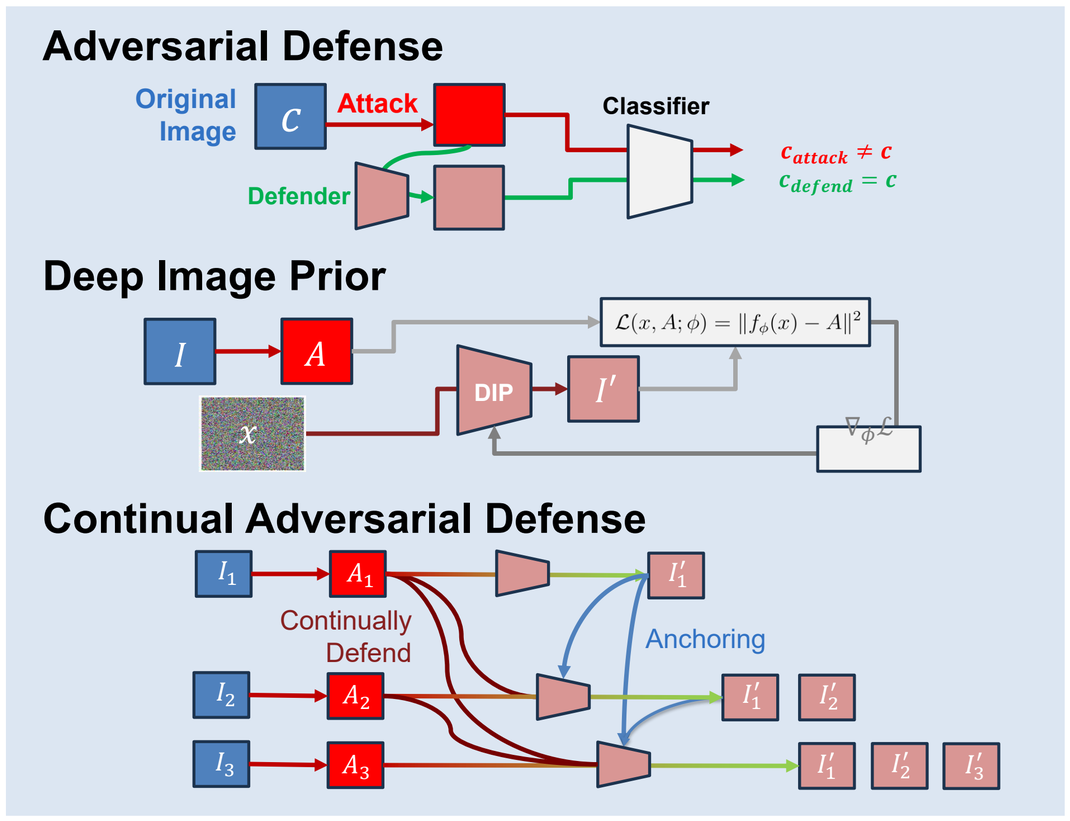
Incremental Deep Image Prior for Continual Adversarial Defense

Author: Dun-Ming Huang

Motivation

In this work, we discuss lightweight methods to defend against a sequence of adversarial attacks. sequence of adversarial attacks. We investigate the usage of Deep Image Prior [1] as a denoiser that removes adversarial noises from a given attack, and prior work has proven it is possible [2]. DIPDefend [2] satisfies all desiderata **but can only correspond to one attacked image**. Therefore, we contribute:

- **Incremental Deep Image Prior**– DIP that retains output for multiple images by training it with continual learning tricks.
- We apply IDIP to defend a classifier from a sequence of adversarial attacks.
- We achieve **continuous adversarial defense** that is equivalently competent to training new defenses per attack.



Related Works

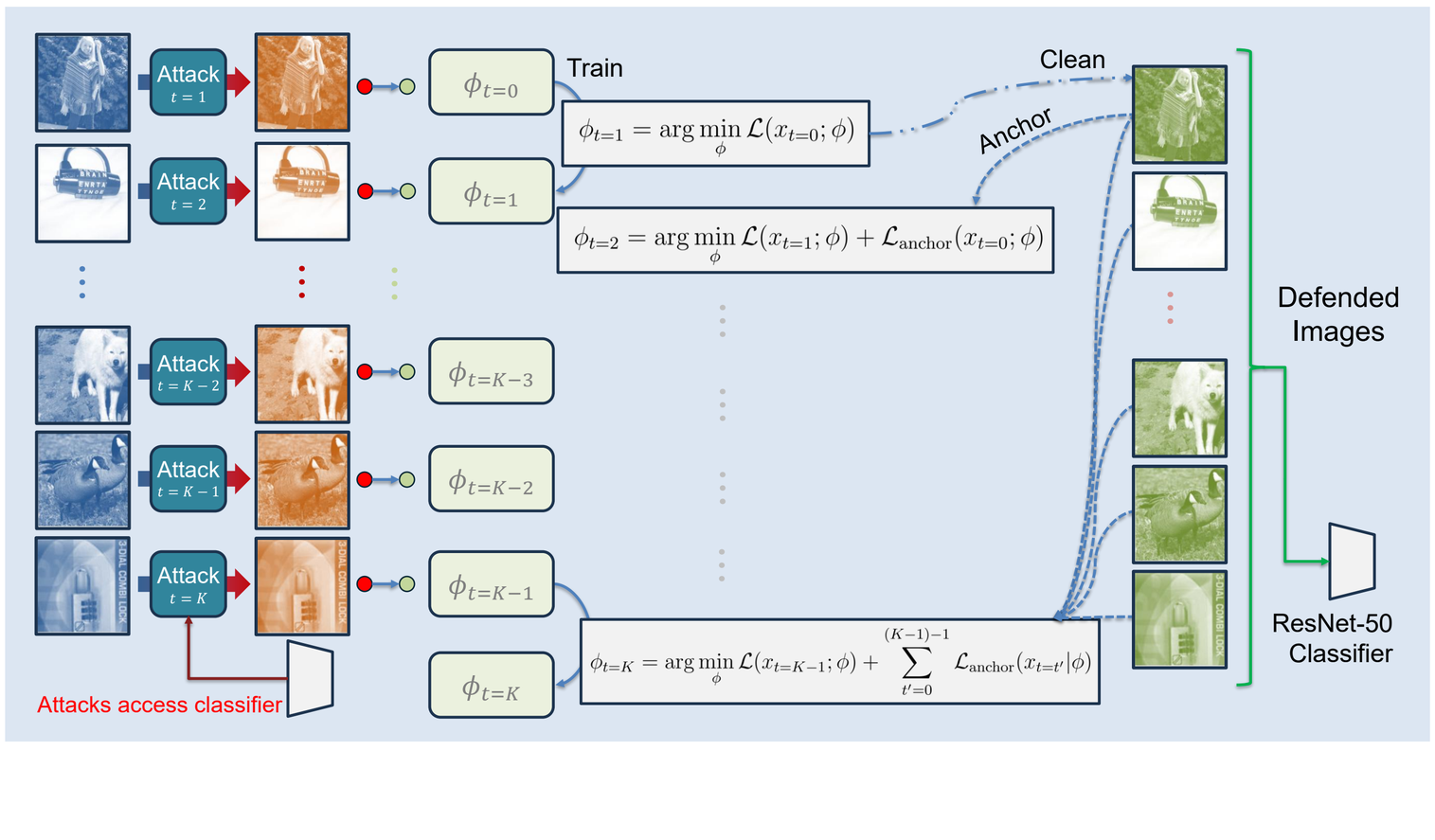
Deep Image Prior [1] constructs a denoised version of its input by the natural bias of U-Nets to find minimal-noise solutions. Provided an objective has a smooth landscape, DIP solves its inverse task well.

DIPDefend. Using DIPs as denoisers for adversarially noised images successfully cleans them for correct model predictions, achieving adversarial defense [2]. Using adaptive PSNR gating avoids recreating adversarial noise.

Methods

Incremental DIP. To retain defenses for past attacks, we train our DIP on a composite objective of (a) defending the current attack, and (b) maintaining the defense output of all past attacks. The term (a) is an MSE [1]. The term (b) can be EWC [3] or the MSE between the current output and the first defense attempt output. We experiment with the MSE of FFT spectrums with exponentially inverse weight on higher-frequency details.

Training Routine. We assign one random noise as the net input for each image. We train each network on (a) the attack of the timestep and (b) all attacks that precede.



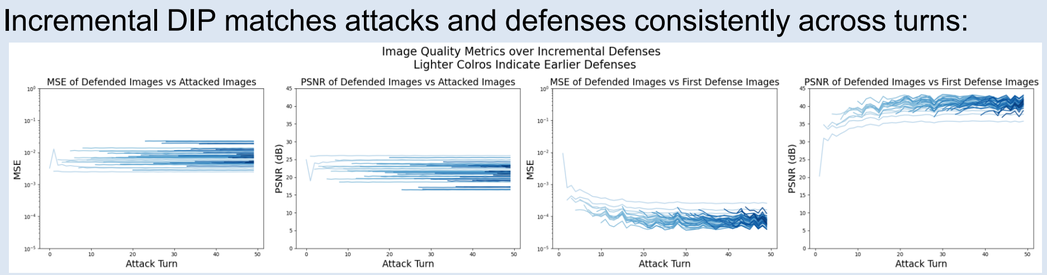
Experiments and Results

We experiment on the ImageNet-small dataset. Using its validation set, we generate adversarial examples with PGD [4] to attack a pretrained ResNet-50 classifier. We answer the following research questions:

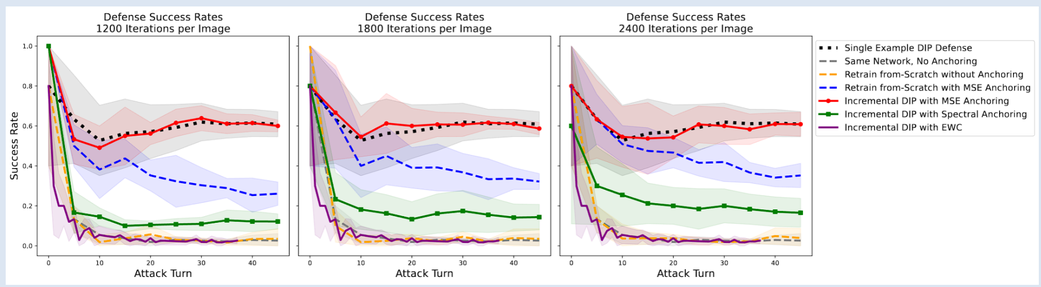
RQ1. Does incremental DIPDefend track each image consistently throughout turns of attacks?

RQ2. Does incremental DIPDefend outperform other naïve approaches or existing continual learning methods?

Incremental DIP matches attacks and defenses consistently across turns:



Amidst an attack success rate of **90%**, IDIP-Defend is equivalently competent to its single-example baseline.



Furthermore, IDIP with MSE Anchoring outperforms the naïve approach of retraining the entire network on all examples seen so far. We conjecture this is because IDIP's training routine naturally preserves the adaptive PSNR filtering in DIPDefend.

However, spectral anchoring does not do well, likely because using "robust features" does not directly correspond to preserving low-frequency details. This suggests DIP's adversarial denoising is more than removing high-frequency noises.

On the right, IDIP's adversarial defenses do not degrade over new attacks.

References. [1] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," International Journal of Computer Vision, vol. 128, no. 7, p. 1867–1888, Mar. 2020 [2] T. Dai, Y. Feng, B. Chen, J. Lu, and S.-T. Xia, "Deep image prior based defense against adversarial examples," Pattern Recognition, vol. 122, p. 108249, 2022 [3] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," Proceedings of the National Academy of Sciences, vol. 114, no. 13, p. 3521–3526, Mar. 2017. [4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu,"Towards deep learning models resistant to adversarial attacks," 2019. [Online].