Enhancing Robustness in Feature Learning through Adversarial Training

Hai Dang Nguyen
School of Data Science
City University of Hong Kong
hdnguyen4-c@my.cityu.edu.hk

Wai Nam Tung
School of Data Science
City University of Hong Kong
wntung4-c@my.cityu.edu.hk

Quoc Minh Duong
School of Data Science
City University of Hong Kong
mduong2-c@my.cityu.edu.hk

***Abstract*—In a world where adversarial attacks occur regularly; neural networks often collapse and produce incorrect predictions due to their sensitivity to these perturbations. In this research, we examined the accuracy performance of the neural networks by introducing two kinds of adversarial attacks on both defended and undefended models. We concluded that adversarial training assists the model in learning features that are both important for prediction and resistant to perturbations.**

***Index Terms*—Adversarial training, feature learning, deep neural networks, adversarial attack**

I. Introduction

Feature learning plays an important role in the performance and effectiveness of a model in many machine learning algorithms. Therefore, in the deployment of machine learning models, much attention and effort are put into feature engineering. However, this method is manual, labor-intensive, and relies heavily on human expertise. The advent of deep neural networks (DNNs) has revolutionized feature learning. Unlike traditional feature engineering, DNNs could learn hierarchical and abstract features automatically through multiple layers of interconnected neurons. This automatic feature learning process has enabled DNNs to perform exceptionally across various domains.

However, a slight modification in the input data could cause the behaviors of DNNs to change remarkably. This leads to DNNs being vulnerable to adversarial attacks, which are crafted perturbations to input data that aim to mislead the model. To the model, these perturbations can lead to misclassification or incorrect predictions, while to the human eye, they are often not observable or should not affect the outputs. This highlights the need for propelling deep neural networks to learn features that are robust and interpretable from the human perspective. A solution is incorporating adversarial training into the model. Adversarial training involves training a DNN on both clean examples and adversarial examples. This enables the model to defend against adversarial attacks and learn robust features.

In this project, we aim to leverage feature learning by performing adversarial training on a deep neural network. More specifically, we perform two types of untargeted white box attacks: Fast Gradient Sign Method (FGSM) and different iterations for Projected Gradient Descent (PGD) on a DNN for image classification and then optimize the model's feature learning by including adversarial examples in training. We seek to build a robust and resilient model that can handle adversarial perturbations while maintaining accurate predictions. The performance of the model before and after adversarial training will then be recorded and evaluated later.

II. Related Works

Alexey Kurakin, Ian Goodfellow, and Samy Bengio's paper "Adversarial Machine Learning at Scale" explores how resilient large deep learning models, particularly Inception v3 trained on extensive datasets (i.e., ImageNet) are when attacked with adversarial examples. Important contributions consist of the effective application of adversarial training to dramatically improve resilience to adversarial attacks that take one step at a time. Yet, their research was concentrated on FGSM attacks, as they claim that iterative attacks are too computationally expensive and don't offer any appreciable advantages. Therefore, we incorporate the iterative method of Projected Gradient Descent (PGD) to increase the diversity of our attacks across different adversarial trained models. This method enables a more sophisticated and stronger adversarial perturbation.

The paper "Towards Deep Learning Models Resistant to Adversarial Attacks" investigates how adversarial examples can weaken deep neural networks. They provide universal and trustworthy neural network training and attack techniques that offer evident security assurances against any adversary. Nevertheless, the research notes that whereas training on adversarial examples in one step increases robustness against these types of attacks, it does not give iterative adversarial examples robustness. We run a similar but modified experiment to verify if our results are consistent with the research publication to confirm whether adversarial training can improve the robustness of the model and whether specialized adversarial training only works well on specific attacks.

A novel training methodology is proposed in the paper "Interpolated Adversarial Training: Achieving Robust Neural Networks Without Sacrificing Too Much Accuracy" to enhance the adversarial robustness of neural networks while preserving high accuracy on unperturbed data. Interpolated Adversarial Training (IAT) is a technique that combines adversarial training with interpolation-based training approaches.

In comparison to conventional adversarial training techniques, the authors show that IAT dramatically lowers the standard test error while preserving similar adversarial robustness. In view of this, we also employed IAT to assess the precision in forecasting unperturbed images as well as those containing FGSM and PGD attacks, and to determine whether our findings align with the existing ones.

An extensive overview of adversarial training techniques for improving the adversarial robustness of deep learning models can be found in the paper "Recent Advances in Adversarial Training for Adversarial Robustness". It aims to emphasize how well adversarial training (AT) strengthens a model's resistance to adversarial attacks and talks about possible future developments in this area. However, the inherent trade-off between adversarial robustness and standard accuracy in AT methods is not addressed in this paper. Thereby, we will investigate whether the issue of the trade-off between robustness and accuracy arises in our experiments and assess whether IAT is sufficient to address this issue.

With an emphasis on deep neural networks (DNNs), the paper "A Review of Adversarial Attack and Defence for Classification Methods" reviews multiple adversarial attack and defense strategies for classification models. It discusses how adversarial examples, which are purposely constructed inputs that can fool a model while seeming normal to humans, can undermine classification models. The study analyses defense strategies like adversarial training, randomization, and projection in addition to classifying attack techniques into gradient-based, score-based, and decision-based approaches. For the sake of simplicity, we limit our experiment to using only adversarial training as a defense strategy, leaving gradient-based attack methods for evaluation.

III. Description

Our experiment's primary objective is to assess how resilient each defense model is to various perturbed images produced by adversarial attacks. Model performance is measured via accuracy rates. An improvement in accuracy rate indicates that adversarial training models are more capable of extracting robust and general features from perturbed images for classification tasks, and vice versa. The dataset, neural networks, attacks, and adversarial training models applied in our experiment are all described in detail below.

A. Dataset

The Kaggle animal image dataset gathered 5400 animal images in 90 different categories or classes. The animals on the list include goats, flamingos, eagles, elephants, and foxes, among others, and are organized into distinct files with label names that match. These equally distributed photos served as natural and adversarial training set for a neural network designed to recognize images and test set after adding perturbations. We divide the data set into two parts with the ratio 8:2 for the training set and the test set with stratified split to ensure that each split of the data represents the overall proportion of the dataset accurately.

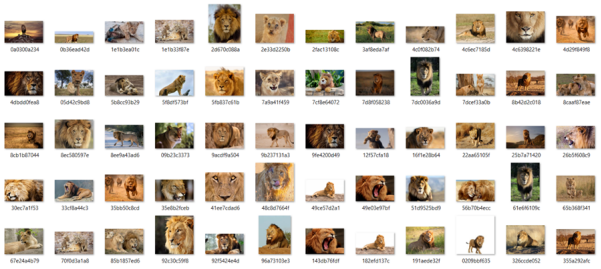

Fig. 1. Structure of the dataset

B. Baseline model

We finetuned ResNet-50 architecture, a deep convolutional neural network comprising 50 layers, which serves as the foundational baseline model for our study. This pre-trained network is capable of classifying images into 1,000 distinct categories, encompassing a diverse array of objects such as various animals, and commonplace items like keyboards, mice, and pencils. The pre-trained model, beneficial for our purposes, comes equipped with weights that have been refined through exposure to a large and varied dataset. Our methodology involves adapting the ResNet-50 model to suit our specific research needs. Initially, this is achieved by modifying the network's output layer to feature 90 outputs. To leverage the pre-acquired learning of the model, we employ a two-phase fine-tuning process. In the first phase, all layers except the last are frozen, and the model undergoes training using the Adam optimization algorithm with a learning rate of 0.0005 across five epochs. Subsequently, in the second phase, we unfreeze all layers and conduct further training for ten epochs, employing a learning rate of 0.0001. This approach aims to fine-tune the model effectively, making it more adept at handling the specific task at hand.

C. Adversarial attacks

Two types of adversarial attacks methods are applied to perturbate the animal images, they are Fast Gradient Sign Method (FGSM) and projected gradient descent (PGD).

Fast Gradient Sign Method (FGSM) is a method for producing adversarial samples. It operates by creating a new input that is only marginally different from the original but is classified differently by the model using the gradients of the model. The term "fast" describes its computational efficiency, which makes it an appealing choice for evaluating a model's resilience.

Projected gradient descent (PGD) is a more advanced and powerful method than the Fast Gradient Sign Method (FGSM). It is an extension of the fundamental gradient descent technique because it generates adversarial samples iteratively. The procedure takes several tiny steps in the direction that maximizes the model's prediction error. The perturbation is typically projected back into a set of permitted perturbations after each step to guarantee that the adversarial example stays within certain limits.

We used two PGD attacks in our research, ranging in strength from seven to twenty iterations. The aim is to investigate the change in the same model's adversarial training accuracy. It is our hypothesis that the accuracy of the model degrades as higher iterations make the attack easier to succeed and less imperceptible.

D. Mix-up technique

Mixup is a data augmentation method that generates virtual training examples by linearly interpolating between pairs of examples and their corresponding labels from the training data. Specifically, for two randomly chosen examples (x_i, y_i) and (x_j, y_j), where x_i, x_j are raw input vectors and y_i, y_j are their one-hot label encodings, the Mixup technique creates new examples $(\tilde{x}, \tilde{y})$ using the formula:

$$\tilde{x} = \lambda x_i + (1\text{-}\lambda) x_j$$

$$\tilde{y} = \lambda y_i + (1\text{-}\lambda) y_j$$

Here, λ is a parameter drawn from a distribution, such as the Beta distribution, within the range [0, 1]. This approach expands the training data range by utilizing the pre-existing understanding that linear combinations of feature vectors should correspond to linear combinations of the related targets. One of the key advantages of Mixup is its simplicity and minimal computational overhead, making it a practical choice for enhancing the training process of machine learning models.

E. Adversarial training models

In this study, we have successfully developed three innovative defense models. We named these models based on the type of adversarial training employed: the FGSM model utilizes the Fast Gradient Sign Method (FGSM) attack, the PGD model employs the Projected Gradient Descent (PGD) attack, and the IAT model incorporates Interpolated Adversarial Training (IAT). The training process for each of these models closely mirrors that of our baseline model, yet there are some critical differences. For the FGSM and PGD models, we enhance the training process by introducing an adversarial component. Each batch of training data is subjected to an adversarial attack (FGSM for the FGSM model, PGD for the PGD model) to produce perturbed images. These perturbed images, along with their original versions, are then processed by the model. The approach in the IAT model is slightly different. Here, we first calculate loss using unperturbed images that have been altered using the Mixup technique (as described by Zhang et al., 2017). After updating the model parameters based on this step, we then apply the PGD attack to the original images, creating perturbed versions. These perturbed images are treated in the same manner as their unperturbed counterparts. An important aspect to highlight is the consistency in the application of the PGD attack across the IAT and PGD models. In both cases, we use a configuration of seven iterations with an epsilon value of 0.1.

IV. Evaluation

In this section, we present a comprehensive evaluation of our project for feature learning with adversarial training applied to our model. We assessed its performance before and after applying adversarial training. The results are presented in the table below:

TABLE I
ACCURACY OF THE MODEL WITH NORMAL INPUTS AND DIFFERENT TYPES OF ATTACKS (%)

	Normal inputs	FGSM attack	PGD attack (7 iterations)	PGD attack (20 iterations)
Baseline Model	94.81	23.61	1.85	2.13
FGM Model	74.35	49.91	24.67	10.37
PGD Model	74.81	50.19	58.24	38.89
IAT Model	79.07	47.87	53.80	28.06

In general, the performance of the model against different adversarial attacks has improved significantly after applying adversarial training. After being trained with adversarial examples, the model's accuracy under FGSM attack increased roughly two times compared to the baseline model. However, against PGD attacks, the model trained with FSGM examples has been shown inferior compared to other types. When being attacked by PGD with 7 iterations, the model trained with FSGM images only has an accuracy of 24.67%, less than half of the corresponding accuracy of the model trained with PGD images (58.24%). This inferiority became even more conspicuous when we tested the models' performance under PGD attack with 20 iterations. The model trained with FSGM examples performed with a mere 10.37% while training with PGD examples returned a decent 38.89% accuracy. This difference could be attributed to the PGD attack being much more sophisticated than FGSM; therefore, training with FGSM images does not provide sufficient defense against PGD. Training using PGD inputs, on the other hand, seems to grant the model the best defense as its accuracy is the highest against all types of adversarial attacks. Nevertheless, training with PGD images is more time-consuming than with FSGM as a trade-off for more defense. Despite performing better against adversarial attacks, adversarial training comes at the expense of the model's accuracy with clean images. The results show that there is a trade-off of around 20% accuracy on normal inputs when applying adversarial training. To partly alleviate this trade-off, we performed interpolated adversarial training (IAT) on the model. The model applied with IAT returned an accuracy of about 5% higher than other types of adversarial training, while also maintaining decent performance against FGSM and PGD attacks. This adversarial training method even enables the model to outperform its FGSM counterpart against PGD attacks, with doubled accuracy.

V. Conclusion

In conclusion, throughout this project, we attempted to increase a deep learning model's robustness against adversarial attacks, aiming to compel the model to learn robust features. The results show initial success: The model's accuracy when attacked by adversarial examples increases remarkably compared to the baseline model. We also discovered that training with stronger adversarial examples helps the model gain stronger defense with the cost of longer training time. However, adversarial training reduces the model's performance on clean inputs as a trade-off for robustness.

References

[1] Animal Image Dataset (90 different animals). (2022, July 17). https://www.kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals

[2] Bai, T. (2021, February 2). Recent advances in adversarial training for adversarial robustness. arXiv.org. https://arxiv.org/abs/2102.01356

[3] Kurakin, A. (2016, November 4). Adversarial machine learning at scale. arXiv.org. https://arxiv.org/abs/1611.01236

[4] Lamb, A., Verma, V., Kawaguchi, K., Matyasko, A., Khosla, S., Kannala, J., & Bengio, Y. (2022). Interpolated Adversarial Training: Achieving robust neural networks without sacrificing too much accuracy. Neural Networks, 154, 218–233. https://doi.org/10.1016/j.neunet.2022.07.012

[5] Li, Y., Cheng, M., Hsieh, C., & Lee, T. C. (2022). A review of Adversarial Attack and Defense for classification methods. The American Statistician, 76(4), 329–345. https://doi.org/10.1080/00031305.2021.2006781

[6] Madry, A. (2017, June 19). Towards deep learning models resistant to adversarial attacks. arXiv.org. https://arxiv.org/abs/1706.0608

[7] Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond Empirical Risk Minimization. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1710.09412