Group ID 15
Murat Çevlik – 152120201067
Onur Dalgıç – 15212021068

Adversarial Robustness Analysis on CIFAR-10 using ResNet-18

1. Introduction

Deep neural networks (DNNs) have demonstrated strong performance across a wide range of computer vision tasks, including image classification, object detection, and autonomous systems. However, despite their success on clean data, these models are known to be highly sensitive to adversarial examples—inputs that contain small, carefully designed perturbations capable of causing incorrect predictions while remaining largely imperceptible to the human eye. This sensitivity poses notable concerns, especially in safety-critical domains such as autonomous driving, biometric recognition, and medical image analysis, where even minor prediction errors may lead to serious consequences.

1.1. Problem Definition

Neural networks trained exclusively on clean data generally exhibit limited robustness when exposed to adversarial attacks such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). Under these conditions, even small input perturbations can result in significant drops in classification performance. Consequently, improving adversarial robustness has emerged as an important and ongoing challenge in the field of modern machine learning.

1.2 Goals and Motivation

This project aims to evaluate several commonly used adversarial defense strategies on the CIFAR-10 dataset using a customized ResNet-18 architecture. The primary motivation is to analyze how different training approaches—ranging from single-step FGSM-based methods to multi-step adversarial training techniques—impact both clean accuracy and adversarial robustness. In addition, the study investigates whether a fine-tuning strategy that incorporates stronger adversarial attacks and regularization techniques can lead to improved performance compared to standard PGD or TRADES-based training.

1.3 Intended Contributions

1. Implementation of a modified ResNet-18 architecture tailored for the CIFAR-10 dataset.
2. Comparative analysis of seven different adversarial training strategies under consistent experimental settings.
3. Examination of a fine-tuning approach that combines PGD-10 adversarial training with label smoothing to enhance robustness at higher perturbation levels.
4. Evaluation of model robustness under increasing FGSM attack strengths ($\varepsilon = 2/255$, $4/255$, and $8/255$).

2. Related Work

Publication Title	Dataset	Preprocessing	Method	Objective
Adversarial Weight Perturbation Helps Robust Generalization – Wu et al. (2020)	CIFAR-10, CIFAR-100, SVHN	Input and weight perturbation	AWP with adversarial training	To improve robust generalization through weight perturbations during training.
Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks – Croce & Hein (2020)	CIFAR-10, MNIST	Standard normalization	AutoAttack (ensemble)	To provide a standardized and reliable framework for robustness evaluation.
Adversarial Attacks and Defenses in Deep Learning – Ren et al. (2020)	MNIST, CIFAR-10, ImageNet	Literature-based	Survey	To review existing adversarial attack methods and defense techniques.
A Survey on Adversarial Attacks and Defences – Chakraborty et al. (2021)	MNIST, CIFAR-10, ImageNet	Literature-based	Taxonomy / Survey	To categorize threat models and defense strategies in adversarial machine learning.
Improving Robustness using Generated Data – Gowal et al. (2021)	CIFAR-10, CIFAR-100	Generated data augmentation	Robust training with synthetic data	To enhance robustness through the use of high-quality generated samples.
Adversarial Robustness via Label Smoothing Revisited – Bai et al. (2023)	CIFAR-10, CIFAR-100	Label smoothing	Regularization-based defense	To analyze the impact of label smoothing on adversarial robustness.
Adversarial Robustness Analysis on CIFAR-10 using ResNet-18 – Murat Çevlik & Onur Dalgıç (2025)	CIFAR-10	Normalization (μ, σ), random crop (pad=4), random horizontal flip (p=0.5)	FGSM, PGD-7, PGD-10, TRADES, label smoothing, fine-tuning	To examine the trade-off between clean accuracy and robustness across multiple adversarial training strategies.

Table 1

3. Methodology

This study focuses on the implementation and evaluation of different adversarial training strategies aimed at improving the robustness of deep neural networks (DNNs) against L_∞-bounded perturbations. All experiments were conducted using the PyTorch deep learning framework.

3.1. Dataset and Preprocessing

The CIFAR-10 dataset was used in all experiments. It consists of 60,000 color images with a resolution of 32×32pixels, evenly distributed across 10 different classes. To support stable training and improve generalization, the following preprocessing steps were applied:

- **Normalization:**
  All images were normalized using the channel-wise mean and standard deviation computed from the CIFAR-10 training set. The normalization parameters were set to
  $$\mu = (0.4914, 0.4822, 0.4465), \sigma = (0.2023, 0.1994, 0.2010).$$
- **Data Augmentation:**
  During training, standard data augmentation techniques were employed, including random cropping with a padding of 4 pixels and random horizontal flipping with a probability of 0.5.

3.2. Model Architecture

A modified ResNet-18 architecture was implemented and adapted specifically for the 32×32input resolution of the CIFAR-10 dataset. Instead of the standard ImageNet-style input layer, which typically employs a 7×7convolution with a stride of 2, the model uses a 3×3convolution with stride 1 and padding 1. This design choice helps preserve the spatial resolution of small input images and reduces information loss in the early layers. The network maintains the standard residual block structure, along with Batch Normalization layers and global average pooling for the final classification stage.

3.3. Attack Algorithms (Threat Model)

The models were trained and evaluated using three commonly applied first-order adversarial attack methods, all operating under an L_∞-bounded threat model with perturbation magnitude ε.

- **Fast Gradient Sign Method (FGSM):**
  FGSM is a single-step adversarial attack that generates adversarial examples by perturbing the input in the direction of the sign of the gradient of the loss with respect to the input. The adversarial example $x_{\text{adv}}$is computed as:
  $$x_{\text{adv}} = x + \varepsilon \cdot \text{sign}(\nabla_x \mathcal{L}(f(x), y)).$$

- **Projected Gradient Descent (PGD):**
  PGD is an iterative extension of FGSM that repeatedly applies small perturbations while projecting the perturbed input back into the L_∞ ball centered at the original input x. In this study, PGD with 7 iterations (PGD-7) was used during adversarial training, while PGD with 10 iterations (PGD-10) was employed during the fine-tuning phase. The update rule is given by:

$$x_{t+1} = \mathcal{B}_\varepsilon(x)[x_t + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(f(x_t), y))],$$

  where $\mathcal{B}_\varepsilon(x)$ denotes the projection onto the ε-bounded L_∞ ball.

- **TRADES Loss (Theoretically Robust Adversarial Defense):**
  TRADES is a defense strategy that explicitly balances clean accuracy and adversarial robustness through a regularized loss formulation. The loss consists of a standard cross-entropy term on clean inputs and a Kullback–Leibler (KL) divergence term that encourages consistency between predictions on clean and adversarial examples:

$$\mathcal{L} = \mathcal{L}_{\text{ce}}(x, y) + \beta \cdot \text{KL}(f(x) \parallel f(x_{\text{adv}})).$$

  Following the original formulation, the trade-off parameter β was set to 6.0 in all TRADES-based experiments.

3.4. Training Strategies

Seven different models were trained to compare the behavior of various adversarial training strategies under consistent experimental settings. Each model follows a distinct training protocol, as summarized below.

ID	Strategy Name	Description
Model 1	Teacher	Standard training using only clean samples without adversarial perturbations.
Model 2	Pure FGSM	Training performed exclusively on FGSM-generated adversarial examples with (\varepsilon = 8/255).
Model 3	Hybrid	Training batches composed of an equal mixture of clean samples and FGSM adversarial examples with (\varepsilon = 8/255).
Model 4	FGSM Mixed v2	Similar to the hybrid approach, but using FGSM adversarial examples with a lower perturbation magnitude of (\varepsilon = 4/255).
Model 5	PGD Training	Adversarial training conducted using PGD with 7 iterations (PGD-7).
Model 6	TRADES Training	Adversarial training based on the TRADES loss formulation with (\beta = 6.0).
Model 7	Fine-Tuned Model	Model initialized from the PGD-trained baseline (Model 5) and further fine-tuned using PGD-10 adversarial examples, combined with label smoothing ((\varepsilon = 0.1)) and a cosine annealing learning rate scheduler ((T_{\text{max}} = 20)).
Model 8	Curriculum PGD Training	Adversarial training using PGD-7 with a curriculum schedule, where the perturbation magnitude is gradually increased up to $\varepsilon = 8/255$.

Table 2

4. Results and Discussion

The trained models were evaluated under increasing adversarial perturbation budgets ε, expressed in units of $1/255$, to analyze their behavior across different attack strengths.

4.1. Quantitative Evaluation

The classification accuracy of all seven models was evaluated under four different conditions: clean inputs and adversarial perturbations with $\varepsilon = 2/255$, $4/255$, and $8/255$. The results are summarized in table below, which provides a direct comparison of clean accuracy and robustness across increasing attack strengths.

Model	**Clean**	$\varepsilon = 2/255$	$\varepsilon = 4/255$	$\varepsilon = 8/255$
Model 1	93.65	33.50	22.27	17.14
Model 2	91.86	79.52	64.46	40.64
Model 3	81.81	72.74	62.93	44.68
Model 4	86.67	74.94	61.02	36.24
Model 5	74.00	67.69	60.89	47.43
Model 6	72.65	66.10	58.86	45.92
Model 7	78.20	71.63	64.41	51.23
Model 8	90.63	%81.23	%73.42	60.42

Table 3

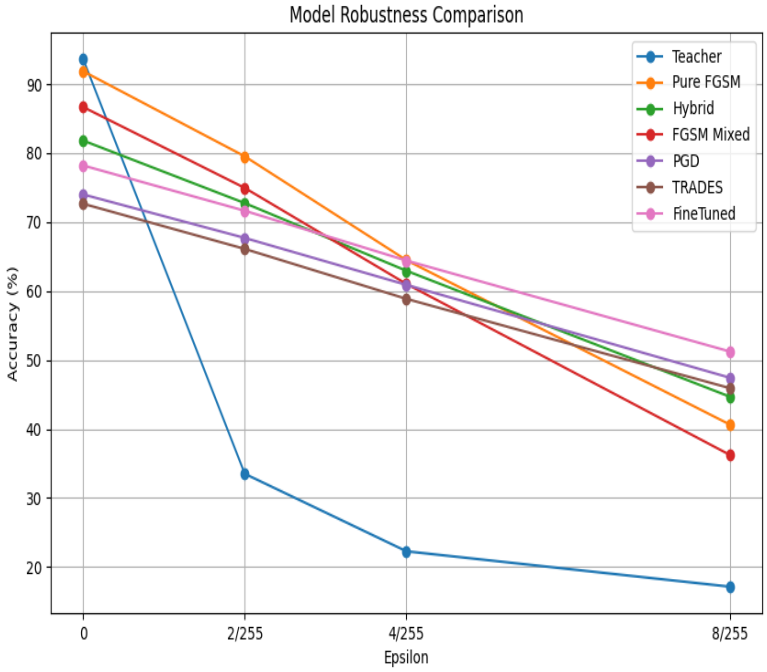

Figure 1

4.2. Discussion

The results highlight the well-known trade-off between clean accuracy and adversarial robustness that arises in adversarial training settings.

- **Teacher Model (M1):**
  The baseline model trained only on clean data achieved the highest clean accuracy (93.65%), indicating strong performance on unperturbed inputs. However, its accuracy dropped substantially under adversarial perturbations, reaching 17.14% at $\varepsilon = 8/255$. This behavior reflects the sensitivity of standard training approaches to adversarial noise.
- **Single-Step Defenses (M2, M3, M4):**
  Models trained using single-step FGSM-based strategies showed improved robustness compared to the Teacher model at lower perturbation levels. For example, Model 2 achieved 79.52% accuracy at $\varepsilon = 2/255$. Nevertheless, their performance declined as perturbation strength increased, indicating limited robustness against stronger, multi-step attacks.
- **PGD and TRADES (M5, M6):**
  Both PGD-based training and TRADES resulted in more consistent robustness across increasing perturbation budgets. At $\varepsilon = 8/255$, Model 5 slightly outperformed Model 6 (47.43% versus 45.92%), while both models exhibited a noticeable reduction in clean accuracy, remaining in the range of approximately 72–74%.
- **Fine-Tuned Model (M7):**
  The fine-tuned model, initialized from the PGD-trained baseline and further trained using PGD-10 adversarial examples with label smoothing, demonstrated a balanced performance profile. It achieved the highest accuracy under the strongest perturbation level ($\varepsilon = 8/255$, 51.23%) while also maintaining a higher clean accuracy (78.20%) compared to the base PGD model. These results indicate that the applied fine-tuning strategy can improve robustness while partially preserving clean-data performance.
- **Curriculum PGD Training (M8):** Model 8 extends the adversarial training analysis by incorporating both a curriculum-based training strategy and a wider convolutional architecture. Unlike Models 1–7, which are based on a modified ResNet-18 backbone, Model 8 employs a WideResNet-28-10 architecture and is trained using PGD-7 adversarial examples with a progressively increasing perturbation magnitude. This combination results in a substantially higher clean accuracy (90.63%) compared to PGD- and TRADES-based ResNet-18 models, while also achieving strong robustness across all evaluated attack strengths. In particular, Model 8 attains 60.42% accuracy under the strongest FGSM perturbation ($\varepsilon = 8/255$), outperforming all other evaluated models by a significant margin. These results suggest that the use of a wider network architecture, together with a curriculum-based adversarial training scheme, can significantly improve the robustness–accuracy trade-off beyond what is achievable with standard PGD or fine-tuning strategies on narrower architectures.

4.3. Encountered Problems and Solutions

During the experimental process, several practical challenges were observed and addressed to ensure stable training and consistent evaluation.

- **Training Stability in TRADES:**
  In the early stages of TRADES-based training, minor convergence issues were observed, which were related to the interaction between the loss formulation and Batch Normalization layers. This issue was mitigated by carefully controlling the use of model.train() and model.eval() modes during adversarial example generation and loss computation.
- **Computational Cost:**
  Adversarial training with multi-step attacks, particularly PGD-based methods, required substantially more computational resources compared to standard training. To manage this overhead, batch sizes were adjusted and

learning rate scheduling techniques, such as cosine annealing, were applied to maintain stable convergence within a reasonable training time.

- **FGSM-Based Training Behavior:**
  Models trained exclusively with single-step FGSM attacks showed limited generalization as the perturbation strength increased. This behavior was alleviated by incorporating hybrid training strategies and transitioning to multi-step adversarial training approaches, which resulted in more consistent robustness across different attack settings.
- **Training Stability in Curriculum PGD (Model 8):** During the curriculum-based adversarial training of Model 8, instability was observed when stronger perturbation levels were introduced abruptly. This issue was addressed by gradually increasing the perturbation magnitude over training epochs, allowing the model to adapt incrementally to more challenging adversarial examples. The curriculum schedule resulted in smoother convergence and improved robustness compared to directly training with the target perturbation strength.
- **Architecture-Dependent Computational Overhead:** The use of a wider network architecture (WideResNet-28-10) in Model 8 significantly increased memory consumption and training time relative to ResNet-18-based models. To accommodate this overhead, gradient clipping was applied to stabilize optimization, and evaluation frequency for computationally expensive robustness checks was reduced. Despite the increased cost, the improved robustness–accuracy trade-off justified the use of the wider architecture.

5. Conclusion

This study examined the behavior of deep learning models under adversarial conditions by evaluating multiple adversarial training strategies on the CIFAR-10 classification task. Using a modified ResNet-18 architecture, the experiments highlighted how different training protocols influence model performance when exposed to gradient-based perturbations. In addition, an extended experiment employing a wider network architecture was conducted to further investigate the robustness–accuracy trade-off under curriculum-based adversarial training.

The results indicate that models trained solely on clean data are considerably sensitive to adversarial noise, while adversarial training methods can improve robustness to varying degrees. Among the ResNet-18-based approaches, PGD-based adversarial training provided a stable baseline in terms of robustness, albeit with a noticeable reduction in clean accuracy. The fine-tuning strategy built on top of the PGD-trained model, which incorporated stronger adversarial examples and label smoothing, exhibited a more balanced performance profile by improving robustness at higher perturbation levels while maintaining reasonable clean accuracy, achieving a robust accuracy of 51.23% at $\varepsilon = 8/255$. Furthermore, the curriculum-based adversarial training approach implemented with WideResNet-28-10 architecture demonstrated a substantial improvement in both clean accuracy and adversarial robustness. This model achieved the highest overall performance across all evaluated perturbation levels, reaching 60.42% accuracy under the strongest FGSM attack ($\varepsilon = 8/255$) while maintaining a clean accuracy of 90.63%. These findings suggest that combining curriculum-based perturbation scheduling with a wider network architecture can significantly improve the robustness–accuracy trade-off beyond what is achievable with standard adversarial training on narrower architectures.

Overall, the results emphasize that adversarial robustness is jointly influenced by both training strategies and model architecture. While adversarial training remains essential for improving robustness, architectural capacity and training dynamics play a critical role in determining the final performance. Future work may extend this analysis by performing architecture-controlled comparisons under identical training protocols, as well as by incorporating stronger evaluation frameworks and adaptive attack strategies to further assess robustness under more diverse threat models.

References

- **Bai, Y., Chen, X., Zhang, H., & Wang, Y. (2023).** Adversarial robustness via label smoothing revisited. *International Conference on Machine Learning (ICML).*
- **Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015).** Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR).*
- **Gowal, S., Rebuffi, S.-A., Kolesnikov, A., & Lucic, M. (2021).** Improving robustness using deep ensembles. *Neural Information Processing Systems (NeurIPS).*
- **Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018).** Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR).*
- **Pang, T., Yang, X., Dong, Y., Xu, T., & Su, H. (2020).** Boosting adversarial training with confidence calibration. *Neural Information Processing Systems (NeurIPS).*
- **Rice, L., Wong, E., & Kolter, J. Z. (2020).** Overfitting in adversarially robust deep learning. *International Conference on Machine Learning (ICML).*
- **Sarih, A. E., Aneja, N., & Hong, O. W. (2025).** Certified accuracy and robustness: How different architectures stand up to adversarial attacks.