Noise in the Network: Comparing Defense Methods for Adversarial Examples in Neural Networks

Yeojun Han
yeojunh@student.ubc.ca

Abstract

This paper explores the vulnerability of Convolutional Neural Networks (CNNs) and Deep Neural Networks (DNNs) to adversarial attacks using CIFAR-10 images. It evaluates the performance of adversarial training methods, revealing that while the original ResNet50 model performs best on unaltered data, it struggles with adversarial attacks. Ensemble adversarial training, which uses adversarial examples designed against other pre-trained models, shows improved robustness but slightly lower performance on standard classification tasks. Future work aims to expand the range of adversarial examples, explore different image classification models, and incorporate a wider array of defense mechanisms. The findings show an improved robustness of defense mechanisms even with small underfit models and this study overall will inform strategies to better defend DNNs from adversarial attacks.

1 Introduction

The rise in popularity of neural network models has sparked interest in their ability to replicate human behaviour. A prime example is the use of Convolutional Neural Networks (CNNs) and Deep Neural Networks (DNN) models in computer vision for tasks such as object recognition and identification. These models classify elements in an image into specific categories, such as objects, scenery, or even species of animals.

However, classification models like object identification, which require extensive training datasets, can inadvertently introduce bias. This bias stems from the challenge of encapsulating the world's complexity within a dataset and can potentially be amplified (Sutherland, 2024). In contrast, adversarial attacks (Szegedy et al. 2014; Papernot, Faghri, et al. 2018; Goodfellow, Shlens, and Szegedy 2015; Papernot, McDaniel, et al. 2015; Tramèr, Kurakin, et al. 2020; Guo et al. 2018; Sutherland 2024) involve the creation of artificially modified datasets. These datasets, visually indistinguishable from unaltered images, can lead various models to misclassify the same example, sometimes even producing identical outputs across different models (Goodfellow, Shlens, and Szegedy 2015). Notably, even state-of-the-art neural networks, known for their generalisation capabilities, are susceptible to such attacks (Szegedy et al. 2014). The deliberate production of adversarial outputs through these attacks can have serious implications, including autonomous car accidents, content filter circumvention, or manipulation of biometric access (Papernot, McDaniel, et al. 2015).

Despite numerous attempts to develop methods to counter adversarial examples, no model has successfully maintained state-of-the-art accuracy on both clean and adversarial inputs (Goodfellow, Shlens, and Szegedy 2015). However, these attempts have resulted in a series of methods that exhibit greater robustness against adversarial examples compared to original models (Goodfellow, Shlens, and Szegedy 2015; Papernot, Faghri, et al. 2018; Papernot, McDaniel, et al. 2015; Guo et al. 2018; Tramèr, Kurakin, et al. 2020).

In this project, we aim to re-implement and evaluate the current state-of-the-art adversarial training methods against a set of adversarial examples. Our goal is to assess the performance of each method

Course project for UBC CPSC 440/550, April 2024.

and identify patterns indicating their robustness and vulnerability. Specifically, we plan to summarise the strengths and weaknesses of each method using a controlled set of adversarial and clean images, thereby gaining a deeper understanding of the nuances of each method of generating adversarial examples. By understanding the performance of the two adversarial training defense mechanisms, we believe that we can better defend DNNs from such attacks.

2 Related Work

2.1 Causes of Adversarial Examples

Neural networks, while powerful, are often complex and difficult to interpret. Szegedy et al. (2014) discovered that adversarial examples are not a consequence of overfitting or an overfit model. Rather, they are a byproduct of linearity in high-dimensional spaces (Goodfellow, Shlens, and Szegedy 2015). This vulnerability of neural networks to seemingly insignificant yet impactful attacks suggests that these classifiers might not be learning the underlying concepts in the same way humans do. Instead, they might be learning patterns from naturally occurring data and subsequently failing when confronted with artificial examples that seldom appear in the data distribution (Goodfellow, Shlens, and Szegedy 2015).

2.2 Creating Adversarial Examples

There are numerous methods for creating adversarial examples, many of which are readily available in libraries such as CleverHans (Papernot, Faghri, et al. 2018). The Fast Gradient Sign Method (FGSM) is one such method that strikes a balance between the likelihood of misclassification and human detection (Goodfellow, Shlens, and Szegedy 2015; Papernot, Faghri, et al. 2018). The Elastic Net Method (EAD) employs elastic-net regularization and builds upon a previous model, the Carlini-Wagner Attack, introduced in 2016 (Papernot, Faghri, et al. 2018). The Jacobian-based Saliency Map Approach (JSMA) calculates the Jacobian adversarial saliency map at each iteration and selectively perturbs the features with high scores (Papernot, Faghri, et al. 2018).

2.3 Defense Mechanisms

Numerous methods exist to defend against adversarial attacks, though none have yet achieved performance on par with state-of-the-art models on clean data. One such method, adversarial training, incorporates adversarial examples during the training process (Papernot, Faghri, et al. 2018). Another approach involves input transformations, which alter the image to counteract tampering. These transformations can include image cropping and rescaling, bit-depth reduction, JPEG compression, total variance minimization, and image quilting (Guo et al. 2018). Neural Network Distillation employs distillation to transfer knowledge from a larger Deep Neural Network (DNN) to a smaller one, enhancing both generalizability and robustness (Papernot, McDaniel, et al. 2015). Lastly, ensemble adversarial training alternates between pre-trained models for each batch (Tramèr, Kurakin, et al. 2020).

3 Experiments

3.1 Dataset Description

There are two distinct stages to our dataset: the control and the adversarial dataset. In this section, we will describe the original dataset that represents images that are found in "nature" as the "clean dataset", and use the term "adversarial dataset" for the dataset specifically designed to break the image classification network (Papernot, McDaniel, et al. 2015).

3.1.1 Clean Dataset

The control dataset is the dataset that represents the set of images that a typical image classification model should correctly label with high accuracy. For this dataset, we used CIFAR-10 (Krizhevsky 2009) as it was the most fitting for an image classification problem as identified the original paper

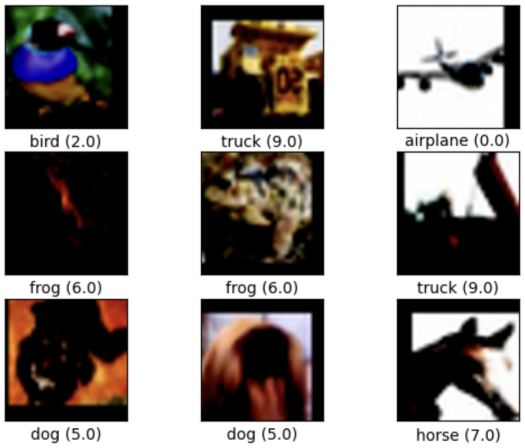

Figure 1: Fast Gradient Sign Method (FGSM) with epsilon 0.001

(Goodfellow, Shlens, and Szegedy 2015). We split the data into train and test set as defined by the original dataset. This data will be used to train the control model, ResNet50.

3.1.2 Adversarial Examples

Second, the adversarial dataset is the set of images that have been modified with one or more of the Fast Gradient Sign Method (FGSM), the Elastic Net Method (EAD), the Carlini Wagner Attack, the Jacobian-based Saliency Map Approach (JSMA) and DeepFool to produce an adversarial classification when given to an image classification model. Each model uses a different approach for creating adversarial examples and we believe that using various methods would help assess the performance of each defense methods from diverse angles. We created adversarial examples using the CleverHans library as implemented by Papernot, Faghri, et al. (2018). The training adversarial dataset will be used to train the defense model for adversarial training and test adversarial dataset will be used for the final models to analyze the model performance.

To create adversarial examples with FGSM, we used the control ResNet50 as the target model and infinity norm. As a result, we generated $128 \times 4 = 512$ untargeted adversarial examples from one training batch of size 128 and five epsilon values: 0.3, 0.07, 0.01, 0.001. A test adversarial dataset is also created for testing with a clean test CIFAR-10 dataset. As seen in Figure 1, the adversarial examples are essentially indistinguishable from a standard CIFAR10 image with a small epsilon. However, with a large epsilon like 0.3, there were significant artifacts that made it obvious that the images have been tampered.

For the Carlini-Wagner Attack, we used the CleverHans library and constructed an attack against the ResNet50, train images and train labels (Papernot, Faghri, et al. 2018; Carlini and Wagner 2016). While this method does not require a hyperparameter, the resulting images seem indistinguishable to the eye, as seen in Figure 2.

Lastly, the Sparse L1 Gradient Method is another adversarial attack method included in the Clever-Hans library (Tramèr and Boneh 2019; Papernot, Faghri, et al. 2018). Similar to the Carlini-Wagner Attack, the Sparse L1 Gradient Method does not require hyperparameters. As seen in Figure 3, the images from this method were the most natural, despite all of the other methods have been difficult to pinpoint as tampered.

3.2 Defense Methods

Choosing a control model for this project was tricky; it needs to be well-known in practice to lead to a meaningful contribution, yet it needs to be feasible with the resource constraints presented by the compute units provided by our only access to a GPU via Google Colab Pro. ResNet was a reasonable image classification model taught in various undergraduate machine learning classes, has a high enough classification accuracy, and most importantly, it is impacted by adversarial image attacks. We will use ResNet50 as the main model for this project including the control and adversarial training.

Figure 2: Carlini-Wagner Attack

Figure 3: Sparse L1 Gradient Method

In a previous study, Osawa et al. (2018) attained an accuracy of 75.1 ± 0.09 % by training a ResNet50 with batch size 4096 and 35 epochs. Given the scope of this project and the compute resource constraints, our goal was to train a ResNet to achieve minimum 75% accuracy. Cross-validation of our ResNet implementation suggested that a batch size of 128 and 64 epochs yield the best performing model of 78.9% accuracy.

There were several approaches to defending our models from adversarial examples, some of which involved directly integrating adversarial examples as a part of model training and others modifying the inputs to the model without altering the model.

Adversarial training is one of the most well known and effective method against adversarial examples (Papernot, Faghri, et al. 2018). It incorporates adversarial examples during training, meaning that the training data should compose of clean images as well as adversarial images against the model. Since adversarial examples are typically generated against a target model, so we used the adversarial training dataset designed against a ResNet50 model.

Ensemble adversarial training is similar to adversarial training in a sense that the training data includes adversarial examples. However, ensemble adversarial training uses adversarial examples designed against other pre-trained models (Tramèr, Kurakin, et al. 2020). This method utilises the property of adversarial examples to transfer between models, which increases model robustness against black-box attacks where an attack is generated without access to a model (Tramèr, Kurakin, et al. 2020; Kurakin, Goodfellow, and Bengio 2016).

Figure 4: Test score of each model against each dataset in percentage

Model, Attack	CIFAR10	FGSM	Carlini-Wagner Attack	Sparse L1 Gradient
ResNet50	78.91	53.90625	38.671875	25.0
Ensemble Adversarial Training	75.16	58.59375	50.78125	67.3828125

4 Discussion

4.1 ResNet50

The original ResNet50 model has the best performance against unaltered CIFAR10 data with 78.91% accuracy. However, as expected, it has the weakest performance against adversarial attacks. Particularly, its significantly low performance against the Sparse L1 Gradient demonstrates the devastating impacts of a well-designed attack against an image classification network (Tramèr and Boneh 2019).

4.2 Ensemble Adversarial Training

Ensemble adversarial training improves upon the black-box attacks from which adversarial training suffers (Tramèr, Kurakin, et al. 2020). Due to its nature to be trained on a dataset that includes adversarial attacks based off of a different model, the model gains robustness. The model used in this project was a ResNet50 model trained on a dataset that includes adversarial examples crafted against a pretrained AlexNet model.

In terms of performance, ensemble adversarial training model has a significantly higher accuracy in the Sparse L1 Gradient attack (67.38%) compared to the naive ResNet50 model. However, it suffers from slightly lower performance in standard classification of the original CIFAR10 dataset (75.16%)compared to the naive ResNet50 (78.91%). Its performance against FGSM and the Carlini-Wagner attack is underwhelming at 58.59% and 50.78%, respectively.

4.3 Future Work

Using incredibly limited resources such as that of this project is both an advantage and a disadvantage. With such limited resources, we are constrained to smaller models. Despite this constraint, we trained and fitted three wo different models: ResNet50 and ResNet50 with ensemble training from AlexNet. As an unexpected side effect of these constraints, we showed that ensemble adversarial training is effective even when trained with batch size 128 and epoch 64 on CIFAR10 dataset. On the other hand, due to the lack of thorough training on our models, we are unable to observe the finer details and quirks of each adversarial attack on the ResNet50 and the ensemble adversarial training model. It is difficult to determine whether the low test accuracy is due to a model's inability to correctly classify given an adversarial attack or simply due to underfitting.

Given more time, we would like to improve the model selection and expand the range of adversarial examples to the Elastic Net Method, the Jacobian-based Saliency Map Approach and the DeepFool method (Papernot, Faghri, et al. 2018; Moosavi-Dezfooli, Fawzi, and Frossard 2015). We would also like to try various different image classification models beyond ResNet50 and AlexNet and carefully fine-tune the models to improve the training accuracy. Most importantly, we would like to incorporate a wider array of defense mechanism, including input transformations and distillation (Guo et al. 2018; Papernot, McDaniel, et al. 2015).

A Supplementary Material

The GitHub repository to this project can be found here:
https://github.com/yeojunh/CPSC440-project/

References

Carlini, Nicholas and David A. Wagner (2016). "Towards Evaluating the Robustness of Neural Networks." *CoRR* abs/1608.04644. arXiv: 1608.04644. URL: http://arxiv.org/abs/1608.04644.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy (2015). "Explaining and Harnessing Adversarial Examples." arXiv: 1412.6572 [stat.ML].

Guo, Chuan, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten (2018). "Countering Adversarial Images using Input Transformations." arXiv: 1711.00117 [cs.CV].

Krizhevsky, Alex (2009). *Learning multiple layers of features from tiny images*. Tech. rep.

Kurakin, Alexey, Ian J. Goodfellow, and Samy Bengio (2016). "Adversarial examples in the physical world." *CoRR* abs/1607.02533. arXiv: 1607.02533. URL: http://arxiv.org/abs/1607.02533.

Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard (2015). "DeepFool: a simple and accurate method to fool deep neural networks." *CoRR* abs/1511.04599. arXiv: 1511.04599. URL: http://arxiv.org/abs/1511.04599.

Osawa, Kazuki, Yohei Tsuji, Yuichiro Ueno, Akira Naruse, Rio Yokota, and Satoshi Matsuoka (2018). "Second-order Optimization Method for Large Mini-batch: Training ResNet-50 on ImageNet in 35 Epochs." *CoRR* abs/1811.12019. arXiv: 1811.12019. URL: http://arxiv.org/abs/1811.12019.

Papernot, Nicolas, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, Rujun Long, and Patrick McDaniel (2018). "Technical Report on the CleverHans v2.1.0 Adversarial Examples Library." arXiv: 1610.00768 [cs.LG].

Papernot, Nicolas, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami (2015). "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks." *CoRR* abs/1511.04508. arXiv: 1511.04508. URL: http://arxiv.org/abs/1511.04508.

Sutherland, Danica J. (2024). *What do we learn?* URL: https://www.cs.ubc.ca/~dsuth/440/23w2/slides/11-what-learn.pdf.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus (2014). "Intriguing properties of neural networks." arXiv: 1312.6199 [cs.CV].

Tramèr, Florian and Dan Boneh (2019). "Adversarial Training and Robustness for Multiple Perturbations." *CoRR* abs/1904.13000. arXiv: 1904.13000. URL: http://arxiv.org/abs/1904.13000.

Tramèr, Florian, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel (2020). "Ensemble Adversarial Training: Attacks and Defenses." arXiv: 1705.07204 [stat.ML].