UNIVERSITY OF CALIFORNIA

Los Angeles

Parameter-free Adversarial Attack via Learned Optimizer

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Computer Science

by

Lalit Bhagat

2023

ABSTRACT OF THE THESIS

Parameter-free Adversarial Attack via Learned Optimizer

by

Lalit Bhagat
Master of Science in Computer Science
University of California, Los Angeles, 2023
Professor Cho-Jui Hsieh, Chair

As the domain of adversarial attack countermeasures continues to expand, the accurate evaluation of these defenses remains a challenge. Adversarial attacks pose significant challenges to the security and robustness of deep learning models. Traditional methods typically depend on predetermined parameters, such as ensembles of certain methods and manually designed rules, which may not be optimal for generating effective attacks. In this research, we propose a parameter-free adversarial attack by leveraging a learning-to-learn (L2L) framework.

We train a recurrent neural network-based optimizer to adaptively update directions and steps, enabling more efficient and adaptive adversarial attacks. We conduct extensive experiments on robust models trained on the MNIST and CIFAR-10 datasets.

Our findings show that the learned optimizer outperforms traditional methods, such as PGD, in generating adversarial attacks for small networks and smaller datasets like MNIST. For larger networks, our method demonstrates improved performance only for smaller attack steps. These results highlight the potential of parameter-free attacks in evaluating and understanding the robustness of deep learning models.

The thesis of Lalit Bhagat is approved.

Aditya Grover

Quanquan Gu

Nanyun Peng

Cho-Jui Hsieh, Committee Chair

University of California, Los Angeles

2023

To my family, friends, and behenji,
whose unwavering support and encouragement
have shaped me into the person I am today.

Table of Contents

List of Tables

Acknowledgments

First and foremost, I would like to express my deepest gratitude to Professor Cho-Jui Hsieh for his unwavering mentorship, support, patience, and encouragement throughout the course of this research. This thesis would undoubtedly not have been possible without his guidance.

I would also like to extend my sincere appreciation to my Master's committee members, Professors Aditya Grover, Quanquan Gu, and Nanyun Peng. Their dedication in reviewing this thesis has played an essential role in refining and enhancing the quality of my work.

CHAPTER 1

Introduction

Deep learning has revolutionized various domains, including computer vision, natural language processing, and speech recognition, due to its ability to learn from vast amounts of data and achieve state-of-the-art performance in many tasks. However, deep learning models have been found to be vulnerable to adversarial attacks, where small perturbations are added to the input data, leading to misclassifications by the model. These attacks pose significant challenges to the security and robustness of deep learning systems, especially in critical real-world applications such as object recognition in self-driving cars.

Adversarial examples and adversarial training were first introduced by Ian Goodfellow [8]. Various algorithms, such as DeepFool [16], FGSM [8], I-FGSM [12] and CW [2], have made it easy to generate adversarial examples. Furthermore, adversarial attacks have been demonstrated to be effective in black-box settings [12], where the attacker has limited knowledge of the target model, and have also been validated through real-world applications [17]. This has prompted the development of numerous defense mechanisms to safeguard deep learning models against adversarial attacks, among which adversarial training remains the most effective.

The Projected Gradient Descent (PGD) attack [14] is a popular method for testing adversarial robustness due to its computational efficiency and effectiveness in many circumstances. However, even PGD has been found to fail [6], resulting in a considerable overestimation of robustness. One of the reasons for potential failure is the estimation of fixed step size. Therefore, automating the step size for adversarial attacks has become an important topic in the field of adversarial robustness.

To address the limitations of the traditional PGD attack, researchers have developed the Auto-PGD, an adaptive version of PGD. Auto-PGD adaptively tunes its parameters, such as the step size and number of iterations, allowing it to generate stronger adversarial examples and provide a more accurate estimate of a model's robustness. Despite its advantages, Auto-PGD has some limitations. For instance, the adaptive nature of Auto-PGD might increase the computation complexity, as it requires tuning parameters during the attack. This can lead to higher computational costs compared to fixed-parameter methods like PGD. Furthermore, although Auto-PGD adapts its parameters to perform better against a particular defense, it might not be universally optimal for all defense mechanisms, requiring further adaptation or the combination of different attack methods.

Considering the limitations of both PGD and Auto-PGD, the adaptive auto attack, AutoAttack, has become the standard for assessing adversarial defenses. As described by Croce and Hein [6], AutoAttack offers a reliable and effective way to evaluate adversarial defenses by combining an ensemble of white-box adversarial attacks. This ensemble includes different attack methods, such as decision-based, gradient-based, and score-based attacks, providing a comprehensive assessment of a given defense's robustness. By identifying weaknesses in defenses, the ensemble approach contributes to the development of more robust and secure deep learning models.

However, AutoAttack also has some limitations. One notable limitation is its computation complexity, as combining multiple attack methods increases the computational complexity and runtime compared to using a single attack method. Additionally, AutoAttack relies on human-designed rules and ensembles, which might not fully capture the adaptability and flexibility needed to address the wide range of adversarial defense mechanisms. This might make it less suitable for real-time applications or large-scale evaluations. Therefore, it is important to design a single adaptive attack that does not use hyperparameters and human-designed rules and ensembles. In the following sections of this thesis, we will explore such an adaptive attack method and evaluate its effectiveness in comparison to traditional attack methods.

1.1 Contributions

In this work, we advance the field of adversarial robustness in deep learning through a set of pivotal contributions. First, we introduce a parameter-free attack package designed to assess the robustness of deep learning models. Unlike existing auto-attack packages, which are reliant on human-designed rules and heuristics [6], our approach uses a learned optimizer. Our approach introduces a Recurrent Neural Network (RNN) based Learn-to-Learn (L2L) adversarial attack. This approach utilizes the capability of RNNs to capture long-term dependencies, as corroborated by prior work [13], and employs them as the optimizer for generating coordinate-wise perturbations. This aligns closely with the framework we propose, thus making it a significant departure from traditional adversarial attack methods.

Finally, we delve into the specifics of optimizer training and present novel insights. Through experimental evaluations, we ascertain that training the optimizer for fewer steps can generate perturbations that are more effective than those produced by PGD. However, the advantage is mitigated and can even turn into a disadvantage when a larger step size is employed for training. This observation serves as an essential guideline for future work regarding step size configuration in optimizer training.

In summary, the contributions of the thesis are as follows:

- Developed a parameter-free Learn-to-Learn (L2L) based adversarial attack, utilizing Recurrent Neural Networks (RNNs) as the optimizer
- Demonstrated that the parameter-free L2L-based adversarial attack outperforms traditional methods like Projected Gradient Descent (PGD) when smaller step sizes are used for training.
- Conducted rigorous evaluations to assess the limitations of existing L2L frameworks, offering valuable guidelines concerning the impact of step size during optimizer training.

1.2 Overview

The remainder of this thesis is organized as follows:

Chapter 2 reviews the related work on adversarial attacks and defense mechanisms, setting the stage for our research in the broader context of adversarial machine learning.

Chapter 3 details our methodology, introducing our RNN-based Learn-to-Learn (L2L) approach for adversarial attacks and explaining the unique aspects of our model.

In Chapter 4, we present experiments and results, showcasing the effectiveness of our approach compared to traditional adversarial methods and state-of-the-art techniques.

Finally, Chapter 5 concludes the thesis, summarizing our key findings and contributions, and outlines directions for future research in this area.

CHAPTER 2

Related Work

2.1 Adversarial Attack and Defense

Adversarial attacks have become a significant area of research within the field of machine learning, particularly concerning the security and robustness of deep learning models. Szegedy et al. [19] were among the first to reveal the susceptibility of neural networks to adversarial examples—inputs crafted with imperceptible perturbations designed to mislead models. This seminal work laid the groundwork for a multitude of subsequent studies aiming to understand and mitigate such vulnerabilities. Goodfellow [8] introduced the concept of adversarial training as a countermeasure, which involves augmenting the training process with adversarial examples. They also presented the Fast Gradient Sign Method (FGSM), a technique for generating adversarial examples rapidly, underscoring the linear nature of deep networks as a potential cause for their fragility.

The quest for more sophisticated attacks led to the development of iterative methods, such as Projected Gradient Descent (PGD) [14], which are more effective at finding adversarial examples than one-step methods like FGSM [12]. [12] demonstrated that adversarial examples could be crafted even in the physical world, not just in digital space, raising concerns about the practical implications of such attacks. Meanwhile, the Carlini & Wagner Attacks [2] offered a set of methods that provided a benchmark for the robustness of defensive strategies, indicating that defensive distillation [18] could be overcome.

In parallel, the pursuit of robust optimization techniques led to the formulation of adversarially robust models. [14] posited that robustness could be viewed as a property emerging from solving a min-max optimization problem, where the inner maximization is responsible

for finding the most adversarial perturbation, while the outer minimization improves the model's parameters against such perturbations. This perspective has fueled research into the interplay between attack generation and model robustness.

In summary, the body of work on adversarial attacks reflects a dynamic tension between the development of attack methodologies and the quest for model robustness. The evolving nature of adversarial tactics—from simple gradient-based approaches to complex learned optimizers—highlights the importance of ongoing research in this field. It underscores the need for a continued focus on devising robust training mechanisms and exploring the theoretical underpinnings of adversarial resilience.

2.2 Learning to Learn

The Learning to Learn (L2L) framework, also known as meta-learning, is an emerging paradigm in machine learning that focuses on designing algorithms capable of learning how to learn. The central idea is to develop models that can generalize learning strategies across various tasks, thereby improving their performance on unseen challenges. One of the seminal works in this area by [1] introduced the concept of training neural networks with a meta-level knowledge of what constitutes an effective learning process.

In the current study, our focus is on a specialized niche within the Learning to Learn (L2L) domain, specifically the learning of optimizers that enhance performance. Instead of relying on predefined update rules crafted by human experts, the L2L framework advocates the use of neural networks to autonomously devise optimization strategies. The roots of this approach can be traced back to the pioneering work of [4] and [10], where the initial models of adaptive algorithms were applied to straightforward convex challenges. Building on this foundation, [1] introduced an LSTM-based optimizer capable of tackling more intricate optimization tasks, including the training of convolutional neural network classifiers.

In the context of adversarial machine learning, L2L approaches have been explored to design models that can learn robust optimization strategies. The work of [**?**] stands out

in this regard, utilizing a CNN generator to transform clean images into their perturbed counterparts. However, as this approach mirrors the one-step, deterministic nature of attacks like FGSM, it has limitations in terms of strength and diversity. Addressing this, [9] refined the algorithm to iteratively generate more potent and varied attacks. A significant challenge with the CNN generator, however, is its extensive parameterization and its inability to effectively account for long-term dependencies, which complicates the optimization process, particularly in the context of the minimax challenges posed by adversarial training. This led to [22] introducing an RNN optimizer-based method, which not only facilitates a more stable training regimen but also offers a more nuanced understanding of the update mechanism. Inspired by these developments, our research adopts a similar RNN-based optimizer but with a specific focus on enhancing the efficacy of adversarial attacks, tailoring it to be more adept at generating sophisticated adversarial examples.

CHAPTER 3

Methodology

3.1 Preliminaries

3.1.1 Notations

In our work, clean images and their labels are denoted by the bold lowercase letters $\mathbf{x}$ and $\mathbf{y}$, respectively. The focus of this paper is on a task related to image classification, where the classifier f is parameterized using θ. The operation $\text{sign}(\cdot)$ is applied element-wise to identify the sign of an input, with the convention that $\text{sign}(0) = 1$. The vicinity of $\mathbf{x}$ and the set of allowable perturbed images are represented by $B(\mathbf{x}, \epsilon)$, defined as $\{\mathbf{x}' : \|\mathbf{x}' - \mathbf{x}\|_\infty \leq \epsilon\}$, here employing the infinity norm as the metric for distance. The projection operator, denoted as Π, is responsible for mapping the perturbed data back into the allowable set. Specifically, $\Pi_{B(\mathbf{x},\epsilon)}(\mathbf{x}')$ is calculated as $\max(\mathbf{x} - \epsilon, \min(\mathbf{x}', \mathbf{x} + \epsilon))$, functioning on an element-wise basis. Finally, $L(\cdot, \cdot)$ symbolizes a loss function used in multi-class scenarios, such as cross-entropy.

3.1.2 Adversarial Attack

In this section, we delve into the formulation of adversarial attacks, adapting certain hand-designed optimizers for this purpose. The goal of constructing a robust adversarial attack is to maximize the loss within a bounded neighborhood $B(\mathbf{x}, \epsilon)$, effectively challenging the classifier's resilience. This process can be conceptualized as a maximization problem, contrasting the minimization approach in adversarial training. The objective is to identify perturbations that lead to the maximal loss, thereby effectively exploiting the weaknesses of the classifier. This concept is formalized as follows:

$$\max_{\mathbf{x}' \in B(\mathbf{x},\epsilon)} L(f(\mathbf{x}'), \mathbf{y}) \tag{3.1}$$

Here, L denotes the loss function, f represents the classifier parameterized by θ, and D is the empirical distribution of the input data. This maximization problem seeks to discover the most effective adversarial examples within the permissible perturbation set $B(\mathbf{x}, \epsilon)$, thereby rigorously testing the robustness of the classifier against adversarial threats.

The effectiveness and efficiency of the inner maximization process are vital for developing an improved adversarial attack. A widely utilized method for this inner maximization is the projected gradient descent (PGD) algorithm. PGD performs a predetermined number of update steps as follows:

$$\mathbf{x}'_{t+1} = \Pi_{B(\mathbf{x},\epsilon)}(\alpha \text{sign}(\nabla_{\mathbf{x}'} L(\mathbf{x}'_t)) + \mathbf{x}'_t) \tag{3.2}$$

In this equation, $L(\mathbf{x}'_t)$ signifies the term we aim to maximize.

3.1.3 Adaptive Step Sizes

The efficacy of adversarial attacks is significantly influenced by the choice of the optimization algorithm used for inner maximization. As suggested by [6, 22], the prevalent Projected Gradient Descent (PGD) method might not be the most effective strategy. Minor modifications to PGD, without altering the core objectives of the adversarial approach, can significantly enhance the attack's impact. Utilizing the CNN architecture outlined in [24], experiments on the MNIST dataset show that a constant step size in a typical 10-step PGD setup may not always be optimal for the attack's progression. Integrating backtracking line search (BLS) dynamically determines the step size for the adversarial attack. Starting with maximum step size α_0 and reducing it iteratively as $\alpha_t = \rho \alpha_{t-1}$ until the condition $L(\mathbf{x}' + \alpha_t \mathbf{p}) \geq L(\mathbf{x}') + c\alpha_t \mathbf{T}\mathbf{p}$ is met, where $\mathbf{p} = \nabla_{\mathbf{x}'} L(\mathbf{x}')$, ensures that each update contributes to a meaningful increase in the loss function. This method stops after a finite number of steps once a sufficiently small α is achieved, following the standard in gradient

ascent optimization (refer to discussions in [22]). With $\rho = 0.5$ and $c = 10^{-4}$, the results indicate that the combination of adversarial attack with BLS produces stronger adversarial examples and diminishes robust accuracy more effectively than the standard PGD approach.

3.1.4 Learning to learn by gradient descent by gradient descent

The Learning to Learn by Gradient Descent by Gradient Descent (L2L-GDGD) [1] approach represents a novel paradigm in neural network optimization. This method involves a neural network, referred to as the 'learner', which is trained to optimize the parameters of another network. The learner network dynamically determines the best strategy for optimization, and this process is mathematically expressed as:

$$\theta_{t+1} = \theta_t + g_t(\nabla f(\theta_t), \phi). \tag{3.3}$$

Here eq. (3.3), θ denotes the parameters of the network being optimized, g_t is the update function determined by the learner, and ϕ represents the parameters of the learner network. The learner's objective is to adaptively modify the optimization path to effectively minimize the loss function of the target network.

The essence of L2L-GDGD lies in its ability to self-improve the optimization process. Rather than relying on predefined rules or fixed algorithms, the learner network identifies and implements the most efficient strategies for optimizing the target network. This approach is transformative as it allows for the development of optimization strategies that are not constrained by traditional limitations, offering a more flexible and potentially more effective approach to neural network training.

3.2 Learning to Learn for Adversarial Attack

In the advancement of adversarial attack methodologies, the role of the maximization process has emerged as a pivotal factor. Prior discussions, particularly in Section 3.1.3, have

acknowledged the efficacy of Backtracking Line Search (BLS) in this context. Nevertheless, the practical application of BLS within adversarial attack frameworks encounters significant computational challenges. These arise primarily from the algorithm's inherent need for repetitive line searches and continuous loss assessments, leading to an escalated computational load. This situation raises a fundamental question in the optimization of adversarial attacks: Is there an approach to autonomously determine an optimal step size for inner maximization that circumvents excessive computational demands? Extending this line of inquiry, one might also consider the potential for crafting a bespoke maximizer, tailored to specific datasets and models, as a viable alternative to general-purpose optimizers like PGD.

In recent years, the field of learning-to-learn has been exploring the potential of machine learning, particularly neural networks, to develop improved optimizers that could supersede traditional hand-designed ones. Despite the promising nature of this research [1, 20, 13], the real-world application of machine learning-driven optimizers faces notable challenges. A primary issue, identified as gradient explosion [15], occurs in extended optimization sequences and hampers the ability of these optimizers to adapt to longer optimization steps. One major issue is the phenomenon of gradient explosion [15] in unrolled optimization, which hinders the generalization of these ML-learned optimizers to scenarios with longer steps. Additionally, truncated optimization often leads to short-horizon bias [21], further complicating their practical application.

In our research, we demonstrate the feasibility and practicality of learning an optimizer specifically for the inner maximization process in adversarial attack strategies. It is important to note that in the context of these strategies, the maximization problems typically exhibit a similar structure: $\max_{\mathbf{x}' \in B(\mathbf{x}, \epsilon)} L(f(\mathbf{x}'), \mathbf{y})$. Here, the loss function L and the network f (both in terms of structure and weights) remain consistent across different instances, with the primary variables being the input $\mathbf{x}$ and the label $\mathbf{y}$. This uniformity presents an opportunity to develop a specialized maximizer, one that is fine-tuned to perform efficiently on a predefined set of optimization challenges inherent to adversarial attacks. Moreover, the requirement for the maximizer is not to generalize across a wide range of problems but

to excel in a fixed set of scenarios specific to the adversarial framework. This distinction allows for the creation of an optimizer that can potentially surpass the performance of more generalized methods like PGD.

To develop a more effective adversarial attack, we have chosen to parameterize our optimizer using a Recurrent Neural Network (RNN), inspired by the learning-to-learn framework [1]. Our approach diverges from conventional methods through the incorporation of specialized design modifications tailored for the specific challenges of adversarial optimization. These enhancements enable the RNN to effectively navigate the complexities inherent in crafting potent adversarial examples. Optimizing the parameters of the RNN, represented by ϕ, is crucial in this process, ensuring that our strategy is finely tuned for the nuanced demands of effective adversarial attacks.

In our work, we have integrated an RNN-based optimizer, denoted as m and parameterized by ϕ, to enhance the adversarial attack methodology. This optimizer is designed to follow a parameterized update rule that aligns with the PGD update rule:

$$\delta_t, h_{t+1} = m_\phi(g_t, h_t), \quad \mathbf{x}'_{t+1} = \Pi_{B(\mathbf{x},\epsilon)}(\mathbf{x}'_t + \delta_t). \tag{3.4}$$

In this formula, g_t represents the gradient $\nabla_{\mathbf{x}'} L(f(\mathbf{x}'), \mathbf{y})$, and h_t is the hidden state of the RNN. A distinctive aspect of our RNN optimizer is its ability to generate perturbations in a coordinate-wise manner. This approach differs from other learning-to-learn (L2L) methods, which typically process the entire image as input. One of the key benefits of our method is the significant reduction in the number of trainable parameters, resulting in a more efficient and expedited training process.

Furthermore, the hidden state of our RNN optimizer plays a critical role in the optimization process. By maintaining a separate hidden state for each coordinate, our model ensures varied update behaviors across different aspects of the input. This structure not only captures rich information, such as the trajectory of loss gradients as mentioned in [9], but also allows for a recursive update mechanism that is simpler in structure yet effective in adapting to the adversarial context.

In adapting the RNN architecture for our purposes, we have taken inspiration from the model outlined in [1] but introduced several key adjustments inspired from [22] to better align it with the requirements of adversarial attack optimization. The perturbation calculation at each step of our modified RNN can be described as:

$$\delta_t = \tanh(Vh_t + b_1), \tag{3.5}$$

$$h_{t+1} = \tanh(Ug_t + Wh_t + b_2) \tag{3.6}$$

Here, h_t is a vector in R^d, with V, U, and W being matrices of dimensions $R^{1\times d}$, $R^{d\times 1}$, and $R^{d\times d}$ respectively, and b_1 and b_2 as biases in appropriate dimensions, facilitating the coordinate-wise update. During the optimization process, as the gradients diminish in magnitude upon nearing local maxima, the perturbation values should stabilize, exhibiting minimal fluctuation between iterations. However, as indicated in equations (3.5) and (3.6), even with minimal gradients g_t, the update mechanism can still generate notable changes due to the influence of $\tanh(b_1)$. In scenarios where an optimal solution is achieved with a zero-initialized hidden state (thus necessitating b_2 to also be zero), the presence of a non-zero $\delta_t = \tanh(b_1)$ can inadvertently shift the adversarial example away from its optimum. Consequently, the inclusion of biases b_1 and b_2 may hinder optimization near-optimal solutions. Given the typically brief horizon of the optimization process in adversarial contexts, it is improbable for the network to autonomously learn to nullify these biases. Therefore, to promote stability and efficacy in training, we have opted to omit these bias terms from the standard RNN structure in all our implementations.

In our framework, we train the RNN optimizer parameters ϕ. The optimal parameters are determined by:

$$\phi^* = \arg\max L(\phi) \tag{3.7}$$

where $\mathbf{x}'_T(\phi^*)$ is computed by iteratively running Eq. (3.4) for T times. Since our aim

is to find better adversarial examples, we define the objective function for training over the horizon T as:

$$L(\phi) = \sum_{t=1}^{T} w_t L(f_\theta(\mathbf{x}'_t(\phi)), \mathbf{y}). \tag{3.8}$$

Setting $w_t = 0$ for all $t < T$ and $w_T = 1$ implies that our learned maximizer m_ϕ will maximize the loss after T iterations as per equation (10). It should be noted that [15] highlighted potential issues with this kind of unrolled optimization, such as the risk of gradient explosion, which remains an unresolved challenge in learning-to-learn (L2L). To mitigate this, we perform unrolling after every 10 iterations of T. For example, unrolling is conducted at $T = 10, 20, 30, \ldots$ for the number of attack steps. We discovered that without this unrolling step, the learned optimizer fails to generate more effective attacks. Algorithm 3.2 contains the complete algorithm.

Algorithm 1 RNN-based adversarial attack

Input: clean data $\{(x, y)\}$, batch size B, step size α, number of inner iterations T, classifier parameterized by θ, RNN optimizer parameterized by ϕ
Output: Learned optimizer m_ϕ
Randomly initialize m_ϕ and initialize f_θ with pre-trained configurations
repeat
    Sample a mini-batch M from clean data.
    for (x, y) in B **do**
        Initialization: $h_0 \leftarrow 0$, $L_\phi \leftarrow 0$
        Gaussian augmentation: $x'_0 \leftarrow x + 0.001 \cdot \mathcal{N}(0, I)$
        for $t = 0, \ldots, T-1$ **do**
            $g_t \leftarrow \nabla_{x'_0} L(f_\theta(x'_t), y)$
            $\delta_t, h_{t+1} \leftarrow m_\phi(g_t, h_t)$, where coordinate-wise update is applied
            $x'_{t+1} \leftarrow \Pi_{B(x,\epsilon)}(x'_t + \delta_t)$
            $L_\phi \leftarrow L_\phi + w_{t+1} L(f_\theta(x'_{t+1}), y)$, where $w_{t+1} = t + 1$
        end for
    end for
    Update ϕ by $\phi \leftarrow \phi + \alpha \nabla_\phi L_\phi / B$
until training converged

CHAPTER 4

Experiments and Results

In this section, we meticulously detail the experimental results obtained from the deployment of our proposed RNN-based adversarial attack framework. Our investigation encompasses a thorough comparison of our method against a series of established baselines, specifically within the context of white-box adversarial attacks. To ensure the robustness and generalizability of our results, we extend our experiments across a variety of datasets and delve into the performance implications of employing different network architectures.

The experimental analysis is methodically organized to facilitate a comprehensive understanding of the attack efficacy and system behavior. Section 4.1.1 delves into the Datasets and Classifier Networks, providing an overview of the experimental canvas upon which our adversarial strategies are tested. Following this, Section 4.1.2 introduces the Baselines for Comparison, where our RNN-based method is juxtaposed with existing adversarial attack strategies to underscore its relative performance and highlight its strengths. Section 4.1.3 then elucidates the Evaluation and Implementation Details, offering transparency into the metrics used for performance assessment and the specifics of the implementation that might influence the reproducibility and comparability of the results.

With the experimental groundwork laid out, Section 4.2 presents the Results of our experiments, where the empirical data is meticulously examined and interpreted. This segment is pivotal, as it not only demonstrates the practical potency of our proposed attack but also places it within the spectrum of current adversarial attack methodologies. Lastly, Section 4.3 offers an Analysis of the findings, delving deeper into the implications, the potential for scalability, and the adaptability of our adversarial attack approach. Here, we dissect the ele-

ments of success and the areas for improvement, setting the stage for subsequent discussions and future research directions.

Through this structured experimental discourse, we aim to substantiate the claims of our thesis with quantifiable evidence and critical insights, hence contributing to the body of knowledge with findings that are both verifiable and instrumental for the advancement of adversarial attack techniques in deep learning.

4.1 Experimental settings

4.1.1 Datasets and classifier networks

In the experimental setup of our study, we meticulously selected the datasets and corresponding classifier networks to evaluate the performance of our proposed adversarial attack model. The experiments are primarily conducted using two benchmark datasets well-established in the domain of machine learning: MNIST [7] and CIFAR-10 [11].

For the MNIST dataset, which comprises grayscale images of handwritten digits, we adopt a Convolutional Neural Network (CNN) architecture with four convolutional layers as described in [2]. This particular architecture is chosen for its proven effectiveness in digit recognition tasks within the MNIST context. It serves as a standard against measuring the impact of adversarial attacks.

Turning to the CIFAR-10 dataset, known for its collection of color images across ten diverse categories, we employ the Wide ResNet architecture [23] as the classifier. The selection of Wide ResNet is deliberate; its extensive use in prior defense papers, including those exploring adversarial training [14] and TRADES [24], provides a rigorous benchmark for evaluating the robustness of our adversarial approach. By utilizing classifier networks that are well-established within the research community, we ensure that our experimental results are relevant and that our findings on the effectiveness of the proposed RNN-based adversarial attack can be compared directly with the current state-of-the-art defensive strategies.

4.1.2 Baselines for Comparison

In our Baselines for Comparison section, the evaluation of our RNN-based adversarial attack extends to a comprehensive set of benchmarks, providing a robust context for assessing its performance. The foundational baseline for our comparisons is the Projected Gradient Descent (PGD) attack, which is universally recognized as a standard method for testing the robustness of neural networks against adversarial examples. To ensure a fair and objective comparison, we utilize the source code provided by the original authors on GitHub, meticulously applying the attack using their recommended set of hyper-parameters. This adherence to the established PGD protocol allows us to set a consistent and transparent benchmark for the initial assessment of our attack's effectiveness.

To align our work with the latest advancements in adversarial machine learning, we further compare our attack strategy with AutoAttack, an ensemble method that represents the forefront of adversarial attack techniques. AutoAttack is acclaimed for its comprehensive evaluation of model robustness, incorporating a variety of attack vectors to challenge deep learning models. By incorporating AutoAttack into our baseline comparisons, we acknowledge and position our work within the current state-of-the-art, ensuring that our findings remain relevant and significant in the rapidly evolving landscape of adversarial research.

Our comparative analysis is conducted in a tiered manner. The first level involves the deployment of our attack against models trained using standard techniques, such as the small CNN architecture for the MNIST dataset and the Wide ResNet architecture for CIFAR-10. This initial stage allows us to determine the base level of robustness that conventional training methods offer against adversarial interventions.

Building on this foundation, we then escalate our comparative analysis to adversarially trained models known for their enhanced defense capabilities. From the AutoAttack [6] and RobustBench [5] library, we select models that have been specifically engineered to withstand adversarial attacks. For MNIST, we examine the robustness of the 'TRADES' model [24], which utilizes a small CNN architecture, while for CIFAR-10, we evaluate the 'Carmon2019Unlabeled' model [3], based on the robust Wide ResNet-24-10 architecture.

These models acknowledged for their robustness against adversarial examples, provide a stringent benchmark for evaluating the strength of our proposed adversarial approach.

Our evaluation employs a two-tiered approach: initially, our RNN-based attack is tested against standard training models, and then it is pitted against top-tier adversarially trained models. This strategy reveals the attack's effectiveness across a spectrum of defenses, providing a clear measure of its potential in real-world applications.

4.1.3 Evaluation and implementation details

In our experimental setup, we evaluated the effectiveness of adversarial attacks on classifier models trained on datasets both with and without perturbations, employing the previously mentioned architectures. The adversarial examples were tested against robust classifiers, setting the maximum L_∞ perturbation strength at $\epsilon = 0.3$ for MNIST and $\epsilon = \frac{8}{255}$ for CIFAR-10.

Our methodology includes a learned attacker, which we trained for 10 and 40 steps to assess how training duration impacts the efficacy of attacks and to explore if longer training can lead to a generalized attacker across various steps.

We conducted attacks using 10, 20, 30, 40, 60, and 100 steps to comprehensively test robustness, with a step size of $\epsilon/4$. In addition to using PGD with a step size of $\epsilon/4$, we tested a variant with a step size of $\epsilon/10$ to examine the sensitivity of robustness against step size alterations. We also integrated the leading-edge AutoAttack into our evaluation, employing both the Auto PGD with Cross-Entropy loss (CE) and the Auto PGD with Difference of Logits Ratio (DLR) loss configurations [6]. Additionally, our study includes various ablation experiments to understand the effects of increasing the number of layers in the attacker, altering the step size, and training the attacker with different step sizes.

Through these multifaceted attack iterations and by incorporating various models and loss configurations from AutoAttack, our analysis offers an in-depth perspective on the adversarial landscape. This extensive array of tests is vital for pinpointing strengths and potential

vulnerabilities across different models and attack methods, thus providing a holistic view of the robustness of the classifiers being studied.

4.2 Results

In the present examination, we meticulously evaluate the robustness of models trained using various defense strategies against white-box attacks. The empirical findings, as detailed in Tables 4.1 and 4.2, showcase the superior performance of our proposed Learn-to-Learn (L2L)-based adversarial attack. Our method outperforms traditional approaches across most of the tested scenarios.

Upon review of the results on the MNIST dataset, our proposed L2L-based adversarial attack demonstrates its effectiveness against the TRADES model. The robust accuracy attained by the model against this attack dips as low as 94.62% for certain attack steps, indicating a significant vulnerability. This is in stark contrast to the PGD attack with a step size of $\epsilon/4$, which maintains a higher robust accuracy, not dropping below 95.07%, suggesting that the model is better defended against this form of attack.

Directly contrasting the L2L method with the PGD attack using a step size of $\epsilon/10$, it becomes evident that our method is more effective, as indicated by the lower robust accuracies achieved across all attack steps. This difference in performance is particularly noticeable at the 40-step mark, where the L2L attack lowers the robust accuracy to 94.65% compared to PGD's 95.30

Attacker	**Attack steps**					
	10	**20**	**40**	**60**	**80**	**100**
PGD ($\epsilon/4$)	95.68	95.28	95.12	95.07	95.08	95.08
PGD ($\epsilon/10$)	97.46	96.05	95.30	95.17	95.16	95.14
L2L	**95.32**	**94.98**	**94.65**	**94.70**	**94.62**	**94.68**

Table 4.1: Comparison with baseline adversarial attacks on the TRADES model with small-CNN architecture for MNIST dataset

In the more challenging CIFAR-10 dataset, the L2L-based attack continues to demonstrate its efficacy, particularly in scenarios with lower attack steps of 10 and 20. The robust accuracy for the 'Carmon2019Unlabeled' model employing the Wide-Resnet-24-10 architecture declines to as low as 63.97% under the L2L attack, compared to a reduction to 62.33% in the most potent PGD attack with a step size of $\epsilon/10$. These results highlight the nuanced yet effective performance of our L2L approach in diverse adversarial contexts.

Attacker	**Attacker Steps**					
	10	**20**	**40**	**60**	**80**	**100**
PGD ($\epsilon/4$)	65.11	64.37	64.15	64.08	64.02	64.03
PGD ($\epsilon/10$)	72.51	64.29	**62.68**	**62.38**	**62.50**	**62.33**
L2L	**64.32**	**64.09**	64.03	63.97	63.97	63.97

Table 4.2: Baseline performance comparison on 'Carmon2019Unlabeled' with Wide-Resnet-24-10 for CIFAR-10

In conclusion, the detailed results demonstrate the L2L-based attack's notable efficiency, especially evident in the MNIST dataset where it consistently surpasses PGD in reducing the model's robust accuracy. In the CIFAR-10 dataset, the attack's performance is particularly effective at lower attack steps, indicating a nuanced relationship with model complexity and defense mechanisms. The step-free nature of the L2L attack adds to its adaptability and effectiveness, marking it as a significant advancement in adversarial attack methodologies. These outcomes not only illustrate the L2L method's potential in varying adversarial contexts but also highlight the importance of continuous innovation in developing robust defense strategies for machine learning models.

4.3 Analysis

The cornerstone of this analysis is a rigorous comparison of our model's performance against the state-of-the-art AutoAttack. It must be stressed that unlike other attacks, which often rely on hand-designed rules and predefined step sizes, our model operates free of such con-

straints. This step-size-free characteristic is integral to our approach, allowing for a more flexible and potentially more potent adversarial strategy.

4.3.1 Comparison with auto-attack

When analyzing the results on the MNIST dataset, as detailed in Table 4.3, our model exhibits a strong initial performance. Specifically, our L2L-based attack outperforms traditional methods for attack steps up to and including 30 iterations. This is indicative of our model's proficiency in quickly identifying and exploiting vulnerabilities within the TRADES model. However, beyond this point, the AutoAttack begins to demonstrate its robustness, overshadowing our model's performance.

Attacker	**Attack Steps**					
	10	**20**	**40**	**60**	**80**	**100**
PGD ($\epsilon/4$)	95.680	95.280	95.120	95.070	95.080	95.080
PGD ($\epsilon/10$)	97.460	96.050	95.300	95.170	95.160	95.140
L2L	**95.320**	**94.980**	**94.650**	94.700	94.620	94.680
Auto PGD-ce	96.420	96.150	95.310	**94.560**	**94.280**	**94.120**
Auto PGD-dlr	96.650	96.410	95.500	94.960	94.510	94.720

Table 4.3: Comparison with state-of-the-art adversarial attacks on the TRADES model with smallCNN architecture for MNIST dataset

Moving to the CIFAR-10 dataset, the scenario presents a more challenging environment for our L2L-based model. While the results, as seen in Table 4.4, do not show our model outperforming AutoAttack, they do indicate that our approach remains competitive. The performance is close to that of the state-of-the-art, suggesting that with further refinement, our step-size-free model could potentially match or surpass the current leading methods.

In both datasets, the implications of a step-size-free adversarial model are profound, offering a new perspective on the development of attack strategies. By removing the reliance on step size, our model demonstrates adaptability and a potential for enhanced efficacy,

Attacker	Attack Steps					
	10	**20**	**40**	**60**	**80**	**100**
PGD ($\epsilon/4$)	65.110	64.370	64.150	64.080	64.020	64.030
PGD ($\epsilon/10$)	72.510	64.290	**62.680**	**62.380**	62.500	**62.330**
L2L	64.320	64.090	64.030	63.970	63.970	63.970
Auto PGD-ce	64.220	63.980	63.680	63.560	63.510	63.530
Auto PGD-dlr	**63.190**	**62.980**	62.720	62.510	**62.470**	62.460

Table 4.4: Comparison with state-of-the-art adversarial attacks on the 'Carmon2019Unlabeled' model with Wide-Resnet-24-10 architecture for CIFAR-10 dataset

which may prove vital in advancing the arms race between adversarial attacks and defenses in machine learning.

4.3.2 Increasing the number of neurons in RNN

In an effort to enhance the efficacy of our Learn-to-Learn (L2L) adversarial model, we explored the impact of scaling up the RNN's complexity by increasing the number of neurons. This adjustment was hypothesized to potentially improve the model's capacity to generate more effective adversarial examples by capturing more complex patterns and dependencies within the data.

The results of this modification are presented in Table 4.5. When applied to the small-CNN model, we observed a trend where the increased RNN size led to a measurable improvement in the attack's performance, particularly at smaller step sizes. For instance, the L2L model with an enhanced RNN consisting of 15 neurons showed a slight increment in performance at the 10 and 20 attack steps compared to the standard L2L model.

This observation suggests that the model's ability to perturb the input data in a way that is more challenging for the defense mechanisms to counteract benefits from a larger neural capacity. It indicates a threshold of complexity within the RNN's architecture that, when exceeded, can yield more potent adversarial examples.

Model	**Attack Steps**					
	10	**20**	**40**	**60**	**80**	**100**
L2L	95.320	94.980	**94.650**	94.700	**94.620**	94.680
L2L (15 RNN)	**95.300**	**94.960**	94.780	**94.640**	94.630	**94.620**

Table 4.5: Impact of increasing the number of neurons in RNN on adversarial attack effectiveness

The implications of this finding are significant for the design of adversarial attacks. It points to the possibility that there exists an optimal size for the RNN that balances the computational efficiency with the attack's effectiveness, advocating for further research into the relationship between RNN size and adversarial success.

CHAPTER 5

Conclusion and Future work

5.1 Conclusion

The research presented in this thesis has made several substantial contributions to the field of adversarial machine learning. Through the development and evaluation of a Learn-to-Learn (L2L) based adversarial attack model, we have demonstrated the potential for step-size-free attacks to produce significant perturbations, challenging the robustness of well-established defense models.

Our findings have consistently shown that the L2L approach can outperform traditional adversarial methods, such as PGD, particularly on the MNIST dataset with the TRADES model. Notably, the model's performance was markedly superior at smaller attack steps, showcasing the effectiveness of the L2L methodology in quickly identifying and exploiting model vulnerabilities. Even when faced with the more complex CIFAR-10 dataset, the L2L model displayed a close competition with state-of-the-art techniques, hinting at the untapped potential of step-size-free adversarial models.

Furthermore, the investigation into the impact of increasing the RNN's neuron count revealed promising results. On both the smallCNN and WideResNet architectures, enhancing the RNN size proved beneficial up to a certain complexity threshold, beyond which the returns diminished. These results have underscored the importance of carefully considering the trade-offs between model complexity and computational efficiency.

The adaptability and flexibility of our step-size-free adversarial model stand out as its most significant advantage. This characteristic not only opens new avenues for crafting

more potent attacks but also underscores the need for developing robust defenses that can withstand such adaptable threats.

In light of these contributions, future work should aim to further explore the limits of step-size-free adversarial training. This includes investigating the optimal RNN size for various model architectures and attack scenarios, as well as examining the potential for real-time application in dynamic environments. The goal moving forward is to refine these adversarial techniques to enhance their efficacy while also bolstering the resilience of machine learning models against such innovative attacks.

The journey through this research has been one of discovery and innovation. The implications of this work are far-reaching, providing a foundation for more secure and reliable machine learning applications in an era where adversarial threats continue to evolve.

This thesis presented a novel approach to generating adversarial attacks in deep learning models, emphasizing a parameter-free methodology utilizing a learning-to-learn (L2L) framework. Our approach innovatively employed a recurrent neural network (RNN) based optimizer, marking a significant departure from traditional methods reliant on predetermined parameters and manual rule designs.

Our extensive experiments demonstrated the efficacy of the learned optimizer, particularly in the context of small networks and datasets like MNIST. The results indicated that our method outperforms traditional attacks, such as Projected Gradient Descent (PGD), especially when smaller step sizes are used. This signifies a notable advance in the adaptability and efficiency of adversarial attacks, highlighting the potential of parameter-free methods in assessing and understanding the robustness of deep learning models.

The research also critically assessed the limitations of existing L2L frameworks in learning an optimizer for adversarial attacks. We found that while the L2L framework showed promise, its applicability was not universal, underscoring the need for caution in its widespread application. This insight is crucial for future research in the field.

Moreover, we observed that training the optimizer for fewer steps can generate effective perturbations, more so than those produced by PGD, although this advantage diminishes

with larger training step sizes. This finding serves as an essential guideline for future work regarding the configuration of step sizes in optimizer training.

In conclusion, this thesis contributes significantly to the field of adversarial robustness in deep learning. By developing a parameter-free L2L-based adversarial attack and exploring its nuances, we provide a foundation for future research aimed at creating more robust and secure deep learning models. The work encourages a shift towards more adaptive, efficient, and effective adversarial attack methodologies, a crucial step in the ongoing development of resilient AI systems.

5.2 Future work

In future work, we aim to extend and refine the capabilities of the Learn-to-Learn (L2L) adversarial framework presented in this thesis. A primary focus will be on adjusting the L2L network architecture to enhance its effectiveness in adversarial attack scenarios, incorporating the latest advancements in machine learning techniques to improve efficiency and potency. This will involve not only optimizing the existing framework but also exploring new strategies and methodologies, particularly in the realm of black-box attacks, to broaden the applicability of our model in more realistic and varied scenarios. Additionally, a significant extension of our research will be the application and testing of our learned optimizer on Vision Transformer (ViT) models. ViTs, with their unique architectural approach, offer a novel testing ground for our methods, potentially revealing insights into the adaptability and effectiveness of the L2L framework across different neural network architectures.

Alongside these developments, we plan to conduct extensive testing across a wider array of datasets and model types. Such comprehensive validation is crucial to ascertain the strengths, limitations, and areas for improvement of our approach, ensuring its robustness and relevance in the rapidly evolving landscape of adversarial machine learning. Through these endeavors, our future work aspires not just to enhance the current model but also to make substantial contributions to the field, addressing both theoretical and practical

challenges in adversarial machine learning.

Bibliography

[1] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29, 2016.

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[3] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *Advances in neural information processing systems*, 32, 2019.

[4] Neil E Cotter and Peter R Conwell. Fixed-weight networks can learn. In *1990 IJCNN international joint conference on neural networks*, pages 553–559. IEEE, 1990.

[5] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *CoRR*, abs/2010.09670, 2020.

[6] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[7] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[9] Yunseok Jang, Tianchen Zhao, Seunghoon Hong, and Honglak Lee. Adversarial defense via learning to generate diverse attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2740–2749, 2019.

[10] Louis Kirsch and Jürgen Schmidhuber. Meta learning backpropagation and improving it. *Advances in Neural Information Processing Systems*, 34:14122–14134, 2021.

[11] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research), 2010.

[12] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[13] Kaifeng Lv, Shunhua Jiang, and Jian Li. Learning gradient descent: Better generalization and longer horizons. In *International Conference on Machine Learning*, pages 2247–2255. PMLR, 2017.

[14] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[15] Luke Metz, Niru Maheswaranathan, Jeremy Nixon, Daniel Freeman, and Jascha Sohl-Dickstein. Understanding and correcting pathologies in the training of learned optimizers. In *International Conference on Machine Learning*, pages 4556–4565. PMLR, 2019.

[16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[17] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[18] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016.

[19] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[20] Olga Wichrowska, Niru Maheswaranathan, Matthew W Hoffman, Sergio Gomez Colmenarejo, Misha Denil, Nando Freitas, and Jascha Sohl-Dickstein. Learned optimizers that scale and generalize. In *International conference on machine learning*, pages 3751–3760. PMLR, 2017.

[21] Yuhuai Wu, Mengye Ren, Renjie Liao, and Roger Grosse. Understanding short-horizon bias in stochastic meta-optimization. *arXiv preprint arXiv:1803.02021*, 2018.

[22] Yuanhao Xiong and Cho-Jui Hsieh. Improved adversarial training via learned optimizer. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 85–100. Springer, 2020.

[23] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[24] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.