CSC2529 Project Proposal

Junbo Huang (1004280422), Jason Tang (1004221326)

1 Introduction

In recent years, machine learning (ML) models have improved dramatically in their ability to perform visual tasks on naturally occurring images. However, most ML systems were designed with minimal consideration of any potential exploits. One class of these exploits focuses specifically on attacking a model's integrity such that it outputs incorrect predictions, potentially in a manner benefiting the attacker. Some potential exploits include: falsifying cheques, bypassing facial recognition systems, and altering road signs.

These attacks generally work by adding small perturbations to an input image which, while indistinguishable to the human eye, generates unexpected model outputs. Moreover, there are even ways to steal black-box models and exploit the transferability of adversarial examples to attack them without access to model architecture or their training data [1]. In our project, we will propose and analyze the robustness of a novel ensemble-based defense system utilizing different input sizes in the white-box setting, where attackers have complete access to our model.

2 Related Work

2.1 Adversarial Attacks

In the white box setting, attacks can be generated by solving a constrained optimization problem, which minimizes, or even bounds, the L_0, L_2, and/or L_∞ norm between the input and adversarial images, while also requiring that the model is actually tricked by the adversarial image. The minimization of image distance encourages imperceptible changes, such that human analysis of adversarial images does not raise concerns.

For standard convolutional neural networks (CNNs), the Fast Gradient Sign Method (FGSM) presents a simple and efficient attack that exploits the existing backpropagation architecture in modern neural networks to descend along the gradient of the loss function for the adversarial target class with respect to the input image [2]. This descent step can also be repeated iteratively, and with random nearby initializations (PGD) [4] to create more powerful attacks. Carlini and Wagner (CW) present a strong attack which breaks several previously effective defenses by directly optimizing the constrained optimization problem using a margin loss [5]. Lastly, the Skip Gradient Method (SGM) is a recent technique which exploits skip connections in neural networks with residual connections to pass gradients through the model more directly [6].

2.2 Ensemble Defenses

To address this, many defenses have been proposed and defeated in an ongoing arms race within this field. We found 3 similar ensemble-based methods to our proposed method. The first method learns a diverse ensemble of models using a regularizer term encouraging orthogonality in non-maximal predictions between ensemble members [10]. The second method also aims to improve robustness through diverse ensembles, this time through the usage of varied numerical precisions [8]. Both of these methods were broken in a survey conducted by Tramèr et et. [7] by running PGD until convergence and by approximating the numerically unstable majority vote with an average, respectively. The last similar method is Ensemble Adversarial Training, which utilizes adversarial examples generated using other ensemble members to perform the adversarial training step [3]. This corresponds to the method attackers used to attack vanilla adversarial training [1], which likely leads to the high level of robustness seen with this defense.

2.3 Denoising Defenses

Many different denoising techniques have been proposed as a defensive mechanism against adversarial attacks. Traditional denoising methods such as Gaussian, Bilateral, Nonlocal Means, and Total Variation have all proven useful in removing noise in adversarially perturbed images. Simply applying defensive denoising with TV and NLM can remove major parts of the universal adversarial perturbations in images and improve classification performance [11]. A more recent feature denoising [12] method was designed to incorporate NLM as intermediate blocks in the convolutional network of the classifiers; these denoising blocks can be trained to directly remove noises caused by adversarial perturbations on intermediate features.

Deep learning based denoisers are also very effective as preprocessors. Convolutional neural network architectures such as Denoising Autoencoders (DAE) and U-Net have great capacity for learning and removing adversarial noise. In order to improve robustness, the deep denoising sparse autoencoder (DDSA) [13] method intends to learn a representation extracted from the autoencoder that is robust to adversarial perturbations by adding a sparsity constraint to enforce the extraction of only meaningful and relevant features. However, DAE has a bottleneck structure between the encoder and the decoder, which might hinder the transmission of fine details necessary for high resolution image reconstruction. A denoising U-net (DUNET)

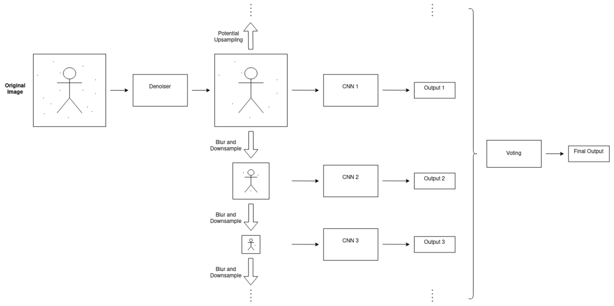

Fig. 1. Illustration of our proposed method.

[14] addresses this problem by adding lateral connections and incorporating residual learning. DUNET also learns the adversarial noise with a loss function guided by high-level representation. The DnCNN denoiser could potentially be a preprocessor to remove adversarial noises, as it also leverages residual learning to effectively estimate noise in the image [15].

3 Proposed Method

We introduce an ensemble-based defense system where each ensemble member receives a different resized version of the input image from a gaussian pyramid (see Fig. 1). We will experiment with replacing the gaussian blur step with a non-differentiable median filter. With additional time, we also intend to potentially explore using bilinear or median-based upsampling to increase image size, as repeated downsampling will likely destroy image information. 1908.00273v2.pdf We also plan on applying a denoiser system before the downscaling process. As seen in recent literature, deep denoisers can be quite effective as a defensive measure. We intend to evaluate the denoising performance of DnCNN compared to no denoising as the baseline. The DnCNN will be trained with a dataset containing adversarial images generated by different attack methods and perturbation levels, which should further improve system robustness. With additional time, we could also compare the denoising performance of the Denoising Autoencoder and U-Net methods.

In terms of combining the resulting ensemble outputs, we intend to explore the following options:

- Uniform Average Outputs: A linear baseline.
- Weighted Average Outputs: Performance scaled weights will likely improve performance on clean inputs, but could lead to targeted attacks on the highest performing models.
- Uniform Majority Vote: Equal votes introduces a non-differentiable step.
- Weighted Majority Vote: Better accuracy on clean inputs, but may be more vulnerable.

4 Experiments and Goals

We will begin with MNIST and CIFAR10 datasets as most works in the literature do. Then, with enough time, we plan on running our system on ImageNet to examine our system's performance on larger images. We also plan on starting with Resnet 34/50 models, and exploring other model options with additional time. For attacks, we plan on using the cleverhans library in evaluating our system against FGSM, PGD, and C&W attacks. We leave the SGM attack as a stretch goal.

For comparison, we plan to run the Ensemble Adversarial Training method and to use a single model vanilla CNN training as a baseline. If we have additional time, we also plan on finding other successful defenses for comparison.

5 Timeline

- Nov 19 - Completed literature review.
- Nov 24 - Complete implementation of our method.
- Nov 28 - Begin experiments.
- Dec 2 - Start preparing Poster and Report.
- Dec 4 - Completed experimentation.
- Dec 8 - Project Presentation and Submission

References

[1] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in Proceedings of the 2017 ACM on Asia conference on computer and communications security, 2017, pp. 506–519.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.

[3] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," arXiv preprint arXiv:1705.07204, 2017.

[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083, 2017.

[5] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in 2017 ieee symposium on security and privacy (sp), 2017, pp. 39–57.

[6] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma, "Skip connections matter: On the transferability of adversarial examples generated with resnets," arXiv preprint arXiv:2002.05990, 2020.

[7] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," Advances in Neural Information Processing Systems, vol. 33, pp. 1633–1645, 2020.

[8] S. Sen, B. Ravindran, and A. Raghunathan, "Empir: Ensembles of mixed precision deep networks for increased robustness against adversarial attacks," arXiv preprint arXiv:2004.10162, 2020.

[9] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in International conference on machine learning, 2018, pp. 274–283.

[10] T. Pang, K. Xu, C. Du, N. Chen, and J. Zhu, "Improving adversarial robustness via promoting ensemble diversity," in International Conference on Machine Learning, 2019, pp. 4970–4979.

[11] Lee, S., Lee, J., and Park, S. (2018). Defensive denoising methods against adversarial attack.

[12] Maaten, L. V. D., Yuille, A. L., and He, K. (2019). Feature Denoising for Improving Adversarial Robustness. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[13] Bakhti, Y., Fezza, S. A., Hamidouche, W., and Deforges, O. (2019). DDSA: A Defense Against Adversarial Attacks Using Deep Denoising Sparse Autoencoder. IEEE Access, 7, 160397–160407.

[14] Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., and Zhu, J. (2018). Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[15] Zuo, W., Chen, Y., Meng, D., and Zhang, L. (2017). Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. IEEE Transactions on Image Processing, 26(7), 3142–3155.