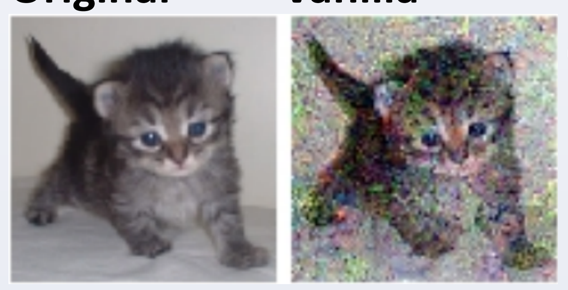
Investigating the Sketchy Effects of Adversarial Training

Simran Kaur[1], Jeremy Cohen[1], Zachary C. Lipton[1]

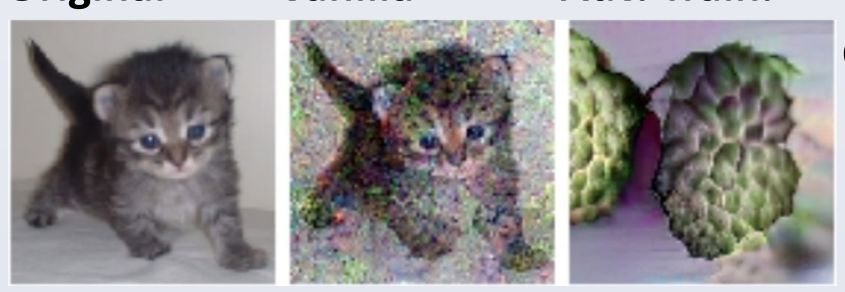
[1]Carnegie Mellon University

Introduction

- **Targeted adversarial attacks** consist of iteratively updating an image by gradient ascent to increase the score of a chosen class.
- Adversarial attacks against vanilla CNNs produce images that appear noisy but are confidently misclassified.



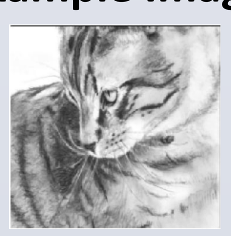
- Santurkar et al [1] showed that adversarially trained neural networks exhibit **perceptually-aligned gradients**: adversarial attacks against these networks produce images that perceptually resemble the target class.



- We investigate how similar phenomena are realized differently between standard and adversarially trained models.
- We trained models on datasets where some features(s) do not vary across images

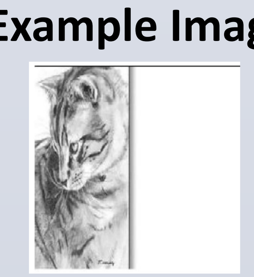
  Restricted ImageNet Sketch Dataset 1: black and white sketches rather than real images [2] [3]

  Example Image

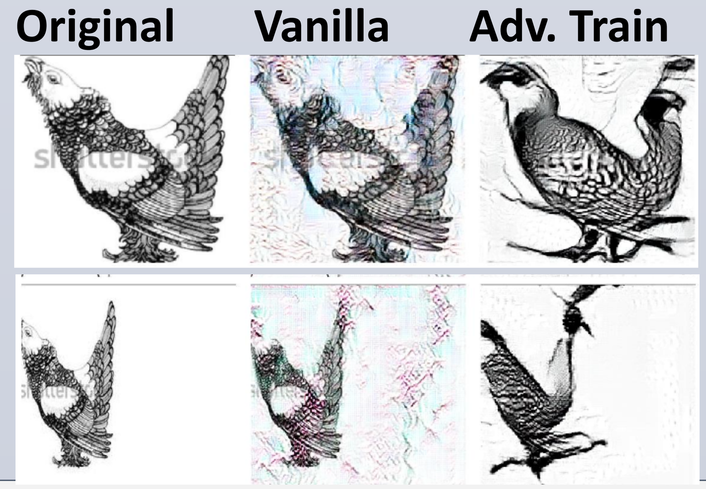


  Restricted ImageNet Sketch Dataset 2: concatenated blank image to each image in Restricted ImageNet Sketch 1 dataset and scaled resulting image to match original dimensions

  Example Image



- We observe that that adversarial attacks against vanilla CNNs fail to maintain the relevant dataset's consistent features(s), whereas these attacks against adversarially trained CNNs maintain these features



Experiments

1) Unconstrained targeted adversarial attacks on Restricted ImageNet Sketch 1 Models

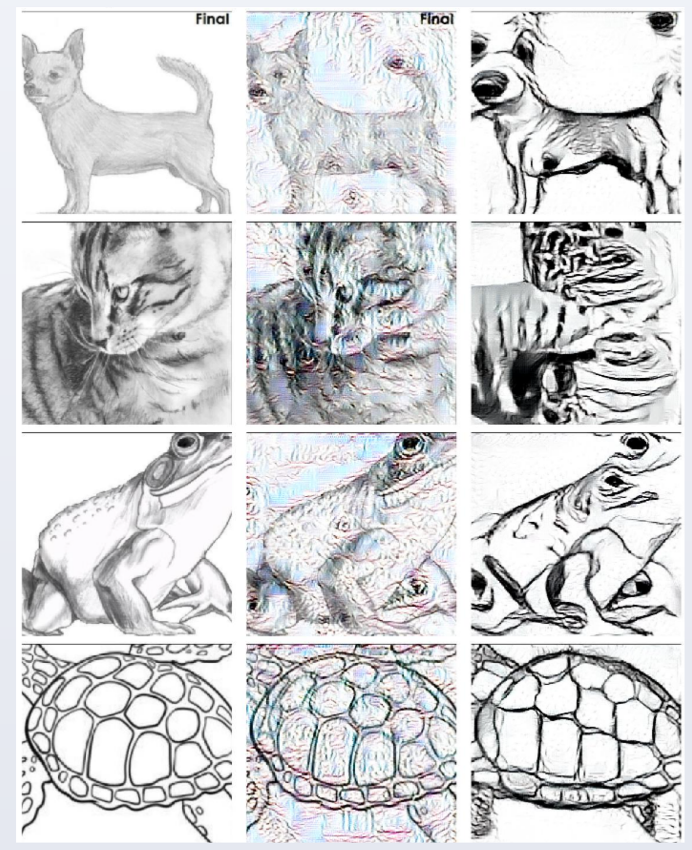


Figure 1: Adversarial attacks against the vanilla CNN produced images that appear noisy with color and are confidently misclassified. However, these attacks against the adversarially trained CNN resemble the target class and maintain the black and white color scheme.

- During training, the adversarial model encounters examples that are created by PGD which are not constrained to just black and white perturbations
- Adversarial training may exploit the fact that non-greyscale perturbations can change the model's prediction.

2) Unconstrained targeted adversarial attacks on Restricted ImageNet Sketch 2 Models

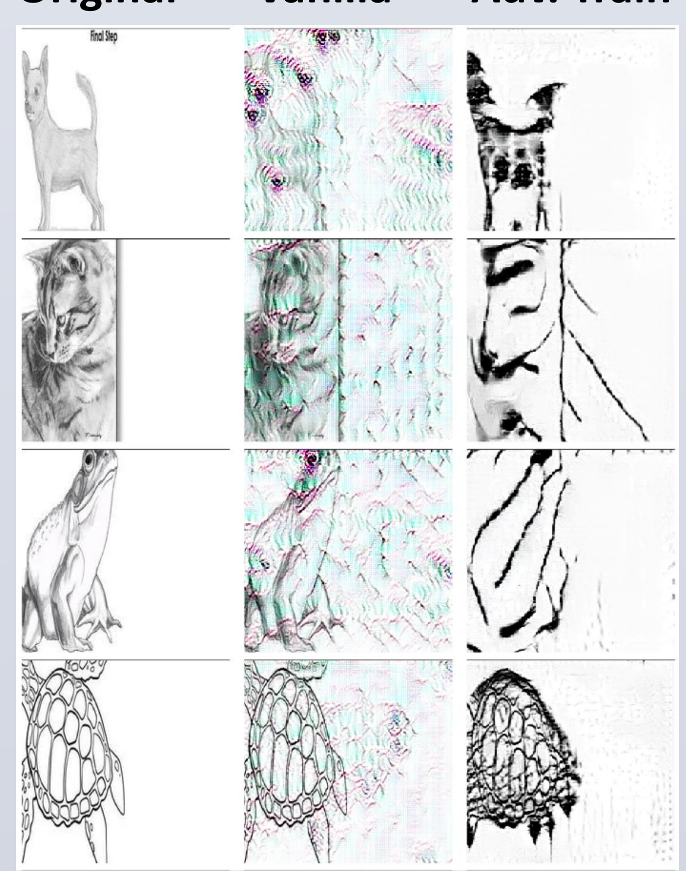


Figure 2: Adversarial attacks against the vanilla CNN produced images that appear noisy with color and are confidently misclassified. The perturbations always exist in the blank right-half of the original image. However, these attacks against the adversarially trained CNN resemble the target class and maintain the blank right-hand side of the original image.

- Adversarial training may exploit the fact that any perturbations to the blank-right half of the image can change the model's prediction.

Experiments

3) Pairwise Frobenius Distance between color filters (Channels 0-2)

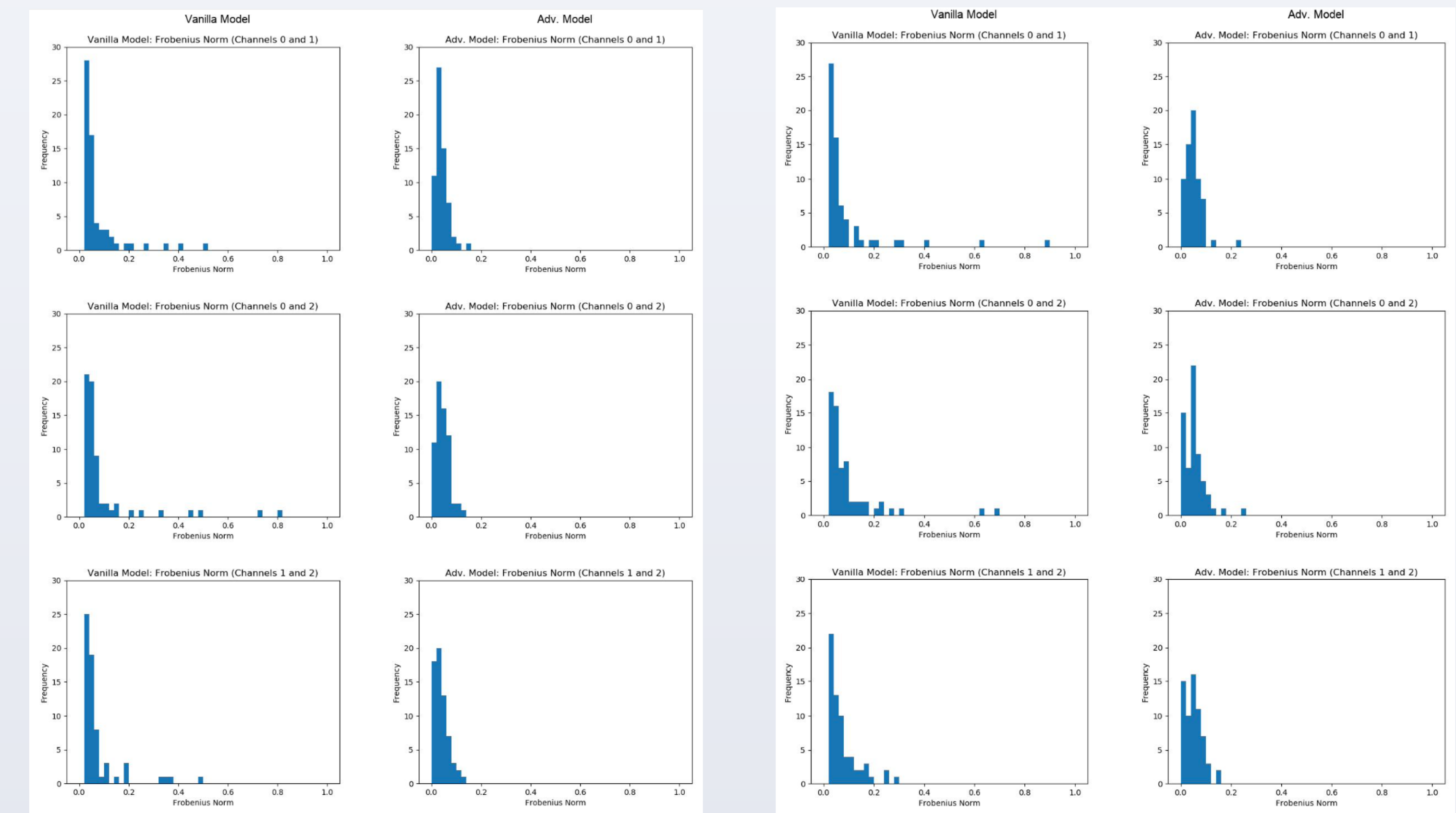


Figure 3: These histograms provide frequency distributions of the Frobenius norm between each pair of channels for each model. For both datasets, the pairwise distance between input channels is never in [0, 0.02) for the vanilla model but is for the adversarial model. These frequency distributions are more widespread for the vanilla models than for the adversarially trained models.

- Each model has a conv1 weight tensor with 3 input channels (Channels 0, 1, and 2) for color and 64 output channels
- While the vanilla models do not have any input channels with pairwise distance in [0, 0.02), the adversarially trained models always do
- For each pair of input channels, the vanilla models have a greater range of pairwise distance than the adversarially trained models do

Future Directions

- Replicate these experiments on more datasets of this nature
- Eventually, we hope to explain the link between standard vs adversarial training and differences in how certain phenomena are realized.

References

[1] Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Image synthesis with a single (robust) classifier. In Advances in Neural Information Processing Systems (NeurIPS), 2019.

[2] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In International Conference on Learning Representations, 2019.

[3] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," in Advances in Neural Information Processing Systems 32, 2019

Contact information

Simran Kaur: skaur@cmu.edu

Jeremy Cohen: jeremycohen@cmu.edu

Zachary C. Lipton: zlipton@cmu.edu