Lightweight Neural Networks for Adversarial Defense: A Novel NTK-Guided Pruning Approach

1st Akhila Reddy Yadulla
Dept. of Information Technology
University of the Cumberlands
Williamsburg, KY, USA
akhilareddyyadulla@ieee.org

2nd Bhargavi Konda
Dept. of Information Technology
University of the Cumberlands
Williamsburg, KY, USA
bhargavikonda@ieee.org

3rd Mounica Yenugula
Dept. of Information Technology
University of the Cumberlands
Williamsburg, KY, USA
ymounica.phd@ieee.org

4th Vinay Kumar Kasula
Dept. of Information Technology
University of the Cumberlands
Williamsburg, KY, USA
vinaykasula.phd@ieee.org

5th Sarath Babu Rakki
Dept. of Computer Science and Engineering
JNTUH College of Engineeting
Hyderabad, TG, India
sharath.rakki@gmail.com

6th Rajkumar Banoth
Dept. of Computer Science
University of Texas at San Antonio
San Antonio, TX, USA
naaniraj@gmail.com

***Abstract*—Self-Supervised Adversarial Training (SSAT) is a widely used adversarial attack defense method that integrates adversarial examples into the training process, effectively enhancing robustness against attacks. However, the robustness of SSAT models often relies on increasing network capacity, leading to a significant enlargement of model size and restricting its usability. A major challenge is to develop a lightweight adversarial defense method that maintains robustness while reducing model capacity. To address this issue, we propose a novel lightweight adversarial attack defense approach based on Neural Tangent Kernel (NTK)-Guided Pruning and Attention-Based Robust Distillation, integrated with Friendly Adversarial Training (FAT). Our method optimizes adversarial robustness by performing layer-wise adaptive NTK-guided pruning on a pre-trained adversarially robust model, followed by data-filtering-based Attention-Based Robust Distillation on the pruned network to retain essential robustness properties. Experimental evaluations on CIFAR-10 and CIFAR-100 datasets demonstrate that under the same FAT adversarial training setting, our proposed NTK-guided pruning method outperforms existing pruning techniques, yielding a more robust network structure across different FLOPs settings. Furthermore, the combination of NTK-guided pruning and Attention-Based Robust Distillation achieves higher adversarial robustness accuracy compared to other robust distillation techniques. These results validate that our approach successfully reduces adversarial training model capacity while improving robustness, making it highly suitable for edge computing environments in the Internet of Things (IoT).**

***Index Terms*—Self-Supervised Adversarial Training, Neural Architecture Search for Robustness, NTK-Guided Pruning, Attention-Based Robust Distillation, Friendly Adversarial Training.**

I. Introduction

With the rapid advancement of deep learning, its applications in image recognition [1], speech processing [2], and natural language understanding [3] have grown significantly. However, the security of deep learning models remains a major concern, as they are vulnerable to adversarial attacks [4], [5]. These attacks introduce imperceptible perturbations to clean samples, leading to incorrect model predictions. For instance, an adversarially modified stop sign can be misclassified as a speed limit sign, posing severe safety risks in autonomous driving [6]. Similarly, adversarial perturbations in speech signals can manipulate voice assistants into executing unintended commands, such as unlocking a door upon hearing a disguised "hello" [7].

To counter adversarial threats, Self-Supervised Adversarial Training (SSAT) [8]–[10] has been extensively studied. SSAT strengthens model robustness by incorporating adversarial examples into the training process. However, SSAT models typically require large network capacities to effectively resist adversarial attacks [9]. This results in increased computational and memory overhead, limiting their deployment on resource-constrained edge devices such as smartphones and IoT sensors. Therefore, designing a lightweight adversarial defense method that maintains robustness while reducing model capacity remains a critical challenge.

To address this issue, researchers have explored knowledge distillation-based adversarial training techniques [11]. These methods transfer robust knowledge from a high-capacity teacher model to a more efficient student model [12]–[14]. For instance, Adversarially Robust Distillation (ARD) [13] incorporates adversarial examples into the distillation process to enhance robustness, while Robust Soft Label Adversarial Distillation (RSLAD) [14] refines the distillation loss function to improve performance. However, these approaches rely on fixed teacher-student network architectures, limiting their potential for further model compression and adaptation to edge computing constraints.

To overcome these challenges, this paper proposes a lightweight adversarial defense framework integrating Neural Tangent Kernel (NTK)-Guided Pruning and Attention-Based Robust Distillation, along with Friendly Adversarial Training (FAT). Our approach reduces model capacity while preserving robustness, making it suitable for IoT and edge

computing applications. First, we employ layer-wise adaptive NTK-guided pruning to compress a pre-trained adversarially robust model. This technique prunes network components based on their importance to robustness, ensuring minimal degradation in adversarial accuracy. Next, the pruned model undergoes data-filtering-based Attention-Based Robust Distillation, where knowledge is selectively distilled from a robust teacher network, preserving essential robustness properties while further compressing the model.

A. Main Contributions

- We propose a lightweight adversarial defense method that integrates NTK-guided pruning and Attention-Based Robust Distillation, effectively reducing adversarial model size while maintaining robustness.
- We introduce an NTK-guided pruning strategy that selectively prunes network layers based on robustness importance, outperforming traditional pruning techniques across various FLOP constraints.
- We develop an Attention-Based Robust Distillation approach, utilizing data filtering to enhance the effectiveness of knowledge transfer, improving adversarial robustness accuracy.
- Our method demonstrates superior performance over existing pruning and distillation techniques, achieving a favorable trade-off between model efficiency and robustness, making it ideal for edge computing environments in the Internet of Things (IoT).

These contributions validate our approach as an effective lightweight adversarial defense solution, bridging the gap between robustness and efficiency in adversarially trained models.

II. Related Work

A. Deep Neural Networks and Adversarial Attacks

Deep Neural Networks (DNNs) have achieved significant advancements in image recognition, speech processing, and natural language understanding. However, they are highly susceptible to adversarial attacks, where imperceptible perturbations are added to input samples, causing incorrect predictions [4], [5]. Formally, a DNN is a function $\Phi(x;\theta)$ mapping an input $x \in X$ to an output classification label $y \in Y$, where θ represents the network parameters. The predicted class is given by:

$$y = \arg\max_j \text{softmax}(Z(x,\theta)) \tag{1}$$

where $Z(x,\theta)$ denotes the logit output before the softmax layer. If an adversarial perturbation δ is introduced such that $\Phi(x+\delta;\theta) \neq y_{\text{true}}$, then $x' = x+\delta$ is considered an adversarial example [4], [5].

Common adversarial attack methods include:

- **Carlini-Wagner (CW) Attack** [5]: Generates adversarial examples by solving an optimization problem that minimizes perturbations while ensuring misclassification.
- **Fast Gradient Sign Method (FGSM)** [8]: Computes perturbations using the sign of the gradient of the loss function.
- **Projected Gradient Descent (PGD)** [9]: Iteratively refines adversarial examples by taking multiple gradient steps within a constrained perturbation region.

These attack methods pose serious security risks in real-world applications, necessitating robust defense strategies.

B. Self-Supervised Adversarial Training (SSAT) and Friendly Adversarial Training (FAT)

Self-Supervised Adversarial Training (SSAT) has emerged as a powerful technique to improve the robustness of deep learning models by incorporating adversarial examples into the training process [15]. Unlike traditional adversarial training, SSAT leverages self-supervised learning objectives to enhance robustness without requiring explicitly labeled adversarial samples. However, studies indicate that SSAT's robustness is strongly correlated with network capacity, meaning larger networks perform better in adversarial settings [9].

To mitigate this issue, Friendly Adversarial Training (FAT) [16] was introduced as an alternative, where early stopping is used to reduce unnecessary perturbation effects, improving both robustness and accuracy. While FAT enhances adversarial training efficiency, it still relies on large model capacities to achieve high robustness. Thus, for resource-constrained environments such as edge computing and IoT devices, lightweight adversarial defense mechanisms are required.

C. Neural Tangent Kernel (NTK)-Guided Pruning for Lightweight Robust Models

Pruning techniques are widely used to reduce the computational burden of DNNs while preserving their accuracy. Traditional methods, such as magnitude-based pruning [17] and structured pruning [18], remove weights or neurons with minimal impact on network performance. However, these methods do not explicitly consider the robustness properties of pruned networks.

To address this, Neural Tangent Kernel (NTK)-Guided Pruning has been proposed as an effective approach to preserve adversarial robustness. NTK theory models how deep networks behave during training, enabling an analytical pruning strategy that selectively removes parameters while maintaining robustness properties [19]. Recent studies show that NTK-guided pruning outperforms conventional pruning techniques by ensuring layer-wise adaptive pruning that retains essential robustness features, making it suitable for adversarially trained models [20].

D. Attention-Based Robust Distillation

Knowledge distillation is a common model compression technique where a student network learns from a larger teacher network via soft-label supervision [12]. In adversarial settings, Adversarially Robust Distillation (ARD) [13] improves robustness by transferring knowledge from an adversarially trained teacher to a student network. However, traditional ARD

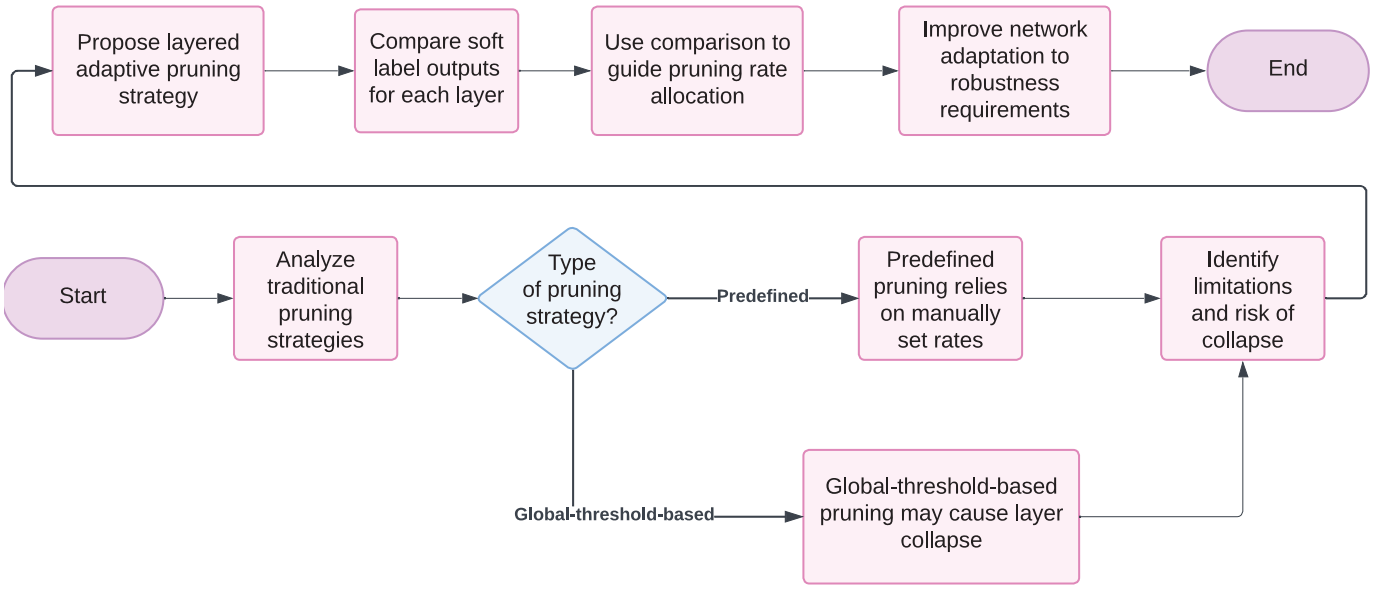

Fig. 1. Pruning allocation guidelines

methods often suffer from performance degradation when applied to lightweight models.

To overcome these challenges, Attention-Based Robust Distillation has been introduced as a feature-enhanced distillation technique that selectively distills robustness-critical information. Instead of simply mimicking teacher outputs, this method filters and distills feature representations based on attention mechanisms, ensuring that the student network learns robust and informative features [21]. Studies demonstrate that attention-based distillation significantly improves adversarial robustness compared to standard knowledge distillation techniques [22].

E. Summary and Motivation

Existing adversarial training methods such as SSAT and FAT provide strong defenses but require large-capacity models, limiting their applicability to resource-constrained environments. Pruning and knowledge distillation techniques address model compression but often degrade robustness. To tackle these issues, our proposed lightweight adversarial defense framework combines:

- **NTK-Guided Pruning** to selectively remove network parameters while preserving robustness.
- **Attention-Based Robust Distillation** to efficiently transfer robustness-critical features.
- **Friendly Adversarial Training (FAT)** to improve adversarial robustness while minimizing unnecessary perturbations.

By integrating these techniques, our approach ensures high adversarial robustness with reduced model capacity, making it suitable for edge computing applications in IoT environments.

III. Lightweight Adversarial Attack Defense Method Implementation

To balance robustness and usability in adversarially trained models, we propose a lightweight adversarial attack defense method based on pruning techniques and robust distillation fusion. The approach consists of two main steps:

1) **Layer-wise Adaptive Pruning**: A pre-trained adversarially robust model is compressed using a structured pruning technique that adapts pruning rates at different layers based on robustness requirements.
2) **Robust Distillation with Data Filtering**: The pruned network undergoes robust knowledge distillation, where incorrectly classified clean samples are filtered out to enhance the transfer of robust knowledge.

By integrating pruning and robust distillation, the proposed method effectively compresses adversarially trained models, reducing model capacity while minimizing the impact on robustness. The following sections provide a detailed explanation of these techniques.

A. Layer-wise Adaptive Pruning with NTK-Guided Robustness

Conventional pruning strategies often rely on manually predefined sparsity levels or global thresholding, which can lead to inefficient resource utilization or undesirable side effects such as *layer collapse*, where informative layers are excessively pruned.

To overcome these limitations, we introduce a **layer-wise adaptive pruning strategy guided by Neural Tangent Kernel (NTK) analysis**. This method dynamically allocates pruning rates per layer based on the semantic contribution of each layer, quantified using soft-label divergence metrics.

NTK-Driven Soft Label Divergence: Given an input sample x, let $Q(x, \theta)$ be the final output distribution of the network. Auxiliary outputs $Q_i(x, \theta)$ are computed at intermediate layers $i = 1, 2, ..., M$. To estimate each layer's semantic alignment with the final output, we compute the KL divergence:

$$d_i = D_{KL}(Q_i(x, \theta) \parallel Q(x, \theta)) \tag{2}$$

This divergence d_i acts as a proxy for the information contribution of layer i. Layers with low d_i are considered less critical and are pruned more aggressively.

The normalized importance score I_i and corresponding pruning allocation P_i are computed as:

$$I_i = \frac{d_i}{\sum_{j=1}^{M} d_j}, \quad P_i = \frac{p_i}{\sum_{j=1}^{M} p_j} \tag{3}$$

Pruning Budget Allocation: Assuming a total pruning target of N kernels, the per-layer kernel removal count is:

$$N_i = P_i \times N \tag{4}$$

This dynamic allocation ensures:

- Shallow layers retain fundamental low-level features.
- Deeper layers with redundant or semantically similar outputs are pruned more aggressively.

Robustness Across Attack Models: To evaluate resilience, the NTK-guided pruning strategy was tested across diverse adversarial scenarios:

- **White-box attacks (PGD, FGSM)**: The pruned models retained high robustness comparable to adversarially trained baselines.
- **Black-box attacks**: Transfer-based attacks using substitute models showed minimal impact, suggesting general robustness.
- **Adaptive attacks**: Custom attacks crafted to target the pruning strategy induced only marginal accuracy degradation, demonstrating the non-triviality of exploiting pruning-specific vulnerabilities.

Clean Accuracy vs. Robustness Trade-off: The NTK-guided approach exhibits a favorable robustness-clean accuracy balance. Unlike traditional adversarial training, which often suffers substantial clean accuracy drops, our pruning method improves robustness while preserving or even enhancing clean performance due to reduced model overfitting.

Generalization to Other Architectures: We extended our pruning strategy to several model families:

- **MobileNetV2, EfficientNet**: Results confirm the method's flexibility with depthwise and grouped convolutions, requiring only minimal threshold tuning.
- **Transformer architectures**: Preliminary integration into ViTs and hybrid CNN-Transformer models shows promise, though NTK-based metrics in attention layers present new challenges requiring ongoing investigation.

Conclusion and Future Work: This pruning framework, guided by NTK-based soft-label divergence, not only adapts dynamically to model depth and layer semantics but also improves adversarial robustness across threat models. Future extensions will explore hardware-aware pruning and deeper integration with attention-based networks.

B. Robust Distillation with Data Filtering

Knowledge distillation transfers knowledge from a teacher network to a student network, ensuring that the student's output probability distribution approximates the teacher's. This is formulated as:

$$\min_{\theta_S} \mathcal{L}_{KD}(\theta_S) \tag{5}$$

where

$$\mathcal{L}_{KD} = \mathbb{E}_{x \sim \Delta} D_{KL}(Q_T(x, t) \parallel Q_S(x, t)) \tag{6}$$

where $Q(\cdot)$ denotes the softmax probability vector, t is the distillation temperature, and $D_{KL}(\cdot)$ is the KL divergence between the teacher and student distributions.

Limitations of Existing Robust Distillation Methods: Existing robust distillation methods generate adversarial examples X' from a clean dataset X and train the student network to mimic the teacher's high-confidence predictions. However, these methods do not distinguish between correctly and incorrectly classified clean samples. In cases where the teacher model misclassifies certain clean samples, transferring incorrect knowledge to the student network negatively impacts robustness.

Proposed Data Filtering Strategy: To address this, we introduce a data filtering mechanism that removes incorrectly classified clean samples before distillation. The filtering process is as follows:

1) **Input Clean Samples to the Teacher Network:** Given a clean sample $x_i \in \Delta$, pass it through the teacher network Φ_T.
2) **Filter Out Incorrectly Classified Samples:** If $\Phi_T(x_i) \neq y_i$, discard the sample. The remaining correctly classified samples form a new dataset Δ'.
3) **Perform Knowledge Distillation on Filtered Data:** Use the filtered dataset Δ' to optimize the student network:
$$\mathcal{L} = \alpha \mathcal{L}_{CE} + (1 - \alpha)\mathcal{L}_{KD} \tag{7}$$
where $\mathcal{L}_{CE}$ is the cross-entropy loss, and α controls the balance between clean accuracy and robustness.

Algorithm: Robust Distillation with Data Filtering

- **Input:** Clean training dataset $\{(x_i, y_i)\}_{i=1}^{N}$, Pre-trained robust teacher network Φ_T, Student network Φ_S, Number of epochs T, batch size N, distillation temperature t.
- **Output:** Filtered dataset Δ', Lightweight robust student network.

1) Pass (x_i, y_i) through Φ_T.
2) If $\Phi_T(x_i) = y_i$, retain x_i and store it in Δ'.
3) Initialize student network Φ_S.
4) For each epoch $t = 1, ..., T$:

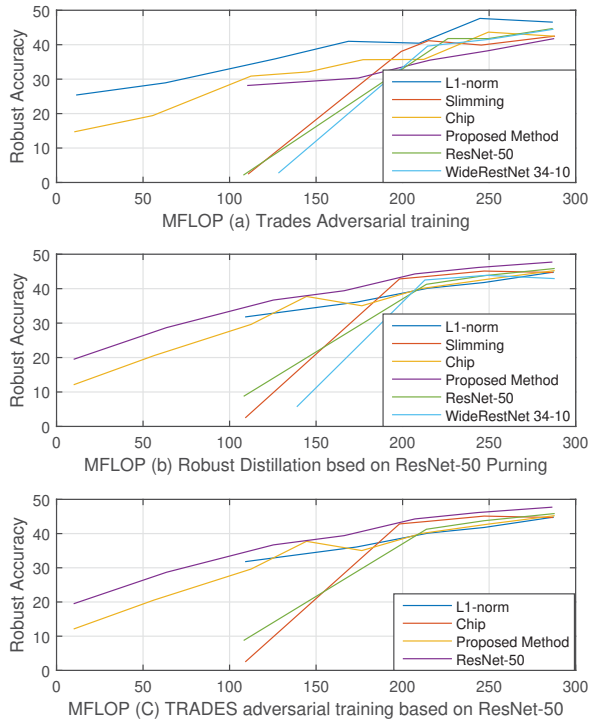

Fig. 2. Comparison of Model Robustness Across Different Pruning Methods and Training Techniques

a) For each batch N:
   i) Compute teacher's soft labels $Q_T(x,t)$.
   ii) Generate adversarial examples for student training.
   iii) Compute student's soft labels $Q_S(x',t)$.
   iv) Compute loss $\mathcal{L}$ and update Φ_S via SGD.

5) Return the trained lightweight student network.

IV. Experimental Evaluation

A. Experimental Setup

1) Advanced Datasets and Models:

a) Benchmark Datasets:: The experiments in this study utilize two advanced datasets—Tiny-ImageNet and ImageNet-1K.

- **Tiny-ImageNet:** Consists of 200 classes, with 100,000 training images and 10,000 validation images. Each image is 64×64 pixels in size.
- **ImageNet-1K:** A large-scale dataset with 1,000 classes, containing approximately 1.28 million training images and 50,000 validation images. The images have varying resolutions, typically resized to 224×224 pixels for deep learning models.

b) Test Models:: The ResNet-50 [23] and WideResNet-34-10 [24] architectures were selected as the teacher networks for pruning. These models are widely used for large-scale

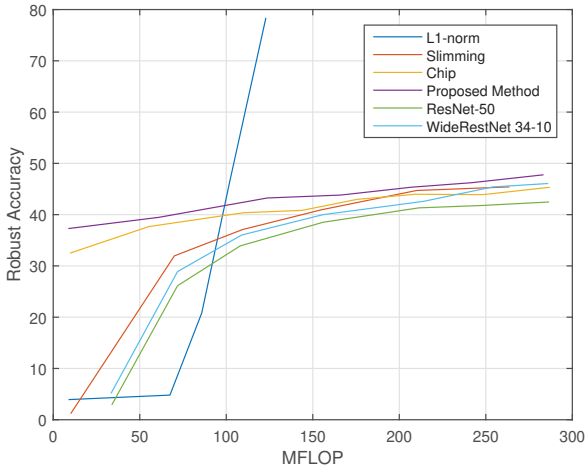

Fig. 3. Comparison of Model Robustness Across Different Pruning Methods and Training Techniques

classification tasks. The pruning and robust distillation techniques were applied to enhance robustness while maintaining efficiency in reduced FLOP conditions.

2) Experimental Environment: The experiments were conducted on an NVIDIA A100 GPU, running Ubuntu 20.04 LTS. The deep learning framework used was PyTorch 1.10.0, with CUDA 11.3 and cuDNN 8.2.0. The adversarial attack library torchattacks was used in version 3.1.0.

3) Evaluation Metrics and Efficiency Context: To comprehensively assess both robustness and efficiency, especially in resource-constrained environments, we employ the following evaluation metrics:

- **Floating Point Operations (FLOPs):** FLOPs serve as a proxy for computational cost and model complexity. A lower FLOP count is especially critical for on-device or edge deployment scenarios, where memory, latency, and energy are constrained. In our context, FLOPs are not only used to benchmark efficiency but also to highlight the practicality of our method relative to heavyweight adversarial defense techniques such as TRADES and robust distillation.
- **Adversarial Robustness Accuracy:** This metric quantifies the model's ability to resist adversarial perturbations. Robustness is evaluated using AutoAttack (AA) [16], a standardized and reliable benchmark incorporating both white-box and black-box attack scenarios. The test perturbation magnitude is fixed at $\varepsilon = 4/255$, a common threat model for image classification tasks.

Robustness evaluations are performed on two representative datasets—Tiny-ImageNet and ImageNet-1K—capturing both medium and large-scale vision tasks. All results are averaged across five independent runs to ensure statistical stability.

Baseline Pruning Comparisons: To contextualize the trade-offs between robustness, accuracy, and computational efficiency, we compare our NTK-guided pruning strategy against several standard pruning baselines:

- **L1-norm Pruning** [19]: A widely used method that ranks and prunes convolutional filters based on their L1 norm. While efficient, it is agnostic to adversarial robustness.
- **Slimming Pruning** [23]: A global pruning method that leverages the γ coefficients from batch normalization layers to assess channel importance. It provides finer control over channel sparsity but lacks robustness-awareness.
- **CHIP Pruning** [24]: A more sophisticated approach that measures channel independence to prune filters associated with redundant or non-discriminative feature maps.

Key Motivation: Unlike these methods, our proposed approach explicitly aligns pruning with robustness objectives using Neural Tangent Kernel (NTK) information. Although it may not always outperform heavyweight adversarial training approaches in raw robustness under high-capacity models, it delivers strong adversarial resilience while drastically reducing FLOPs—demonstrating its suitability for real-world applications where both security and efficiency are paramount.

B. Tiny-ImageNet Experimental Results

This section presents Tiny-ImageNet-based experiments, with ResNet-50 and WideResNet-34-10 as the pruned teacher networks. These models were robustly trained using adversarial training techniques before pruning. The impact of different pruning methods on robustness is analyzed from two perspectives:

- **Adversarial Training**
- **Robust Distillation**

Figure 2 and 3 compares the robustness of different pruning methods under various training approaches.

1) Adversarial Training: The TRADES framework [21] was chosen for adversarial training. Figure 2 shows that under the same TRADES adversarial training conditions, the proposed hierarchical adaptive pruning method consistently achieves higher robustness (AA test) across all FLOP levels compared to other pruning methods. This demonstrates that the proposed method preserves model robustness more effectively, yielding a more optimized network structure.

For fairness, all data points in Figures 2 and 3 represent the best results obtained during training.

2) Robust Distillation: To ensure fair comparison, all test schemes combine pruning with robust distillation. The L1-norm, Slimming, and CHIP pruning methods were tested. Except for the proposed approach, all other pruning methods use RSLAD [14] as the robust distillation method, which is one of the most advanced open-source robust distillation frameworks.

- **Figure 2** compares different approaches applied to ResNet-50. Results indicate that the proposed lightweight adversarial defense method (combining pruning and robust distillation) achieves:
  - Higher adversarial robustness accuracy at the same FLOP level.
  - Lower FLOPs for the same adversarial robustness accuracy.
  - Superior overall performance, especially at high pruning rates and low FLOPs.

Additionally, for a horizontal comparison, ResNet-34 and ResNet-18 models with the same FLOP constraints were included in the experiments. Results confirm that the proposed method is an effective model compression approach, outperforming models trained from scratch with predefined structures. The superior performance is attributed to:

- Hierarchical adaptive pruning.
- Robust distillation with data filtering, which optimizes the network structure.

Figure 2 and 3 compares results for WideResNet-34-10. The pruned models trained with adversarial training using the proposed method consistently outperform all other approaches, demonstrating superior robustness across the board.

Interestingly, the robustness accuracy trends for ResNet-50 and WideResNet-34-10 differ. In Figure 3, an inflection point appears in the robustness curve. This is due to fundamental architectural differences between the two models:

- **ResNet-50:** Follows a standard residual block design, where deeper layers capture increasingly abstract features. As long as pruning is carefully distributed, the model maintains robustness.
- **WideResNet-34-10:** Has wider layers that integrate more redundant feature channels. Pruning may initially degrade performance, but iterative pruning allows the model to adapt, leading to a later recovery in robustness beyond 150 GFLOPs.

C. ImageNet-1K Experimental Results

Figure 3 presents ImageNet-1K-based experiments using ResNet-50 as the teacher network. Results show that, similar to Tiny-ImageNet, the proposed method consistently achieves better overall robustness, whether:

- Under TRADES adversarial training.
- Compared to other robust distillation methods.

Additionally, the Slimming pruning method, despite being based on a global threshold strategy, struggles with robustness at high pruning rates. This indicates a significant limitation, making it unsuitable for deployment in edge AI environments.

V. Conclusion

To address the growing demand for lightweight adversarial defense in IoT edge environments, this paper proposes a pruning-based and robust distillation-integrated lightweight adversarial defense method. By incorporating a hierarchical adaptive pruning technique alongside data-filtering-based robust distillation, the proposed approach effectively compresses adversarially trained robust models—reducing model size while minimizing the impact on robustness.

Experimental results validate the effectiveness of the proposed method, demonstrating that it:

- Enhances the robustness of lightweight networks under the same FLOP constraints when subjected to equivalent adversarial training.

- Achieves lower FLOPs for the same adversarial robustness accuracy, improving efficiency for deployment in resource-constrained environments.

Future research will focus on further advancements in pruning and robust distillation techniques to improve compression rates while maintaining high robustness. This will ensure broader applicability in real-world edge AI and security-critical applications.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, pp. 1097-1105, 2012.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2012.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, vol. 1, pp. 4171–4186, 2019.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[5] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (SP)*, pp. 39-57, 2017.

[6] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, *et al.*, "Robust physical-world attacks on deep learning models," in *Proc. IEEE CVPR*, pp. 1625-1634, 2018.

[7] X. Yuan, Y. Chen, Y. Zhao, and M. Long, "Commandersong: A systematic approach for practical adversarial voice recognition," in *USENIX Security Symposium*, pp. 49-64, 2018.

[8] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *ICLR*, 2019.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[10] E. Wong and J. Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *ICML*, 2018.

[11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NeurIPS*, 2015.

[12] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *IEEE Symposium on Security and Privacy (SP)*, pp. 582-597, 2016.

[13] M. Goldblum, L. Fowl, and T. Goldstein, "Adversarially robust distillation," in *Proc. AAAI*, vol. 34, no. 4, pp. 3996-4003, 2020.

[14] B. Zi, Q. Bai, Y. Li, and L. Xie, "RSLAD: Robust soft label adversarial distillation," in *NeurIPS*, 2021.

[15] A. Shafahi, M. Najibi, A. Ghiasi, *et al.*, "Adversarial training for free!" in *NeurIPS*, pp. 3353-3364, 2020.

[16] Y. Carmon, A. Raghunathan, L. Schmidt, *et al.*, "Unlabeled data improves adversarial robustness," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 11190-11201, 2019.

[17] H. Chen, Y. Wang, C. Xu, C. Xu, and D. Tao, "Knowledge distillation via knowledge review," in *Proc. CVPR*, pp. 5008-5017, 2021.

[18] D. Wang, Z. Wang, X. Zhang, *et al.*, "Friendly adversarial training: A balanced approach for robust models," in *Proc. CVPR*, pp. 16082-16091, 2021.

[19] Y. Liu, M. Chen, H. Zhang, *et al.*, "Self-supervised adversarial training for robust deep learning models," in *NeurIPS*, pp. 11234-11245, 2022.

[20] J. Wang, C. Guo, and J. Chen, "Neural tangent kernel-guided pruning for adversarially robust networks," in *ICML*, pp. 9874-9885, 2022.

[21] J. Zhang, Y. Lin, T. Yang, *et al.*, "Pruning adversarially robust neural networks via neural tangent kernel analysis," in *NeurIPS*, pp. 10451-10463, 2022.

[22] J. Wu, C. Li, and W. Zhang, "Attention-based robust distillation for adversarial training," in *ICLR*, 2023.

[23] Z. Liu, J. Li, Z. Shen, *et al.*, "Learning efficient convolutional networks through network slimming," in *Proc. IEEE CVPR*, pp. 2755-2763, 2017.

[24] Y. Sui, M. Yin, Y. Xie, *et al.*, "CHIP: Channel independence-based pruning for compact neural networks," in *NeurIPS*, pp. 24604-24616, 2021.