Adversarial Training with Synthesized Data: A Path to Robust and Generalizable Neural Networks

Reza Bayat [1 2 3] **Irina Rish** [1 2 3]

Abstract

Adversarial Training (AT) is a well-known framework designed to mitigate adversarial vulnerabilities in neural networks. Recent research indicates that incorporating adversarial examples (AEs) in training can enhance models' generalization capabilities. To understand the impact of AEs on learning dynamics, we study AT through the lens of sample difficulty methodologies. Our findings show that AT leads to more stable learning dynamics compared to Natural Training (NT), resulting in gradual performance improvements and less overconfident predictions. This suggests that AT steers training away from learning easy, perturbable spurious features toward more resilient and generalizable ones. However, a trade-off exists between adversarial robustness and generalization gains, due to robust overfitting, limiting practical deployment. To address this, we propose using synthesized data to bridge this gap. Our results demonstrate that AT benefits significantly from synthesized data, whereas NT does not, enhancing generalization without compromising robustness and offering new avenues for developing robust and generalizable models.

1. Introduction

Adversarial examples (AEs) have emerged as a critical concern for neural networks, posing significant challenges to deployed systems. These examples are indistinguishable to the human eye but significantly deceive neural networks, leading them to misclassify and make erroneous predictions. The consequences of AEs are substantial, as they can undermine the integrity of various systems, including image recognition (Xie et al., 2019), autonomous vehicles (Xiong et al., 2021), medical applications (Ma et al., 2021; Hirano et al., 2021; Bortsova et al., 2021; Apostolidis & Papakostas, 2021), large language models (Greshake et al., 2023), and multimodal systems (Cui et al., 2023b; Zhao et al., 2024; Dong et al., 2023). Early explorations into this phenomenon by Szegedy et al. (2013) shed light on the vulnerability of neural networks to adversarial perturbations. Furthermore, adversarial training (AT), proposed by Goodfellow et al. (2014), has emerged as the de facto method for enhancing the robustness of networks against these vulnerabilities.

On the other hand, recent studies have demonstrated the intriguing potential of leveraging AEs to augment training data for purposes other than robustness, such as improving image recognition performance (Xie et al., 2020; Rebuffi et al., 2023) and generalization (Alhamoud et al., 2022; Rebuffi et al., 2022; Deng et al., 2021). This repurposed usage of AEs has also been beneficial for debiasing the visual system (Zhang & Sang, 2020), one of the primary and ongoing challenges of computer vision. This highlights the complex nature of AEs: even though they pose threats to neural networks, they also offer opportunities for advancing learning dynamics for robustness and other purposes.

Despite these advancements, a significant trade-off persists between achieving adversarial robustness and improving other essential capabilities. For instance, Kireev et al. (2021) demonstrated the effectiveness of AEs in enhancing model performance on commonly corrupted data, illustrating their potential benefits for better generalization. However, their findings also indicate that employing AEs necessitates using a much smaller ϵ attack during adversarial training ($1/255$ compared to $8/255$), which substantially limits the robustness of the models and hinders their practical application.

Studying the impact of AEs on the learning dynamics of networks is crucial for understanding what makes them effective in developing models that balance robustness and generalization effectively. Therefore, we begin to study AT from different perspectives of example difficulty, such as entropy (Shannon, 2001) and a pointwise framework of learning by Kaplun et al. (2022), and reveal a much more stable learning behavior with less overconfident prediction compared to Natural Training (NT). Then, we leverage synthesized data to overcome the shortcomings of AT and reduce the gap in the trade-off between robustness and generalization.

[1]Mila – Quebec AI Institute [2]Université de Montréal [3]CERC-AAI. Correspondence to: Reza Bayat <reza.bayat@mila.quebec>.

2. Example Difficulty Methodologies

Understanding and quantifying the difficulty of samples within a dataset is crucial for improving model performance and interpretability. Various methodologies have been developed to measure this difficulty. However, in this section, we review two approaches, *entropy* and the *pointwise framework of learning*, which provide different perspectives on model uncertainty and performance.

Entropy Entropy is a fundamental concept in information theory, often used to quantify the uncertainty or unpredictability in a probability distribution (Shannon, 2001). Entropy can be applied to the output probabilities of a model to measure the uncertainty associated with its predictions (Wang et al., 2016; Sorscher et al., 2022; Simsek et al., 2022). Specifically, for a given input x with output probabilities $\{p_1, p_2, \ldots, p_n\}$, the entropy $H(x)$ is defined as:

$$H(x) = -\sum_{i=1}^{n} p_i \log p_i, \tag{1}$$

where p_i is the probability assigned by the model to the i-th class, and n is the total number of classes. Higher entropy indicates greater uncertainty and can be interpreted as the model finding the example more difficult to classify.

Pointwise Framework of Learning The Pointwise Framework of Learning introduced by Kaplun et al. (2022) offers a novel perspective by evaluating the performance of a collection of models on *individual* data points rather than averaging performance over a distribution of inputs. This methodology provides a more granular understanding of model performance, revealing how different models perform on specific inputs as resources such as training time, dataset size, and model complexity increase.

The core of this framework is the concept of *learning profiles*: for a given input point z, the learning profile captures the performance of models on z (output probability) as a function of increasing resources. Formally, for an input point $z = (x, y)$ and a family of classifiers T, the learning profile maps the global accuracy of classifiers to their performance on z. This approach allows for the identification of four distinct categories of points: "easy," "hard," "compatible (or monotone)," and "non-monotone." Easy points are those for which even low-accuracy models perform well. Hard points are those for which even high-accuracy models struggle. Monotone points closely track the global accuracy of the models, while non-monotone points exhibit behavior where higher-accuracy models might perform worse on these inputs, and there is no clear pattern of improving performance on individual samples as resources increase. Appendix A shows instances of learning profiles of these categories. Details and a Python implementation for assigning a data point to categories are provided in Appendix B.

3. Experimental Setup

We assessed the learning dynamics of Natural Training (NT) and Adversarial Training (AT) on the CIFAR-10 and CIFAR-100 datasets (Krizhevsky et al., 2009). For each method, we trained five models with different random seed initializations to both reduce the variance in the results and to have a collection of models for defining point categories. We used the SGD optimizer with a Multi-Step learning rate schedule (decaying at epochs 70 and 90), starting with an initial learning rate of 0.1, a batch size of 128, and a weight decay of 1×10^{-4}, for 120 epochs. For AT, adversarial examples were generated using the Projected Gradient Descent (PGD) attack (Madry et al., 2017) with 10 iterations and various epsilon ϵ values, with each attack initialized randomly.

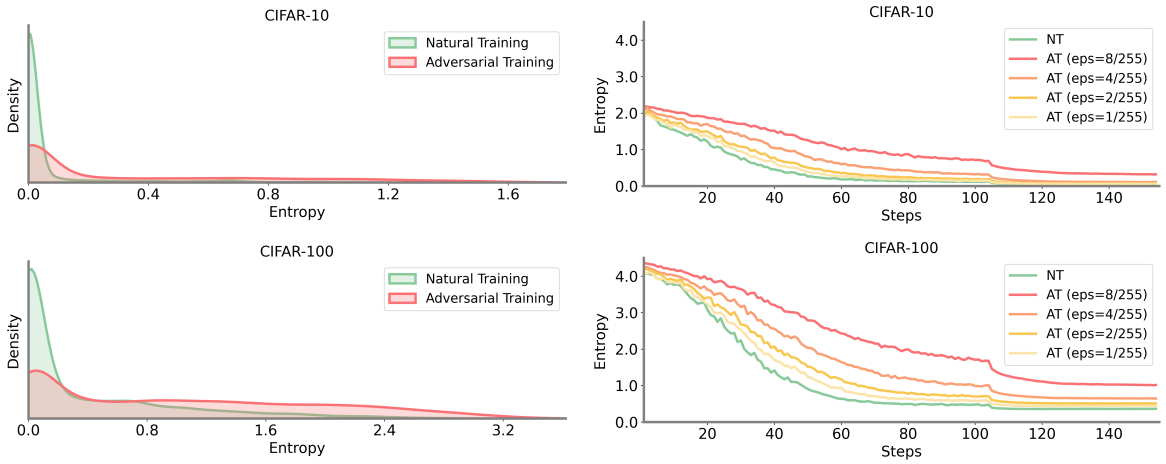
4. Learning Dynamic

We delve into the specifics of AT and NT learning dynamics by examining various methodologies for quantifying example difficulty. We aim to provide a comprehensive view of their different learning dynamics at the sample level. We first focus on entropy as a measure of uncertainty (Section 4.1), then on the pointwise framework of learning for categorizing data points based on their learning profiles (Section 4.2). Later, we show that there is a trade-off between attack strength and monotonicity in learning (Section 4.3). Finally, we test both NT and AT on the specific case study of the CIFAR10-Neg dataset (Kaplun et al., 2022), which involves samples with a negative correlation between accuracy and increased resources (Section 4.4).

4.1. Entropy Analysis

Figure 1a illustrates the entropy distributions of predictions for models trained using NT and AT. The entropy values are higher and more spread out in the AT models compared to the NT models across both datasets. This broader distribution indicates that AT encourages models to maintain more cautious predictions, with probabilities more distributed across classes rather than showing overconfident predictions concentrated on a single class.

Furthermore, as shown in Figure 1b, the evolution of mean entropy over test data points during training varies significantly with different ϵ values of adversarial examples. Notably, models trained with larger ϵ exhibit a slower decrease in entropy over training steps. This gradual reduction in entropy implies that the models trained under AT remain uncertain for longer durations, which could indicate an ongoing engagement with more complex or less obvious features of the training data. This prolonged uncertainty prevents premature convergence on simple, easily perturbable, and non-generalizable features, such as textures or background elements that are not essential for accurate classification.



(a) **Entropy Distribution** (b) **Mean Entropy Evolution**

Figure 1. **Entropy:** Comparison of entropy distribution (a) and mean entropy evolution (b) between Natural Training (NT) and Adversarial Training (AT) for CIFAR-10 (top) and CIFAR-100 (bottom) datasets. The left subfigures show that AT models have higher and more spread-out entropy values, indicating more cautious predictions. The right subfigures depict slower decreases in mean entropy for models trained with larger ϵ values, suggesting prolonged uncertainty and engagement with complex and generalizable features.

Therefore, we can hypothesize that using a larger ϵ not only improves robustness but also provides a dynamic environment that prioritizes the predictivity of features over simplicity in learning. However, as we see in Section 5, this is not always the case; we often observe a performance drop with larger ϵ. We will discuss the root cause and present a promising approach using synthesized data to tackle it.

4.2. Point Categories

Based on Section 2, a *learning profile* for an individual sample can be defined by mapping the average overall performance of a collection of models to their pointwise performance (the probability assigned to the sample for the true class) as resources increase. We examine the case where time (or training steps) increases and categorize all test data points accordingly, as detailed in Appendix B. Figure 2 illustrates the distribution of these categories under Natural Training (NT). Our observations, consistent with (Kaplun et al., 2022), reveal significant levels of non-monotonicity in NT. This non-monotonicity suggests that improvements in overall model accuracy can sometimes lead to degraded performance on specific data points, likely due to the failure to learn stable features.

4.3. Attack Strength and Monotonicity

On the other hand, AT exhibits more monotonic behavior, as shown in Figure 3. This behavior becomes more pronounced with larger ϵ, showing a positive correlation between the ϵ used for AEs and monotonicity. However, this also presents a trade-off between monotonicity and hardness. This learning dynamic could be considered a desired behavior if we

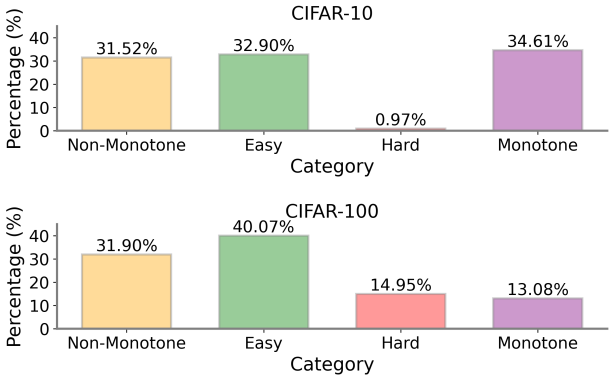


Figure 2. **Point Categories Distribution:** The distribution of point categories for models trained with Natural Training (NT) on the CIFAR-10 (left) and CIFAR-100 (right) datasets.

had not observed an overall performance degradation, as it also indicates less overconfident predictions. Similar to Section 4.1, we desire to use a large attack strength ϵ in practice both for adversarial robustness and to gain the benefits of monotonicity in learning dynamics. In Section 5, we discuss this matter and address this gap by using synthesized data.

4.4. CIFAR10-Neg

According to Miller et al. (2021), there is a strong positive correlation between out-of-distribution (OOD) performance and in-distribution (ID) performance for a wide range of models and distribution shifts. However, Kaplun et al. (2022) challenged this assumption by using learning profiles to create a dataset called CIFAR-10-NEG, which negatively correlated with accuracy on the CIFAR-10 test,

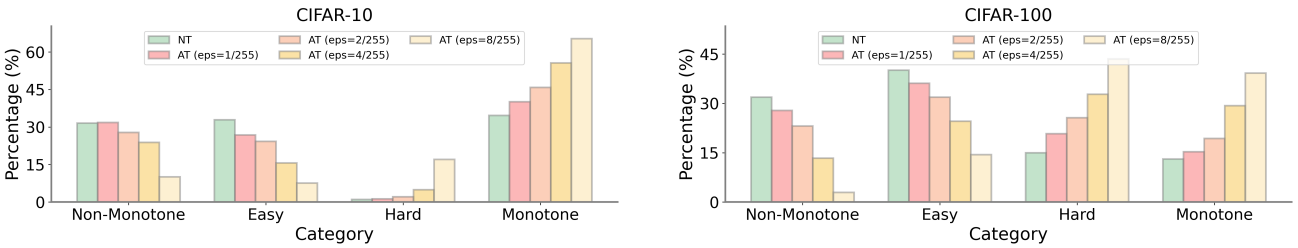


Figure 3. **Attack Strength and Monotonicity:** The percentage of categories for CIFAR-10 (left) and CIFAR-100 (right) using NT and AT with varying perturbation strengths (ϵ). Increasing attack strength correlates with more monotonic behavior, indicating a positive correlation, but also highlights a trade-off between hardness and monotonicity.

thereby showing a reverse correlation between ID and OOD performance for the first time. This dataset, a subset of CINIC-10 (Darlow et al., 2018) samples with the highest non-monotonicity score, revealed intriguing learning dynamics where the models' performance initially improves up to 60% but then starts to decline to 20% as the overall performance on the ID set continues to improve.

Figure 4 shows our analysis of evaluating a collection of NT and AT models on CIFAR-10-NEG dataset. We observed two phases of accuracy evolution. In the first phase, consistent with (Kaplun et al., 2022), we observe a rise and decline in performance on CIFAR-10-neg until it reaches about 25% accuracy. However, in the second phase, as we continue further training, we see a rise in performance again, though lower than its peak in the first phase. Our analysis shows a complementary observation to (Kaplun et al., 2022), and we speculate the second rise in performance is due to training models until convergence.

Despite these complementary results, we tested AT models and observed a trend of more stable performance improvement. Larger ϵ values used for AEs led to gains in both stable accuracy improvement on CIFAR-10-NEG and higher accuracy overall. This supports our insight that AEs provide a dynamic improvement for learning stable features that can enhance both robustness and generalization.

5. Benefit of Synthesized Data on AT

Thus far, our observations indicate that incorporating adversarial examples results in less overconfident predictions due to higher entropy (Section 4.1) and more monotonic learning behavior (Section 4.2). However, as previously mentioned, increasing the attack strength ϵ in AT leads to an overall performance drop, despite increasing robustness. This issue poses a limitation to the practical usage of AT. For example, Kireev et al. (2021) highlights the potential of AEs in enhancing out-of-distribution (OOD) generalization of models on common data corruptions in CIFAR-10-C. However, they demonstrated a trade-off between attack strength and improvement gain. Notably, for this dataset, the optimal

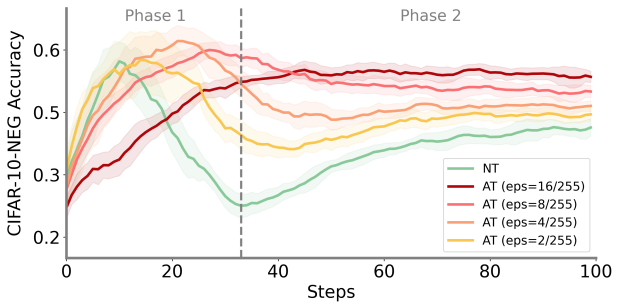


Figure 4. **CIFAR-10-NEG:** Performance evaluation of NT and AT models on the CIFAR-10-NEG dataset. It illustrates two distinct phases of accuracy evolution. In Phase 1, for NT, there is a rise and subsequent decline in performance consistent with (Kaplun et al., 2022). In Phase 2, performance increases again but remains lower than the initial peak, showing complementary results. On the other hand, for AT models, larger ϵ values result in consistent performance improvement, showing more stable behavior.

benefit is achieved with a small ϵ value of $1/255$. However, this is significantly smaller than the commonly used ϵ value of $8/255$ for building robust models.

The performance drop associated with higher ϵ values has been extensively studied in the literature, with *robust overfitting* identified as the primary issue (Yu et al., 2022). Increasing attack strength exacerbates this problem. Various approaches have been developed to mitigate robust overfitting, but one of the most promising solutions is the use of synthesized or generated data (Gowal et al., 2021; Wang et al., 2023; Bartoldson et al., 2024), which effectively increases the number of training samples.

In our study, we leverage synthesized data generated by diffusion models provided by (Wang et al., 2023). Our experimental setup remains consistent, with each training batch comprising half original dataset samples and half generated data. This method incurs no additional computational overhead since the training process is conducted over a fixed number of batches per epoch. For the rest of the settings, we followed our initial setup mentioned in Section 3.

Figure 5 illustrates the benefits of synthesized data in en-

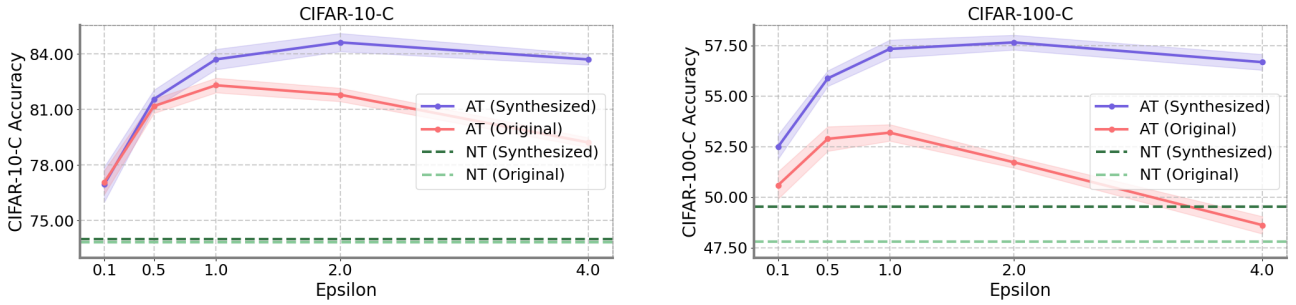


Figure 5. **CIFAR-C Datasets.** Accuracy of NT and AT models on CIFAR-10-C (left) and CIFAR-100-C (right) at various ϵ levels. AT models with synthesized data (blue line) consistently outperform those with original data (red line), especially at higher ϵ values. NT models (light green and dark green dashed lines) show slight improvements with synthesized data. Synthesized data effectively mitigates robust overfitting in AT, enhancing robustness and generalization. Further details on individual corruptions can be found in Appendix C.

hancing the performance of AT models on CIFAR-10-C and CIFAR-100-C. These datasets are extensions of CIFAR-10 and CIFAR-100, respectively, designed to evaluate model performance on common image corruptions. They consist of test set images processed with 19 different types of corruptions, such as noise, blur, weather effects, and digital distortions, applied at five levels of severity.

We do not observe significant performance improvements when using synthesized data for NT, with almost no enhancement on CIFAR-10-C and only a slight improvement on CIFAR-100-C. However, AT benefits greatly from synthesized data, showing substantial improvements over both NT and AT with original data. Notably, with synthesized data, we can use larger ϵ values for AEs, resulting in more robust models while maintaining high accuracy on corrupted data. This suggests that mitigating robust overfitting with additional data is particularly effective, allowing us to create models that not only remain robust but also enhance generalization. This opens new possibilities for developing more reliable models for real-world applications.

6. Related Work

Adversarial Training and Examples Adversarial Training (AT), is the standard for enhancing neural network robustness against adversarial examples (AEs) (Goodfellow et al., 2014). Subsequent research has refined AT, developing stronger attacks and defenses (Madry et al., 2017; Zhang et al., 2019). However, using synthesized data in AT is a promising solution to mitigate robust overfitting and improve robustness (Gowal et al., 2021; Wang et al., 2023; Bartoldson et al., 2024). Beyond robustness, AEs have also been used to enhance model performance. Xie et al. (2020) showed that augmenting training data with AEs improves image recognition accuracy. Similarly, Rebuffi et al. (2023) and Alhamoud et al. (2022) found that AEs boost generalization. This dual use of AEs highlights their potential for developing more resilient and generalizable models.

Example Difficulty Quantifying sample difficulty is crucial for improving model performance and interpretability (Baldock et al., 2021; Cui et al., 2023a). For example, Cui et al. (2023a) show that by penalizing overconfident predictions based on sample difficulty, we can enhance model accuracy. One simple approach is entropy estimation from information theory, which measures prediction uncertainty and provides insights into model difficulty (Shannon, 2001; Wang et al., 2016; Sorscher et al., 2022; Simsek et al., 2022). Another approach is the pointwise framework of learning by Kaplun et al. (2022), which evaluates the performance of models on individual data points and introduces learning profiles that can be used to measure difficulty.

7. Conclusion & Discussion

We explored the impact of adversarial examples on the learning dynamics of neural networks, focusing on both robustness and generalization. Our findings indicate that Adversarial Training (AT) leads to more stable learning dynamics compared to Natural Training (NT), resulting in gradual performance improvements and less overconfident predictions. This suggests that AT encourages models to focus on more both resilient and generalizable features. We identified a trade-off between adversarial robustness and generalization gains, primarily due to robust overfitting, which limits the practical deployment of AT. To address this issue, we used synthesized data into the training process. Our results demonstrated that AT benefits significantly from extra data, which enhances generalization without compromising robustness. This approach offers new avenues for developing robust and generalizable models for real-world applications.

Future work should extend the use of synthesized data to other types of out-of-distribution (OOD) tasks to validate its effectiveness across different scenarios. Additionally, testing our findings with other variations of adversarial training could provide deeper insights and further optimize the balance between robustness and generalization.

Impact Statement

Adversarial examples pose significant challenges to neural networks, undermining their reliability in critical applications like autonomous vehicles and healthcare. While Adversarial Training (AT) has been effective in enhancing model robustness, it often compromises generalization capabilities. Our study addresses this tradeoff by incorporating publicly available synthesized data, enabling the development of models that are both robust and generalizable. This improvement is vital for ensuring the safety and reliability of AI systems and accelerating the adoption of AI technologies for a safer society.

8. Acknowledgements

We acknowledge the support from the Canada CIFAR AI Chair Program and from the Canada Excellence Research Chairs (CERC) Program. This research was enabled in part by computational resources provided by the Digital Research Alliance of Canada and Mila Quebec AI Institute.

References

Alhamoud, K., Hammoud, H. A. A. K., Alfarra, M., and Ghanem, B. Generalizability of adversarial robustness under distribution shifts. *arXiv preprint arXiv:2209.15042*, 2022.

Apostolidis, K. D. and Papakostas, G. A. A survey on adversarial deep learning robustness in medical image analysis. *Electronics*, 10(17):2132, 2021.

Baldock, R., Maennel, H., and Neyshabur, B. Deep learning through the lens of example difficulty. *Advances in Neural Information Processing Systems*, 34:10876–10889, 2021.

Bartoldson, B. R., Diffenderfer, J., Parasyris, K., and Kailkhura, B. Adversarial robustness limits via scaling-law and human-alignment studies. *arXiv preprint arXiv: 2404.09349*, 2024.

Bortsova, G., González-Gonzalo, C., Wetstein, S. C., Dubost, F., Katramados, I., Hogeweg, L., Liefers, B., van Ginneken, B., Pluim, J. P., Veta, M., et al. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Medical Image Analysis*, 73:102141, 2021.

Cui, P., Zhang, D., Deng, Z., Dong, Y., and Zhu, J. Learning sample difficulty from pre-trained models for reliable prediction. *Advances in Neural Information Processing Systems*, 36:25390–25408, 2023a.

Cui, X., Aparcedo, A., Jang, Y. K., and Lim, S.-N. On the robustness of large multimodal models against image adversarial attacks. *arXiv preprint arXiv: 2312.03777*, 2023b.

Darlow, L. N., Crowley, E. J., Antoniou, A., and Storkey, A. J. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv: 1810.03505*, 2018.

Deng, Z., Zhang, L., Vodrahalli, K., Kawaguchi, K., and Zou, J. Y. Adversarial training helps transfer learning via better representations. *Advances in Neural Information Processing Systems*, 34:25179–25191, 2021.

Dong, Y., Chen, H., Chen, J., Fang, Z., Yang, X., Zhang, Y., Tian, Y., Su, H., and Zhu, J. How robust is google's bard to adversarial image attacks? *arXiv preprint arXiv: 2309.11751*, 2023.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gowal, S., Rebuffi, S.-A., Wiles, O., Stimberg, F., Calian, D. A., and Mann, T. A. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. *AISEC@CCS*, 2023. doi: 10.1145/3605764.3623985.

Hirano, H., Minagi, A., and Takemoto, K. Universal adversarial attacks on deep neural networks for medical image classification. *BMC medical imaging*, 21:1–13, 2021.

Kaplun, G., Ghosh, N., Garg, S., Barak, B., and Nakkiran, P. Deconstructing distributions: A pointwise framework of learning. *arXiv preprint arXiv:2202.09931*, 2022.

Kireev, K., Andriushchenko, M., and Flammarion, N. On the effectiveness of adversarial training against common corruptions. *Conference On Uncertainty In Artificial Intelligence*, 2021.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., and Lu, F. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332, 2021.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, pp. 7721–7735. PMLR, 2021.

Rebuffi, S.-A., Croce, F., and Gowal, S. Revisiting adapters with adversarial training. *arXiv preprint arXiv:2210.04886*, 2022.

Rebuffi, S.-A., Wiles, O., Shelhamer, E., and Gowal, S. Adversarially self-supervised pre-training improves accuracy and robustness. 2023.

Shannon, C. E. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

Simsek, B., Hall, M., and Sagun, L. Understanding out-of-distribution accuracies through quantifying difficulty of test samples. *arXiv preprint arXiv:2203.15100*, 2022.

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Wang, K., Zhang, D., Li, Y., Zhang, R., and Lin, L. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016.

Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., and Yan, S. Better diffusion models further improve adversarial training. *International Conference on Machine Learning*, 2023. doi: 10.48550/arXiv.2302.04638.

Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A., and Le, Q. V. Adversarial examples improve image recognition. *Computer Vision and Pattern Recognition*, 2019. doi: 10.1109/cvpr42600.2020.00090.

Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A. L., and Le, Q. V. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 819–828, 2020.

Xiong, Z., Xu, H., Li, W., and Cai, Z. Multi-source adversarial sample attack on autonomous vehicles. *IEEE Transactions on Vehicular Technology*, 70(3):2822–2835, 2021.

Yu, C., Han, B., Shen, L., Yu, J., Gong, C., Gong, M., and Liu, T. Understanding robust overfitting of adversarial training and beyond. *International Conference On Machine Learning*, 2022. doi: 10.48550/arXiv.2206.08675.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Zhang, Y. and Sang, J. Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 4346–4354, 2020.

Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M. M., and Lin, M. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024.

A. Point Categories

We provide additional details on the point categories identified in the Figure 6. The learning profiles depicted in Figure 6 represent the behavior of models on specific data points across various stages of training. These profiles help to categorize data points into four types:

- **Easy Points:** These are points that models can learn to classify correctly very early in the training process. Even models with lower overall accuracy perform well on these points.
- **Hard Points:** These points are challenging for models to classify correctly. High-accuracy models still struggle with these points, indicating inherent difficulty in learning from these examples.
- **Monotone (Compatible) Points:** These points show a clear trend where the performance of models improves consistently with more training. The accuracy on these points generally follows the global accuracy trend of the model.
- **Non-Monotone Points:** For these points, the performance does not follow a consistent pattern. In some cases, higher-accuracy models might perform worse on these points, and improvements are not straightforward as training progresses.

These categories provide insights into the behavior of models on individual data points, offering a more nuanced understanding of model performance beyond aggregate metrics.

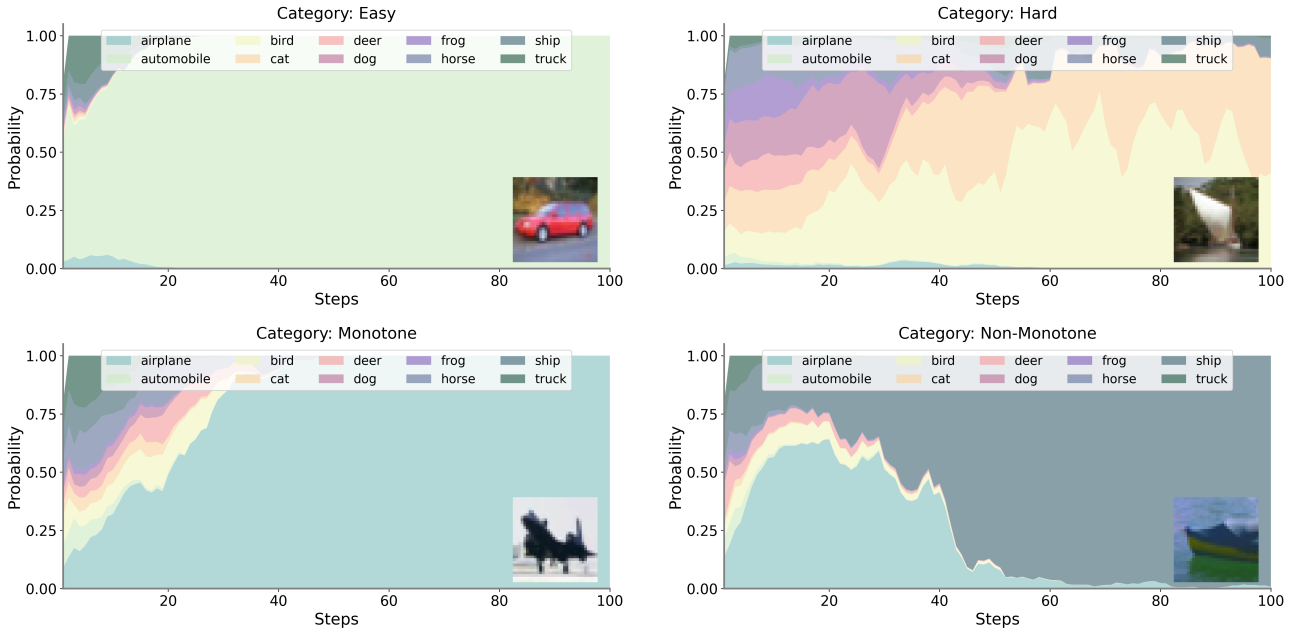


Figure 6. **Learning Profiles:** The figure illustrates the learning profiles of four distinct categories of data points: Easy, Hard, Monotone (Compatible), and Non-Monotone. The width of each color represents the probability assigned to the corresponding class. **Top Left:** Easy points where models quickly achieve high accuracy. **Top Right:** Hard points where models struggle to achieve high accuracy even with extensive training. **Bottom Left:** Monotone points where performance consistently improves with more training steps. **Bottom Right:** Non-Monotone points where performance varies irregularly with more training steps, sometimes decreasing despite overall model improvements.

B. Assigning Samples to Categories

In the Pointwise Framework of Learning introduced by Kaplun et al. (2022), data points are categorized based on their *learning profile*. To assign each data point to a category, we monitor the performance drops of models on that data point throughout the training process from one step to another. If the amount of drop in performance exceeds a certain threshold, the data point is categorized as "non-monotone," indicating irregular performance where higher accuracy models might perform worse on these inputs. For data points that do not exhibit significant performance drops, we compare their learning profiles to predefined profile templates. These templates are as follows: "easy" points have consistently high performance (denoted by $p = 1$), "hard" points have consistently low performance (denoted by $p = 0$), and "compatible" points have performance that tracks the global average accuracy of the models (denoted by $p = c$, where c is the confidence of the model's prediction for the true class). By matching the data point's learning profile to these templates, we can classify it into the most similar group, providing a nuanced understanding of model performance on individual data points.

```python
import numpy as np

def point_category(avg_accuracies, avg_y_hats, y, threshold):
    """
    Assign a category to a sample based on the confidence of the models on a sample throughout the training.

    :param avg_accuracies: The average accuracy of the models on a sample throughout the training.
    :param avg_y_hats: The average predicted probabilities of the models on a sample throughout the training.
    :param y: The true label of a sample.
    :param threshold: The threshold for the confidence drop to be considered non-monotone.
    :return: The category of a sample.
    """

    # Calculate the total confidence drop
    confidences = avg_y_hats[:, y]
    confidence_drop = np.diff(confidences)
    total_confidence_drop = np.sum(np.abs(confidence_drop[confidence_drop < 0]))

    # If the total confidence drop is greater than the threshold, assign the category as non-monotone
    if total_confidence_drop > threshold:
        return 'non-monotone'

    # Calculate the distance of the confidences to the template learning profiles (easy, hard, monotone)
    distances = {
        'easy': (np.ones_like(confidences) - confidences).sum(),
        'hard': (confidences - np.zeros_like(confidences)).sum(),
        'monotone': np.abs(confidences - avg_accuracies).sum()
    }

    # Assign the category as the one with the minimum distance
    return min(distances, key=distances.get)
```

C. CIFAR10-C and CIFAR-100-C Individual Corruptions

Figures 7 and 8 show the accuracies of NT and AT models for different corruption types for CIFAR-10-C and CIFAR-100-C, respectively, showing the significant benefit of synthesized data for AT.

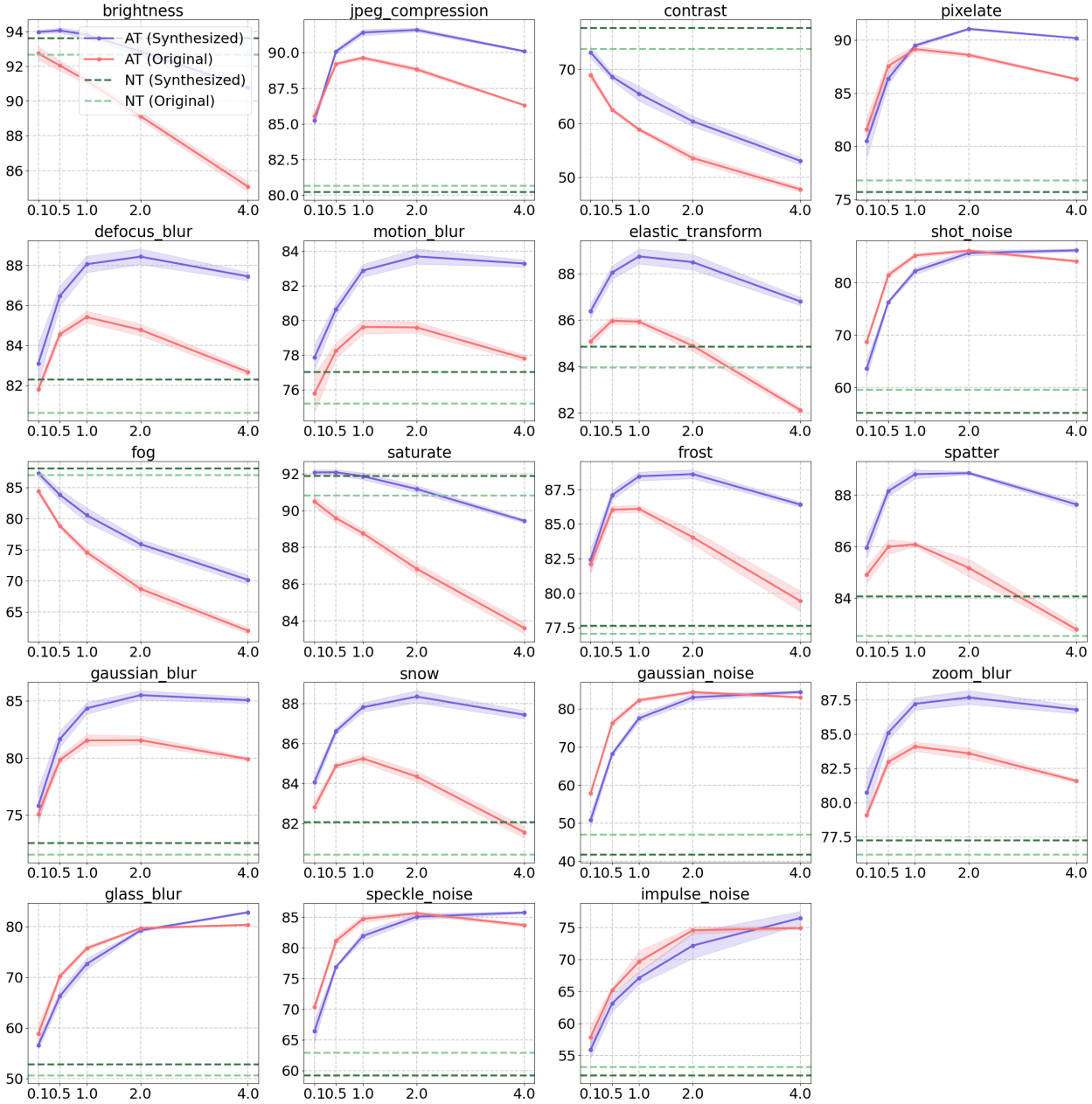


Figure 7. **CIFAR10-C Dataset:** Individual corruptions accuracies of NT and AT models on CIFAR-10-C at various ϵ levels.

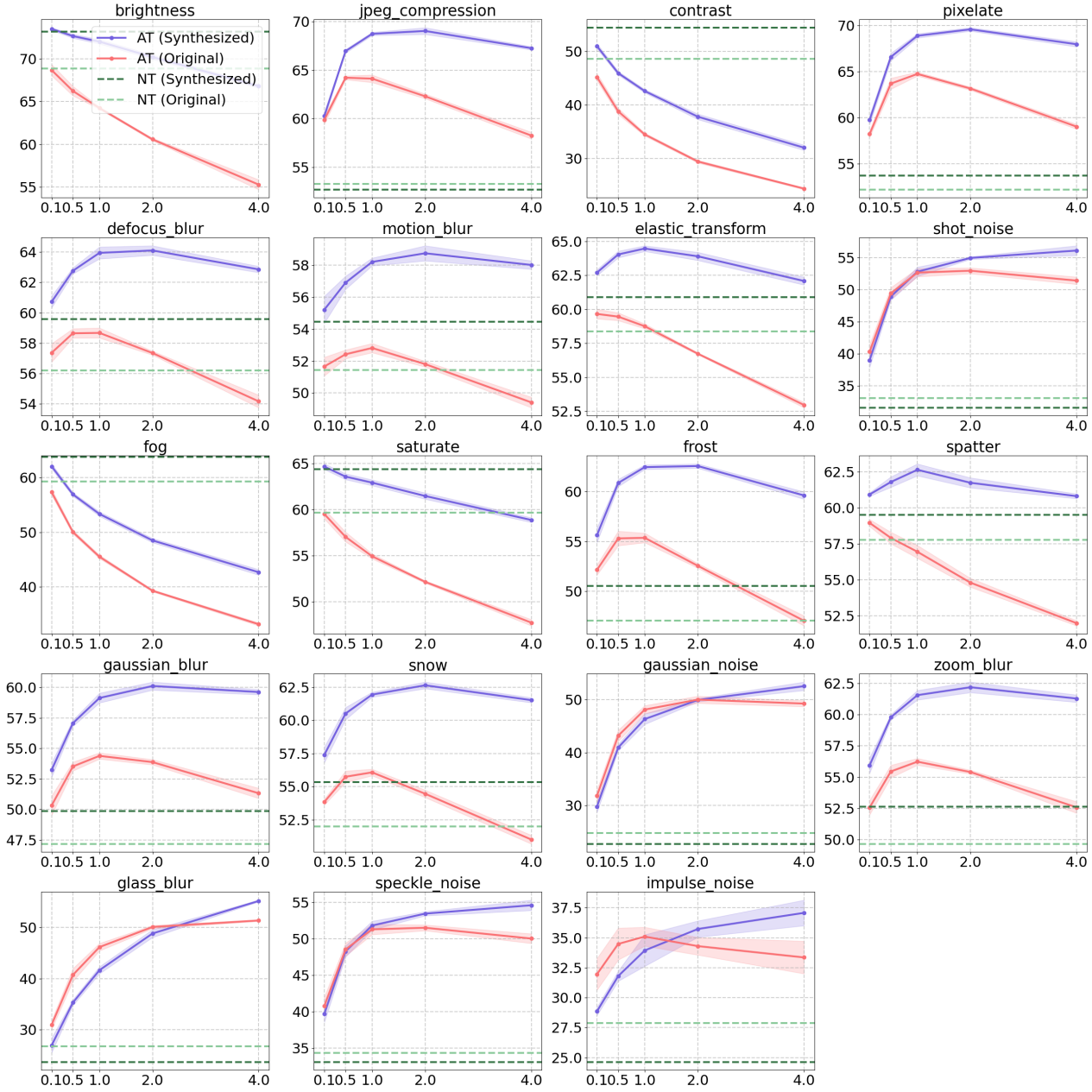


Figure 8. **CIFAR100-C Dataset:** Individual corruptions accuracies of NT and AT models on CIFAR-100-C at various ϵ levels.