setup, the model failed to converge, producing nearly random outputs on both clean and perturbed test data.

These results suggest that label relaxation, when directly applied to binary classification tasks like SST-2, may be detrimental to learning.

VII. Discussion

Vision Domain Generalization. Our experiments on CIFAR-100 demonstrate that TDAT works effectively across different vision architectures. While originally proposed for ResNet-18, we found that the method generalizes well to transformer-based models like DeiT-Small. The training remained stable, label flipping was avoided, and robustness under multi-step attacks significantly improved. With further hyperparameter tuning, we believe that the clean accuracy loss observed in ViTs can also be minimized, making TDAT a strong candidate for single-step adversarial robustness in vision tasks.

Limitations in NLP Settings. Our attempt to apply TDAT's dynamic label relaxation in the NLP setting revealed a critical limitation. In tasks like SST-2, where labels are about binary sentiment polarity, softening the ground-truth label appears harmful. The model failed to learn discriminative features, and accuracy collapsed rapidly. We hypothesize that this is due to the nature of NLP tasks, where decision boundaries are sharper and features are less common between classes than vision.

VIII. Conclusion

In this work, we reproduced the Taxonomy-Driven Adversarial Training (TDAT) framework and tried to extend its application across diverse domains. Our experiments validated that TDAT effectively mitigates catastrophic overfitting in vision tasks, including both convolutional and transformer-based architectures. Through improved taxonomy dynamics and stable training behavior, TDAT enhances robustness under stronger adversarial attacks, especially on CIFAR-100.

However, our results also reveal that direct application of TDAT's label relaxation strategy to NLP tasks like SST-2 may be counterproductive. This highlights an important limitation and opens the door for future adaptations of taxonomy aware AT, but in discrete text input space, in order to mitigate the catastrophic overfitting that even language models face.

Overall, TDAT provides a promising direction for robust single-step adversarial training, especially in vision. Domain specific adaptations could generalize the philosophy of taxonomy aware AT even better.

References

[1] K. Tong, Z. Yang, X. Chen, Y. Liu, and B. Li, "Taxonomy driven fast adversarial training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5233–5241.