Gaussian Pyramid Ensemble Defense

Jason Tang, Junbo Huang
University of Toronto

Motivation

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 potential exploits. One class of these attacks focuses specifically on tampering with a model's integrity such that it produces erroneous predictions, potentially in a manner benefiting the attacker. Some potential attacks 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, significantly alters model outputs. Moreover, there are even ways to steal black-box models and exploit the transferability of adversarial examples to attack models without access to their architecture, weights, or training data.

In our project, we 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.

Related Work

Adversarial Attacks:

- **FGSM** (Fast Gradient Sign Method): exploits existing backpropagation architecture to descend along the gradient of the loss function for the adversarial target class with respect to the input image. [1]
- **PGD** (Projected Gradient Descent): an iterative method that repeatedly runs FGSM. [2]
- **C&W** (Carlini & Wagner): a stronger iterative method that directly optimizes the constrained problem using a margin loss. [3]

Ensemble Defense Method:

- Ensemble Adversarial Training: performs adversarial training with examples generated from ensemble members, which corresponds to the transferability attack originally used against vanilla adversarial training. [4]

Denoising Methods:

- Defensive denoising using basic TV and NLM methods can remove major parts of the universal adversarial perturbations in images. [5]
- Deep denoising sparse autoencoder (DDSA) learns a representation that is robust to adversarial perturbations by adding a sparsity constraint to enforce the extraction of only meaningful and relevant features. [6]
- Denoising U-net (DUNET) learns the adversarial noise with a loss function guided by high-level representation. [7]

References

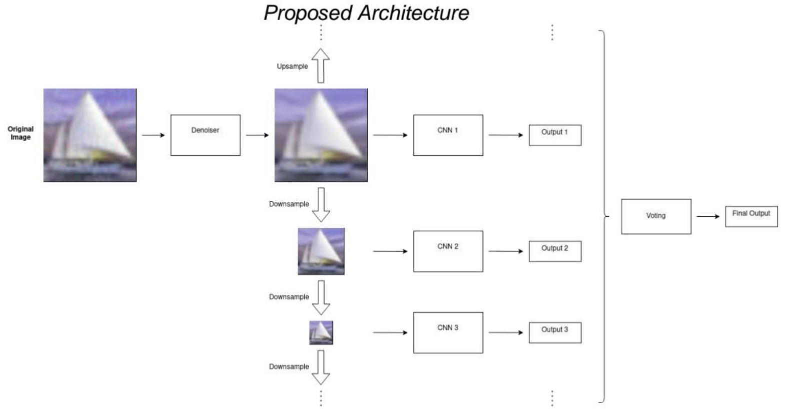
[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.
[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083, 2017.
[3] 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.
[4] 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.
[5] Lee, S., Lee, J., and Park, S. (2018). Defensive denoising methods against adversarial attack.
[6] 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.
[7] 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

Proposed Method

We introduce GPEnsemble, a gaussian pyramid inspired ensemble-based defense system where each ensemble member receives a different resized version of the input image. Each input is then passed through vanilla Resnet18 models trained on their respective input sizes. The resulting outputs are then combined either through a simple linear combination or through a non-differentiable voting system.

We also apply a DnCNN denoiser system before the resizing process, which has seen some success in recent literature. The DnCNN model is trained using adversarial images generated by various attack methods, which should further improve robustness.

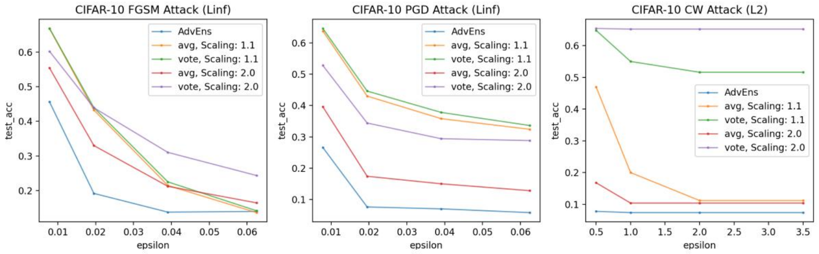
Proposed Architecture



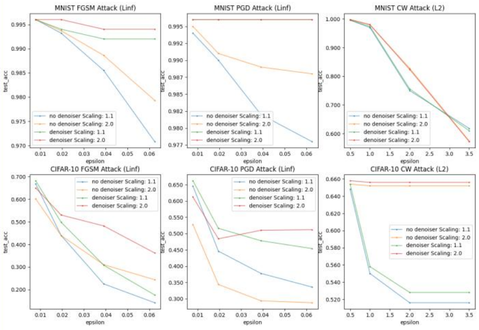
Experimental Results

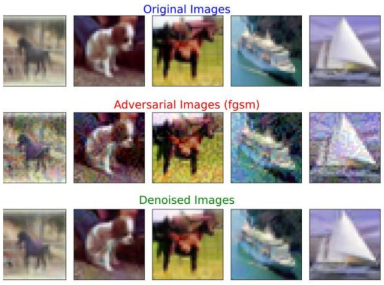
We examine the robustness of our proposed model against an implementation of Ensemble Adversarial Training method (AdvEns) on the commonly used L_∞ and L_2 norms. Specifically, we evaluate on L_∞ limits of [2, 5, 10, 16]/256 per pixel in FGSM and PGD attacks, and on the corresponding L_2 limits of [0.5, 1.0, 2.0, 3.5] in C&W attacks. We also explored using different input scaling factors [2.0, 1.1], as smaller scales allow for more ensemble members with better space efficiency.

Since voting methods have no definable gradient, we generate adversarial examples on a differentiable substitute and transfer attacks to the target model.



As seen, our method is able to outperform the AdvEns model at the cost of running time. We also note that the non-differentiable voting methods are fairly robust to the iterative but less transferable PGD and CW attacks.





PSNR: 30.711dB, SSIM: 0.967 (averaged over 1024 images)

The DnCNN denoiser is trained with adversarial examples and some added gaussian noises for better smoothing and generalization performance. Overall, we observe up to a 21% accuracy increase (12-17% on average) by adding an adversarially trained denoising preprocessor, which provides an additional layer of robustness to our system.