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

Yeojun Han
Student Number: 88591664
yeojunh@student.ubc.ca

March 29th 2024

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., 2013; Goodfellow et al., 2014; Papernot et al., 2015; Papernot et al., 2016; Tramèr et al., 2017; Guo et al., 2017; 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 et al., 2014). Notably, even state-of-the-art neural networks, known for their generalisation capabilities, are susceptible to such attacks (Szedegy et al., 2013). 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 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 et al., 2014). However, these attempts have resulted in a series of methods that exhibit greater robustness against adversarial examples compared to original models (Goodfellow et al., 2014; Papernot et al., 2015; Papernot et al., 2016; Guo et al., 2017; Tramèr et al., 2017).

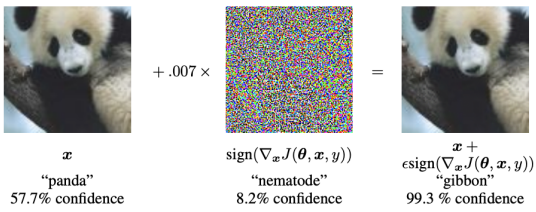

Figure 1: an adversarial example that outputs a visually indistinguishable image that leads to a misclassification with high confidence.

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 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 model.

2 Related Work

2.1 Causes of Adversarial Examples

Neural networks, while powerful, are often complex and difficult to interpret. Szegedy et al. (2013) 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 et al., 2014). 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 et al., 2014).

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 et al., 2016). The Fast Gradient Sign Method (FGSM) is one such method that strikes a balance between the likelihood of misclassification and human detection (Goodfellow et al., 2014; Papernot et al., 2016). The Elastic Net Method (EAD) employs elastic-net regularization and builds upon a previous model, the Carlini-Wagner Attack, introduced in 2016 (Papernot et al., 2016). 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 et al., 2016).

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 et al., 2016). 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., 2017). 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 et al., 2015). Lastly, ensemble adversarial training alternates between pre-trained models for each batch (Tramèr et al., 2017).

3 Implementation

3.1 Adversarial Example Implementation

We will generate adversarial examples using the CIFAR-10 dataset and the Fast Gradient Sign Method (FGSM) (Papernot et al., 2016).

3.2 Defence Method Implementation

We plan to implement four defence methods as outlined in the previous section: adversarial training, input transformations, distillation, and ensemble adversarial training. We hypothesise that some of these methods will outperform others. By exploring these methods of varying complexity, we aim to gain insights into whether more complex models tend to exhibit superior performance against adversarial attacks.

4 Experimental Setup

4.1 Tasks, Datasets, and Evaluation

The objective of this project is to ascertain whether certain defence mechanisms are more effective than others and, if so, how they surpass their counterparts. Consequently, the task involves performing image classification on the baseline model and the models utilizing defence mechanisms, and analyzing the quality of each model's classification. We will use the CIFAR-10 dataset, modified with the FGSM. Each model's performance will be evaluated based on the accuracy of the classification and the confidence level. For instance, if a model makes an incorrect prediction, we expect the confidence level to be low.

4.1.1 Baselines

Our baseline will be a CNN image classification model with ResNet. We will evaluate each model's performance on both adversarial examples and untouched, clean data to assess their baseline performance and generalizability.

4.1.2 Expected Results

We anticipate that the benchmark CNN model will exhibit the lowest accuracy with adversarial examples but the highest accuracy with clean data. Conversely, we expect the defence methods to slightly surpass the benchmark model in terms of adversarial example accuracy, albeit with a marginally lower clean data accuracy. We also foresee that each model will have its own advantages and disadvantages, which will factor in based on the trade-offs prioritised by the user. These insights will inform future research in adversarial training regarding the strengths, weaknesses and possible limitations of each model.

5 References

Biggio, B. and Roli, F. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84:317–331, 2018. doi: 10.1016/j.patcog.2018.07.023.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014, December 20). Explaining and harnessing adversarial examples. arXiv.org. https://arxiv.org/abs/1412.6572

Guo, C., Rana, M., Cisse, M., & Laurens, V. D. M. (2017, October 31). Countering Adversarial Images using Input Transformations. arXiv.org. https://arxiv.org/abs/1711.00117

Papernot, N., Faghri, F., Carlini, N., Goodfellow, I., Feinman, R., Kurakin, A., Xie, C., Sharma, Y., Brown, T., Roy, A., Matyasko, A., Behzadan, V., Hambardzumyan, K., Zhang, Z., Juang, Y., Li, Z., Sheatsley, R., Garg, A., Uesato, J., . . . McDaniel, P. (2016, October 3). Technical report on the CleverHans v2.1.0 Adversarial Examples Library. arXiv.org. https://arxiv.org/abs/1610.00768

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2015, November 14). Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. arXiv.org. https://arxiv.org/abs/1511.04508

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013, December 21). Intriguing properties of neural networks. arXiv.org. https://arxiv.org/abs/1312.6199

eTramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017, May 19). Ensemble Adversarial Training: attacks and defenses. arXiv.org. https://arxiv.org/abs/1705.07204