UNIVERSITY of WASHINGTON

Corruption and Adversarial Robustness of Point Cloud Classifiers

Masa Nakura-Fan, Chung Yik Edward Yeung

Motivation

- Dual robustness in point cloud classifiers is **underexplored**, yet **crucial for safety-sensitive uses**.
- Aims to balance adversarial and corruption robustness, moving away from single-metric improvement focus.
- Evaluates various models on ModelNet-C and **AutoAttack** for a comprehensive, unbiased robustness evaluation against a variety of attacks
- AutoAttack's [2] APGD offers a **more thorough adversarial example discovery than PGD**, representing a novel direction in adversarial training.
- Propose to use an **APGD for adversarially training** point cloud classifiers, which achieves an average error rate (ER) lower than any other model we tested.
- Encourage further research into **improving both adversarial and corruption accuracy** in point cloud classifiers.

Dataset

- ModelNet40, a collection of 3D CAD models across 40 categories - used to generate adversarial examples.
- ModelNet40-C adds 15 corruption types at 5 severity levels, addressing density, noise, and transformation issues.
- ModelNet40-C includes 185,000 unique point clouds for comprehensive evaluation.

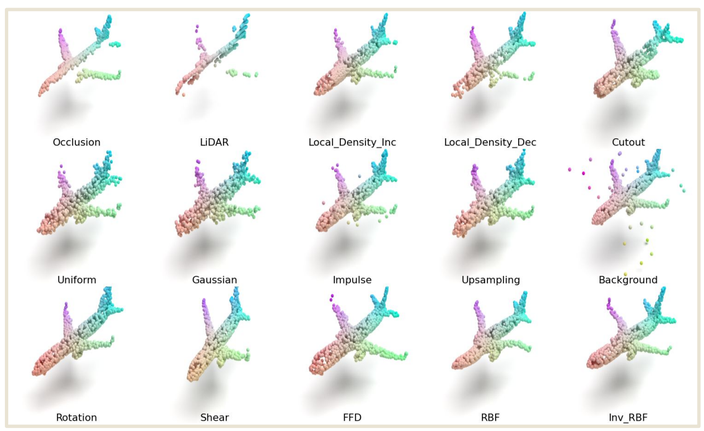

Figure 1: 3D point cloud model of an airplane subjected to various types of corruptions pulled from the ModelNet40-C dataset [3]

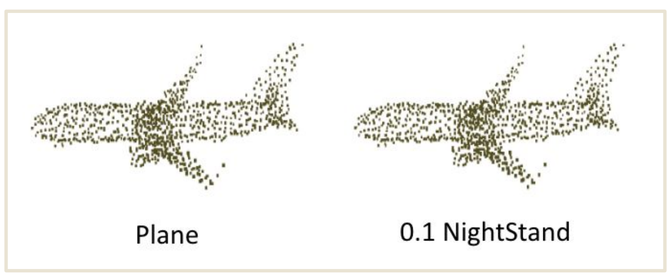

Figure 2: Utilizing ModelNet40 to assess standard robustness and also to create adversarially perturbed data. For example, the image of a plane from ModelNet40 (left) is compared with its adversarially altered version generated by AutoAttack (right).

References

[1] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).
[2] Croce, Francesco, and Matthias Hein. "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks." *International conference on machine learning*. PMLR, 2020.
[3] Jiachen Shen. ModelNet40-C: A Cleaned and Complete Version of ModelNet, 2023.

Threat Model

- Evasion Attacks: Misclassify semantically similar point clouds.
- Whitebox attack using the AutoAttack framework
- Allowed perturbation to a point is an L-infinity-ball of distance ε

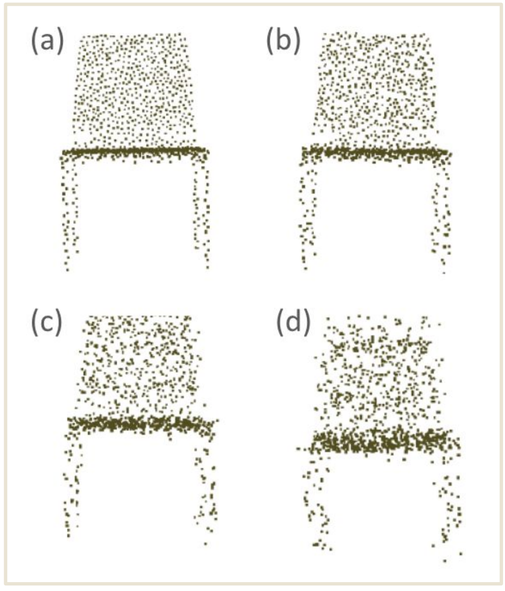

Figure 3: (a) A point cloud of a chair from ModelNet 40. This chair was perturbed using AutoAttack with (b) ε = 0.1, (c) ε = 0.25, (d) ε = 0.5. Under a standard DGCNN model for point clouds, they were classified as (a) chair, (b) night stand, (c) night stand, and (d) plant

Methods

Benchmarking adversarial robustness on corruption robust models:

- Implement an adversarial robustness evaluation framework for point clouds classifiers by leveraging AutoAttack

Adversarial Training using APGD:

- Objective Function as defined by Madry et. al[1]:

$$\min_{\theta} \rho(\theta) \text{ where } \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y) \right]$$

- Implemented APGD data augmentation in DGCNN and PointNet models for maximizing the inner objective function
- Idea: APGD performs better than PGD, so better inner maximization.

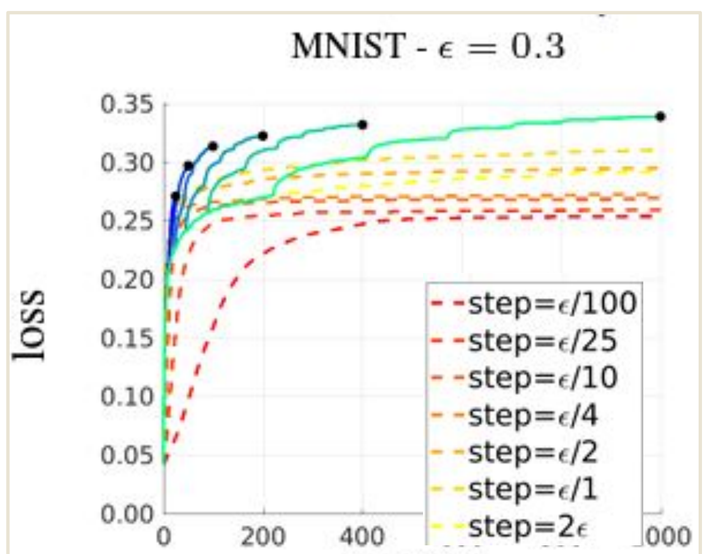

Figure 4: Loss of Madry et. al.[1] model on MNIST dataset for PGD with momentum (dashed lines) with different step sizes and APGD (solid iterations) with different number of iterations. APGD almost always reaches a higher value

Results

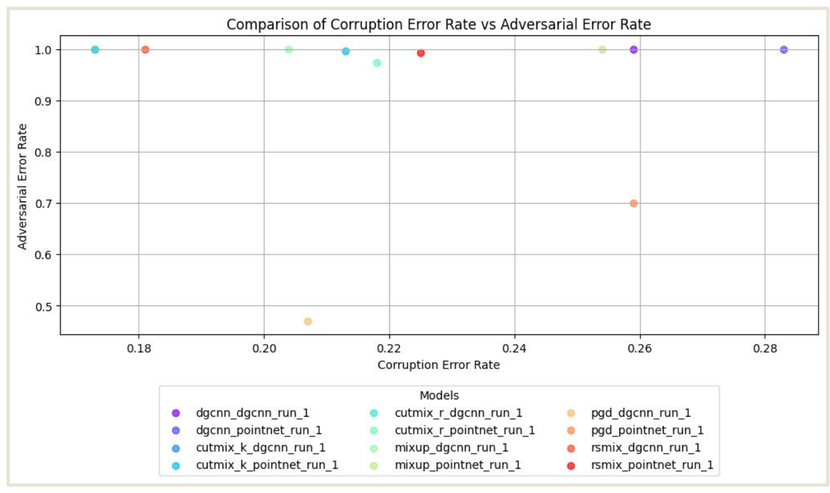

Figure 3: Comparison between Adversarial Error and Corruption Error Rate for DGCNN and Pointnet models using various data augmentation such as PGD, Cutmix-R etc.

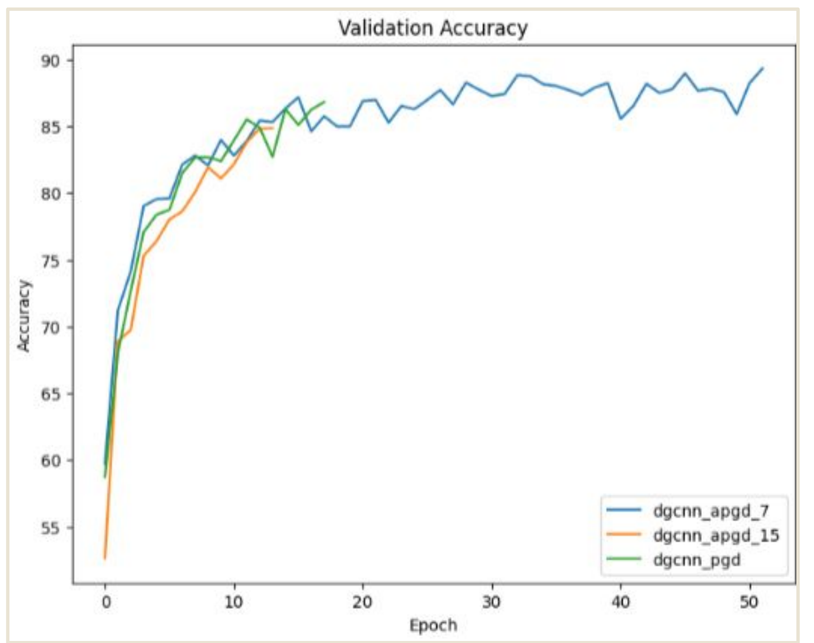

Figure 5: Validation Accuracy over training epochs for DGCNN with APGD using 7 iterations, APGD using 15 iterations, and PGD. Typically trained for 300 epochs but cut short due to lack of time and resource.

Model	ER_{CL}	ER_{CR}	Density	Noise	Transformation	ER_{Adv}	Avg. ER
PointCutMix-K-PointNet	9	21.3	**26.8**	21.8	15.4	99.71	43.34
PGD-PointNet	11.8	25.9	28.8	28.4	20.5	69.92	35.87
PointCutMix-K-DGCNN	**6.8**	**17.3**	28	**11.4**	11.5	100	41.37
PGD-DGCNN	8.1	20.7	36.8	13.8	11.5	46.97	25.26
APGD-DGCNN (ours)	11.92	20.03	33.68	16	**10.42**	**42.28**	**24.74**

Table 1: Comparison of different models' error rates for clean, corrupted, and adversarial data. Corrupted error rates is further broken down to density, noise, and transformation type corruption. Finally, we report the average of all error rate. The APGD was trained with iterations of 7 for ~50 epochs.

Summary

- High corruption robustness does not indicate high adversarial robustness
- When training models, we encourage a focus in all robustness metrics for clean, corrupt, and adversarially perturbed data (i.e. an average. or weighted avg. of all ERs).
- APGD as an data augmentation/data augmentation technique has potential
- Future work:
  - Train APGD models for longer epochs
  - Train models with higher APGD iterations which we predict increases adversarial robustness.
  - Is adversarial robustness a good indicator for corruption robustness?