Adversarial Patch Attack

Sarwari Das, Mohammad Anas, Peining Yang

Duke PRATT SCHOOL of ENGINEERING

Duke MIDS Master in Interdisciplinary Data Science

Introduction

The last two decades have witnessed a tremendous increase in the applicability and performance of deep neural networks, with accuracy rates of CNNs matching or even out-performing human perception.

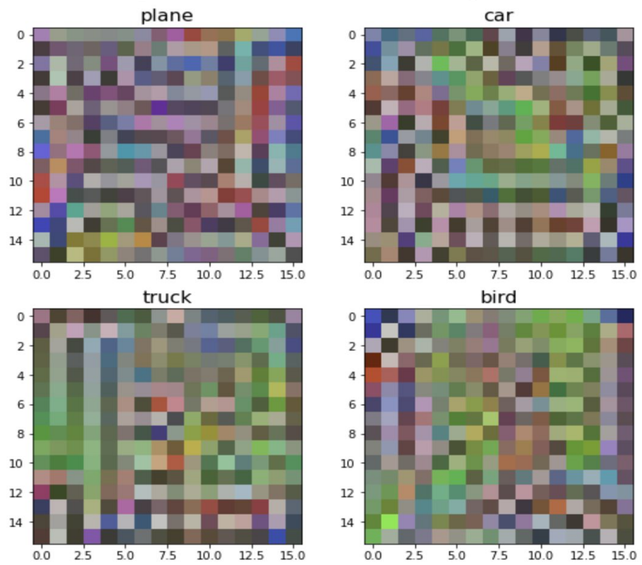
At the same time, a myriad of adversarial attacks has demonstrated the vulnerability of these systems by exposing the threats associated with implementing them in the real world. While most of these attacks are image-dependent and focus on making imperceptible changes to the input, in this project, we develop a "physical" attack in the form of adversarial patches that can be printed and applied to attack image and video classification models. These patches have shown to generalize to the real world, and will continue to be adversarial to classifiers even under different lighting conditions and orientations.

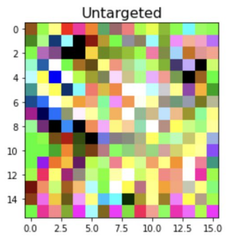
In this project, we take the role of an attacker to craft and train these patches to then mislead neural networks. We will:

- ➔ Develop "physical" attacks, or patches, that could be printed and applied in the real world without prior knowledge of the scene.
- ➔ Generate patch attacks in untargeted (failure to classify) and targeted (incorrect classification) fashions.
- ➔ Test the effect of rotation, location and transferability of the patches.

Methods

1. The CIFAR-10 data set consists of 60,000 colored images with 10 label classes. Each image is of the size 32 by 32 pixels in terms of height and width. The dataset was split into training and testing data with 50,000 images in the training data and 10,000 images in the test data
2. A convolution neural network following the Resnet-20 architecture was trained on this data and tuned to achieve an accuracy of 90% on the test dataset. This trained model was then used to generate targeted and untargeted patches. Due to a lack of computing resources, we were only able to generate a limited number of patches. The targeted patches were generated for the **car**, **bird**, **truck**, and the **plane** classes.
3. A rectangular patch of the required size (3x3, 5x5, 7x7 or 16x16) was initialized and placed at a random location in the training images. The patch was then optimized over 10 epochs with a learning rate of 0.1 and Adam optimizer.
4. The 16x16 patches are shown in the figure above.

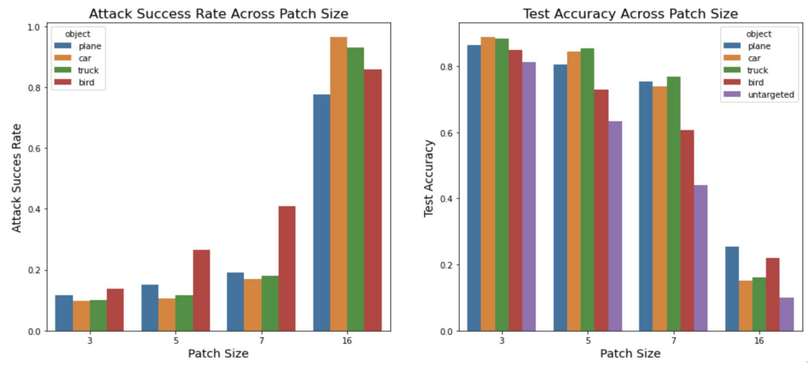




5. The generated patches were then placed at a random location inside the test images. The untargeted patches were evaluated using test accuracy and the targeted patches were evaluated using both test accuracy and attack success rate (ASR). ASR was defined as the proportion of images that were predicted as the target class.

Results

Larger patches were more successful in fooling our white box model as compared to smaller patches. As the patch size increased, our test accuracy dropped, and the attack success rate increased. This was found to be **true for all targeted and untargeted attacks.** Rotating untargeted and targeted patches of all 4 sizes 90, 180, and 270 degrees did not significantly improve ASR.



Transferability

We used a VGG 16 model architecture as our black box model. We used transfer learning from a model pre-trained on ImageNet to result in validation accuracy of about 92% on CIFAR10. On applying our patches, validation accuracy fell sharply for all patches, indicating that they are transferable.

Validation Accuracy for various patch sizes in targeted attacks				
Class name	***3x3***	***5x5***	***7x7***	***16x16***
Bird	26.08%	26.06%	23.76%	14.88%
Car	25.22%	24.27%	22.53%	12.64%
Plane	26.67%	24.51%	24.14%	13.04%%
Truck	25.98%	25.42%	24.96%	12.95%

# of times target was predicted for various patch sizes in targeted attacks				
Class name	***3x3***	***5x5***	***7x7***	***16x16***
Bird	1391	1684	2759	325
Car	462	690	1309	378
Plane	1424	1270	1409	26
Truck	100	107	112	175

Conclusion

As you can see from the images on the left, there are no discernible patterns in our patches. Given our current resources, the larger the patch size the higher the attack success rate. The location and orientation of the patches did not pose a significant effect on ASR, indicating that our patch attack has high generalizability. The patch attack is also transferable across models. We hypothesize that the experiment would produce even higher ASR if we further increase the patch sizes.

References

A. Kurakin, I. Goodfellow, and S. Bengio. *Adversarial examples in the physical world.* arXiv preprint arXiv:1607.02533, 2016.

A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. *Synthesizing robust adversarial examples.* arXiv preprint arXiv:1707.07397, 2017.

I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song. *Robust physical-world attacks on deep learning models.* arXiv preprint arXiv:1707.08945, 2017.

Brown, T. B., Mane, D., Roy, A., Abadi, M., and Gilmer, J. ´ *Adversarial patch.* CoRR, abs/1712.09665, 2017. URL http://arxiv.org/abs/1712.09665.