CS282 Final Project: Attacks Which Do Not Kill Training Make Adversarial Learning Stronger

Ken Chen
ShanghaiTech University
2020533036
chenken@shanghaitech.edu.cn

Boke Chen
ShanghaiTech University
2020533035
chenbk@shanghaitech.edu.cn

Chunbo Xu
ShanghaiTech University
2022233271
xuchb2022@shanghaitech.edu.cn

Abstract

This report serves as the final documentation for our CS282 project. The chosen article for our analysis is titled 'Attacks Which Do Not Kill Training Make Adversarial Learning Stronger.' Within this report, we provide a comprehensive review of the research problem (Section 1). Additionally, we delve into the background, significance, and practical applications of the problem (Section 2). We also examine the methodology employed by the author (Section 3) and highlight the article's innovative aspects (Section 4). Furthermore, our report emphasizes our own contributions to this project. Firstly, we verified the findings of the article by re-implementing the code and presenting the results of our verification (Section 5). Additionally, we proposed novel optimization ideas that build upon the concepts presented in the article (Section 6).

1 Research problem

Based on the minimax formulation, adversarial training is crucial for ensuring the robustness of a trained model. However, the conventional adversarial training strategy employs PGD to generate adversarial samples for training. These samples are exceedingly "strong", causing them to be classified on the opposite side of the decision boundary by the current model. Consequently, training with such samples can adversely affect the natural accuracy and even impede model convergence.

To address this issue, the author conducted research on identifying appropriate adversarial samples and proposed a novel approach termed "friendly adversarial training." Instead of maximizing the loss using the most adversarial data, this approach utilizes the least adversarial data (referred to as friendly adversarial data) to minimize losses during training. The author's investigation revealed that terminating the search algorithm for the most adversarial data at an earlier stage yields favorable results. The rationale behind this lies in the observation that prolonged iterations tend to push the adversarial samples further away from the decision boundary. Hence, early stopping effectively alleviates this phenomenon.

Spring 2023, CS282, Machine Learning, ShanghaiTech University.

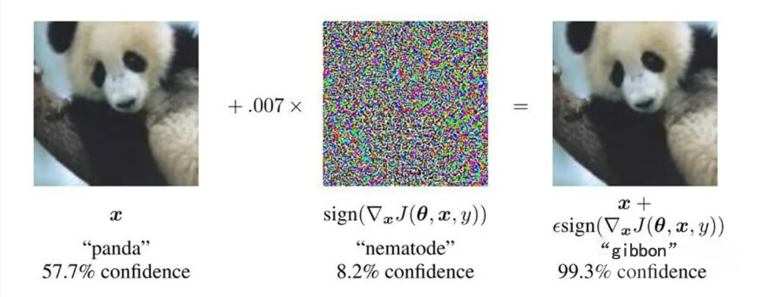
2 Basic Information

2.1 Background

2.1.1 Adversarial Data

Adversarial data refers to the combination of input natural data of deep neural networks and artificially introduced noise.

Deep neural networks can be easily confused by adversarial data. For instance, if we take a picture of a panda and add artificially synthesized noise to it, we obtain a new picture of a panda. While the human visual system would correctly classify this new picture as a panda, the neural network may misclassify it and mistakenly identify it as a picture of a gibbon.

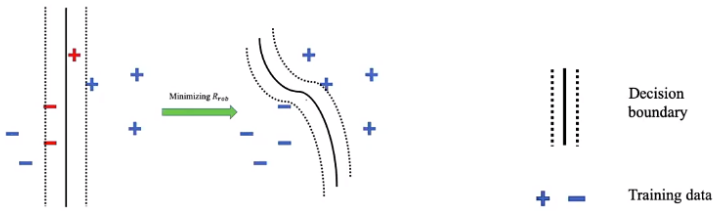


Misclassification of the Neural Network

The impact of adversarial data can be significant. Deep neural networks are extensively employed in various domains, including driverless cars, medical diagnosis, financial analysis, and more. Consequently, enhancing the adversarial robustness of neural networks is of utmost importance.

2.1.2 Adversarial Learning

Adversarial learning stands out as the most effective approach to enhance adversarial robustness. The fundamental concept revolves around generating adversarial data during the network training process and subsequently updating the network parameters based on the acquired adversarial data. This iterative process enables the network to learn and adapt to the adversarial nature of the data.



What the Adversarial Learning can do

The ultimate goals of adversarial learning can be summarized as follows:

- Accurate Data Classification:
  The primary objective is to ensure correct classification of data points, where the network can accurately assign the appropriate labels to input samples.
- Expanding Classification Boundary:
  Adversarial learning aims to widen the classification boundary, represented by the two dashed lines, in order to increase the margin of separation between different classes. By expanding the boundary, the network becomes more resilient to adversarial perturbations and reduces the likelihood of misclassifying data points that lie within the boundary region.

2.2 Significance

Once a neural network is trained in deep learning, it may exhibit high classification accuracy, but its robustness against adversarial attacks can be weak. These attacks involve introducing subtle perturbations to input images that are imperceptible to the human eye but significantly impact the network's classification accuracy. To address this vulnerability, training the network using friendly adversarial data can be highly effective. By incorporating such data during the training process, the classification accuracy of the neural network can be improved, leading to overall enhanced network performance.

2.3 Applications

The adversarial robustness of deep neural networks is of paramount importance in safety-critical domains such as medicine and autonomous driving. Current research in this area employs two primary adversarial training methods: maximal and minimal optimization. For instance, Projected Gradient Descent (PGD) generates the most adversarial data to maximize the loss, but this approach tends to be overly conservative. The authors of the paper propose a minimax optimization technique that significantly improves the training outcomes, ensuring greater precision and safety in domains like medicine and autonomous driving.

By enhancing the adversarial resistance of deep neural networks, we can achieve various benefits in these fields. In medicine, this improvement translates to more accurate medical imaging, enabling doctors to identify lesions and abnormalities with higher precision, thereby improving diagnostic accuracy and speed. Similarly, in autonomous driving systems, the networks' ability to accurately identify obstacles is enhanced, leading to optimized driving behavior and path planning. These advancements contribute to overall safety and efficiency in these critical domains.

3 Methodology

Quoted from Nietzsche's famous aphorism, 'What doesn't kill me makes me stronger,' the author draws inspiration and questions the conventional approach to adversarial training. In response, the author introduces a novel strategy called friendly adversarial training.

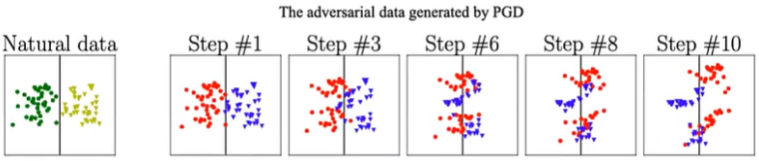
Firstly, the author delves into the fundamentals of traditional adversarial learning, which relies on the minimax formulation:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(\tilde{x_i}), y_i), \qquad where \quad \tilde{x_i} = \arg \max_{x \in B(x_i)} \ell(f(\tilde{x}), y_i)$$

In the inner layer of the minimax formulation, adversarial data is obtained by maximizing the loss function ℓ. Subsequently, in the outer layer, the adversarial data is utilized to minimize the loss function. An effective technique that employs this formulation is Projected Gradient Descent (PGD), which iteratively discovers adversarial data that maximizes the loss function. Consequently, the adversarial data generated through PGD has a high likelihood of causing errors in deep neural networks.

The author identifies several challenges associated with the traditional adversarial learning approach. These challenges include the trade-off between robustness and accuracy, wherein improving robustness may come at the expense of accuracy. Additionally, the computation of the loss function for data derivation is highly time-consuming.

Next, the author critically examines the traditional adversarial learning model and raises concerns about its foundations, particularly the minimax formulation. The author identifies a key issue with the approach of using PGD to search for highly adversarial data through multiple iterations. The resulting data is an approximate solution to the inner maximization problem. The accompanying diagram illustrates this point: the red points represent adversarial data corresponding to the green points, while the blue points represent adversarial data corresponding to the yellow points. At Step #10 of the PGD process, the found adversarial data causes the blue points to cross over into the green region, and the red points to cross over into the yellow region. This phenomenon results in a significant mixing of natural and adversarial data, exacerbating the problem.



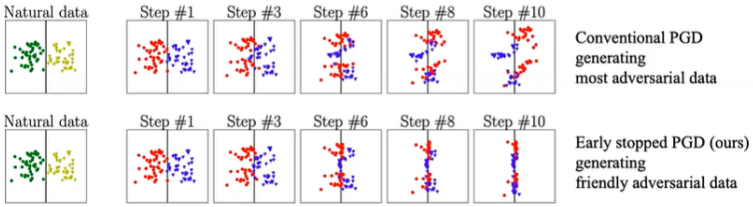
The traditional outer minimization process in adversarial learning, when attempting to learn from the red and blue points, inevitably leads to the learning of an inaccurate classification model. This is the root cause of the significant accuracy sacrifice associated with adversarial learning.

To address this issue, the author introduces a novel adversarial learning formulation known as the min-min formulation:

$$\tilde{x}_i = \arg\min_{\tilde{x} \in B(x_i)} \ell(f(\tilde{x}), y_i) \qquad s.t. \quad \ell(f(\tilde{x}), y_i) - \min_{y \in \mathcal{Y}} \ell(f(\tilde{x}), y_i) \geq \rho$$

The min-min formulation introduced by the author modifies the inner maximization process, which aims to find friendly adversarial data instead of the most adversarial data (as in traditional approaches). In this new formulation, the objective is to minimize the loss function under specific conditions when searching for friendly adversarial data.

In their quest to implement the min-min formulation, the authors devised a method that approximates the desired formula through iterative experimentation. They achieved this by introducing an early stopping mechanism during the PGD search process, hence the term "early stopped PGD." When conducting the search for adversarial data, if PGD detects that the current data is already misclassified, it terminates the search early. The resulting adversarial data generated using this approach is referred to as "friendly adversarial data" by the authors.



Comparasion of the Two PGD

4 Innovative Aspects

4.1 What fascinates us

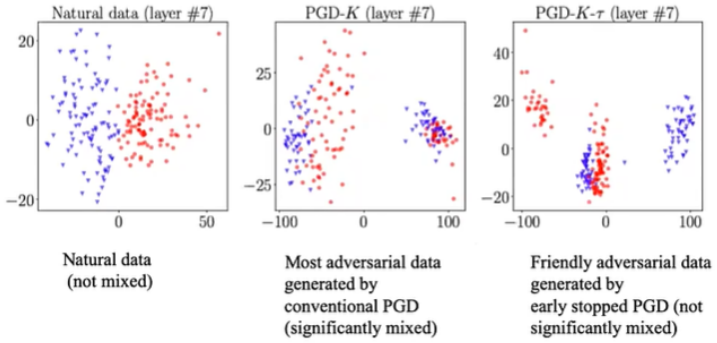
The underlying philosophical principle of this article, encapsulated by the famous quote "What doesn't kill me will make me stronger," serves as a compelling motivation to delve deeper into its content. Observing the author's progression from abstract concepts to concrete implementation is truly captivating and holds our interest throughout the reading experience.

4.2 Innovations and Highlights

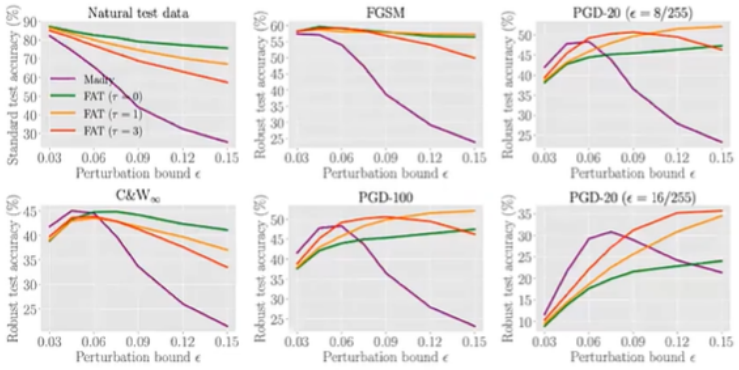
One of the key innovations of this article lies in the author's enhancement of the conventional adversarial learning approach. By introducing a novel set of formulas, the author has successfully refined the process of selecting and filtering training data. This innovative approach sets the article apart from existing methods and contributes to the advancement of adversarial learning techniques.

- In contrast to traditional adversarial learning, friendly adversarial learning offers a solution to the cross-over mixture problem. As depicted in the figure, while this problem may not

arise in the input space of an image classification task, it is often encountered in intermediate layers of the neural network. However, by employing friendly adversarial data, as proposed by the author, this issue can be effectively mitigated. This approach demonstrates its efficacy in addressing the challenge of cross-over mixing within the network architecture.



- In contrast to traditional adversarial learning, friendly adversarial learning offers the advantage of being more time-efficient. This is achieved through the utilization of an early-stopping PGD algorithm, which eliminates the need for further backward propagation. As soon as misclassified friendly adversarial data is identified, the search process is immediately halted. This early stopping mechanism significantly reduces the computational time required for training, making friendly adversarial learning a more time-saving approach.
- In comparison to traditional adversarial learning, friendly adversarial learning enables the attainment of a larger protection radius.



The minimax formulation in traditional adversarial learning aims to find highly adversarial data, which can result in significant cross-mixing issues. As the protection radius increases, this cross-mixing problem becomes more pronounced, eventually leading to training failure (as indicated by the purple line in the above picture). In contrast, friendly adversarial learning (FAT) effectively mitigates the cross-mixing problem, allowing for a wider protection radius to be accommodated. By focusing on the generation of friendly adversarial data, FAT ensures that the impact of adversarial perturbations is minimized, enabling the model to tolerate a larger protection radius without compromising its performance or training stability.

5 Re-implement and Experiments Results

Although we faced the problem of limited equipment resources, we managed to generate a smaller dataset to replace the original cifar10 and run it successfully. Here's the result:

Epoch	Natural Test Acc	FGSM Acc	PGD20 Acc	CW Acc
1.000000	0.365600	0.188600	0.112100	0.114200
2.000000	0.554100	0.279800	0.182300	0.193500
3.000000	0.650100	0.339900	0.205900	0.219800
4.000000	0.685300	0.350100	0.211100	0.222400
5.000000	0.710800	0.388600	0.238100	0.253400
6.000000	0.725500	0.426200	0.283700	0.297300
7.000000	0.756700	0.420800	0.258500	0.275700
8.000000	0.753500	0.423700	0.264800	0.280000
9.000000	0.744300	0.455300	0.298700	0.316900
10.000000	0.735400	0.415300	0.245600	0.273000

Results for FAT

Epoch	Natural Test Acc	FGSM Acc	PGD20 Acc	CW Acc
1.000000	0.481200	0.283400	0.223400	0.209300
2.000000	0.557100	0.333700	0.252300	0.232000
3.000000	0.615900	0.375500	0.288800	0.272700
4.000000	0.633900	0.403100	0.323200	0.304300
5.000000	0.667300	0.418000	0.319300	0.301500
6.000000	0.671100	0.445200	0.345300	0.327100
7.000000	0.700800	0.460000	0.364900	0.348200
8.000000	0.704100	0.462800	0.359400	0.349900
9.000000	0.745200	0.502800	0.386500	0.364600
10.000000	0.756500	0.523700	0.389600	0.386700
11.000000	0.761100	0.528000	0.395500	0.389300
12.000000	0.763200	0.513400	0.392700	0.381500
13.000000	0.751200	0.514800	0.395200	0.383200

Results for FAT_for_TRADES

From the running results, compared with FAT, FAT_for_TRADES shows significant advantage under "CW Acc", "PGD20 Acc" and "FGSM Acc", which is in line with the results given by the authors.

The code is in a separate folder at "./Friendly-Adversarial-Training".

6 Novel Optimization Ideas

- Further explore alternative methods for generating friendly adversarial data:
  While the author used early-stopped PGD algorithm to generate friendly adversarial data, it would be beneficial to investigate other generation methods such as those based on Generative Adversarial Networks (GANs) or other optimization algorithms. This would broaden the range of options for generating friendly adversarial data and potentially enhance the effectiveness of adversarial training.
- Incorporate transfer learning:
  Transfer learning can improve the generalization performance of adversarial samples. By combining pre-trained models during the training process, existing knowledge can be leveraged to enhance the classification accuracy and robustness of adversarial samples.
- Consider the adaptability to different tasks and datasets:

The performance of friendly adversarial data may be influenced by the specific task and dataset. Therefore, it would be valuable to study the adaptability of friendly adversarial data to different tasks and datasets, and optimize it accordingly for specific tasks and datasets.

- Combine different adversarial training strategies:

  In addition to friendly adversarial data, it would be worthwhile to explore the combination of other adversarial training strategies with friendly adversarial data to further improve model robustness. For example, incorporating Generative Adversarial Networks (GANs) or other adversarial training methods could leverage the strengths of different strategies and achieve stronger adversarial training results.

- Consider domain-specific optimization strategies:

  For specific domains of application, it is important to consider domain-specific optimization strategies. For instance, in the field of medicine, incorporating prior knowledge and specific requirements of medical imaging or medical data could lead to the design of tailored methods for generating friendly adversarial data and conducting adversarial training.

7 Conclusion

In conclusion, this report explores the concept of friendly adversarial data and its implications for adversarial learning. The author addresses the limitations of traditional adversarial training methods and proposes the min-min formulation as an alternative approach. By focusing on generating friendly adversarial data instead of extremely adversarial data, the aim is to improve the robustness and accuracy of deep neural networks.

In addition, we have replicated the author's work within the limitations of our resources, including code modifications, environment setup, code execution, etc. We have successfully obtained our validation results, confirming the correctness of the author's work. Furthermore, based on our understanding, we have provided novel optimization suggestions for this article.