Jailbreaking Deep Models
Image Classification

Alex Gonzalez, Chloe Kim, Richard Zhong

New York University, New York, USA

https://github.com/Adgonzalez2018/DL-final-project

Abstract

This project aims to degrade the performance of production grade, publicly posted models by launching effective adversarial attacks and exploring the subtle perturbations. Specifically, the ResNet-34 model that is trained to classify the ImageNet-1K dataset will be attacked. First, a common algorithm called Fast Gradient Sign Method (FGSM) will be implemented for each image in the test dataset for $\varepsilon = 0.02$. Then to further improve attacks, iterative FGSM/PGD attacks were implemented. We explored the impact of different hyperparameter configurations– including epsilon, alpha, and different iterations to achieve fast and better attacks. Lastly, we tried patch attacks where instead of perturbing the whole test image, a random patch of size 32x32 is attacked. Evaluating another model ResNet-50 with the three adversarial dataset led us to achieve a classification accuracy of __**%**. These findings contribute to the broader effort of mitigating transferability and ensuring security so classifiers are less brittle.

Introduction

In recent years, deep neural networks, particularly in image classification, have demonstrated remarkable successes. However, deep classifiers are susceptible to adversarial attacks. The attacks involve the introduction of carefully crafted perturbations to the input data that cause a model to misclassify well-recognized images. With deep models increasingly deployed in real-world application, the integrity of models raises significant concerns.

The project explores the phenomenon of jailbreaking deep models and their brittleness - strategically degrading the performance of production-grade, publicly available image classifiers through adversarial attacks. Attack strategies that are both effective and subtle are explored. Specifically, we focus on the two common norms: L_∞, which restricts the maximum allowable change to any individual pixel attacks, and the L_0 norm, which limits the number of pixels that may be altered.

By evaluating the impact of such attacks, this project highlights the urgent need for robust defenses and the importance of designing models that are not only accurate, but also resilient. Through this work, we aim to deepen our understanding of adversarial vulnerabilities and contribute toward the broader goal of secure and trustworthy deep models.

Methodology

Pixel-wise Attacks: Fast Gradient Sign Method (FGSM) is a common and simple algorithm for mounting an L_∞ attack. The method performs a single step of gradient ascent and truncates the values of the gradients to at most ε, the scalar attack budget that controls the magnitude of the perturbation. The equation is: $x \leftarrow x + \varepsilon\, sign(\nabla \mathrm{xL})$.

The gradient is normalized via a sign function to constrain the perturbation within an L_∞ norm, effectively restricting the maximum per-pixel change. The hyperparameter is the scalar ε, where larger values yield more effective attacks but degrade the image quality. For images with pixel values of [0-255], the value must be no greater than 0.02 to roughly correspond to changing each pixel value in the raw image by at most +/-1.

Advantage of FGSM is that it is fast and simple, as it only requires one gradient step. However, the perturbations are often large and easily detectable.

Improved Attacks: Projected Gradient Descent (PGD) is an iterative adversarial attack that improves upon FGSM by applying multiple small perturbation steps, each followed by a projection back onto an ε around the original input. This ensures the adversarial example stays within a bounded norm while allowing more precise manipulation of the model. It is considered one of the strongest first-order attacks in the white-box setting.

Its adjustable hyperparameters are epsilon ε, alpha α, and iterations. Increasing ε will make attacks stronger, but it risks the quality of the image. Increasing alpha will lead to fast convergence, but can overshoot and produce less

precise results. Increasing iterations improves attack success at the cost of computation. The project experimented with ε = 0.02, α = 0.005, and iters = 10. Advantage of PGD is that the attack is stronger due to iterative refinement,but it is slower because it requires multiple forward and backward passes.

Patch Attacks: Unlike other attacks that make subtle changes across the entire image, patch attacks locally alter the image in a set, defined size. This project used a patch size of 32x32. Hyperparameter ε can be now increased to more than 0.2, and the values chosen were ε = 0.5, α = 0.02, and iters = 20. Its advantage is that it produces minimal perturbations despite it being computationally expensive and sensitive to regularization constant. It is also robust to transformations such as scaling, rotation, and occlusion.

Results/Discussion

Resnet 34	Baseline	FGSM	PGD	Patch
Top 1	76.00%	26.40%	4.00%	37.80%
Top 5	94.20%	50.60%	18.80%	57.20%
Size of perturbations (ε)	NA	0.02	0.02	0.5
Training times (secs)	NA	3.54	22.78	47.78

Table 1
Results of Attacks on ResNet-34

Figure 1
Visualization of FGSM Attack on ResNet-34

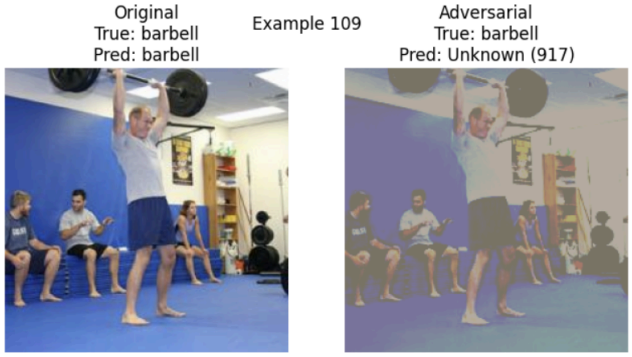

Figure 2
Visualization of PGD Attack on ResNet-34

I

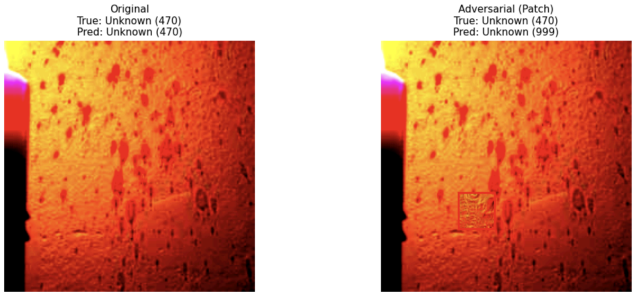

Figure 3
Visualization of PGD Patch Attack on ResNet-34

Resnet 50	Baseline	FGSM	PGD	Patch
Top 1	80.20%	38.80%	36.20%	42.20%
Top 5	94.60%	56.40%	54.60%	61.80%
Size of perturbations (ε)	NA	0.02	0.02	0.5

Table 2
Results of Attacks on ResNet-50 with Perturbed Examples developed from ResNet-34

The results demonstrate that adversarial attacks significantly degrade the performance of both ResNet-34 and ResNet-50 models, highlighting the vulnerability of deep neural networks to carefully crafted perturbations. The classification accuracy of ResNet-34 drops from 76% (baseline) to as low as 4% under PGD attacks, while ResNet-50's baseline accuracy of 80% decreases to 36.2% under the same attack method.

Comparing different attack strategies, FGSM and PGD consistently outperform patch attacks in terms of

adversarial success, with PGD generally achieving higher top-1 accuracies due to its iterative refinement process. However, PGD's computational cost is notably higher than FGSM, which might limit its deployment in scenarios requiring rapid inference times. Patch attacks, while effective, introduce larger perturbations (0.5 compared to 0.02 for PGD), suggesting a trade-off between attack strength and image quality.

The results also indicate that ResNet-34 is generally more susceptible to adversarial attacks than ResNet-50. This difference may be attributed to architectural variations—ResNet-34's depth versus Resnet-50's additional layers could influence how perturbations propagate through the network, affecting classification accuracy. Furthermore, the choice of attack method influences which part of the network is targeted, as seen in the visualization figures, where FGSM primarily affects specific regions, while PGD and patch attacks may alter different areas due to their localized or iterative nature.

Conclusion

This project presents a comprehensive investigation into the effectiveness of various adversarial attack methods on two prominent deep neural networks, ResNet-34 and ResNet-50. The experiments reveal that all three types of attacks—FGSM, PGD, and patch attacks—significantly degrade model performance, with PGD achieving the most pronounced effect on ResNet-34, while ResNet-50 showed a more moderate decline with the same perturbed examples.

The study highlights differences in vulnerability among the two models, suggesting that architectural variations between ResNet-34 and ResNet-50 play a crucial role in how perturbations affect classification. This finding underscores the potential impact of model architecture on adversarial robustness.

While FGSM offers computational efficiency with its fast processing, it results in larger and more detectable perturbations compared to PGD. In contrast, PGD provides refined attacks but at the cost of higher computation time. Patch attacks, though effective, introduce larger perturbations and are computationally expensive, yet they maintain some image quality by altering only a small section of the input.

The results emphasize the urgent need for robust defense mechanisms in deep learning models to counter adversarial attacks. The findings also contribute to the broader goal of mitigating transferability and ensuring model integrity, particularly in real-world applications where reliability is paramount.

Acknowledgment

We acknowledge the use of OpenAI's ChatGPT for certain sections of the report.

References

Carlini, N., & Wagner, D. (2017). *Towards evaluating the robustness of neural networks*. In *IEEE Symposium on Security and Privacy (SP)* (pp. 39–57). IEEE. https://arxiv.org/abs/1608.04644

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples*. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1412.6572

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). *Towards deep learning models resistant to adversarial attacks*. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1706.06083