Adversarial Patch Attack on CIFAR-10 Classifiers

William Chown, Malcolm Smith Fraser, Wei Wu
Department of Electrical and Computer Engineering, Duke University

Abstract

Many deep learning systems are vulnerable to adversarial attacks. In this project we investigate the potential of an adversarial patch attack on images from the CIFAR-10 dataset as previously examined in Brown et al. [3]. A key component to the patch attack is that by not restricting the attack to be imperceptible, a scene-independent patch can be generated that is effective across various lightings and backgrounds – solving a limitation found in other attacks like the Fast Gradient Sign Method [1] and Projected Gradient Decent [5].

We find that a patch attack is successful in two modes of use: fooling a model into predicting a specific targeted class and simply fooling the model into predicting any erroneous class. We also demonstrate that a patch attack can remain effective when transferred to different models.

Objectives

- Implement adversarial patch attack to mislead CIFAR-10 classifiers in both untargeted and targeted fashion
- Evaluate the effect of the patch size
- Evaluate the transferability of adversarial patches

Methods

Dataset. CIFAR-10 [60,000 32x32 color images, 10 classes]. Images were normalized and converted to tensors.

Models. Target model: ResNet34. ResNet18 and Vgg16 used to evaluate the attack transferability. Models were pre-trained [6].

Patches. Pixel sizes: 2x2, 4x4, 8x8, and 16x16. Patch pixel values were normalized to fall within the range of the CIFAR-10 data. Patch locations were generated at random.

Training. Patches were trained as a new parameter in the pre-trained model over five epochs and initialized with zeros. The training procedure is shown below:

Algorithm 1 General Workflow for Training Adversarial Patch Attack

```
Initialize Patch with SGD optimizer, target class
for each training epoch do
    Place patch on image
    Generate predictions
    Compute cross entropy loss of model w.r.t. patch target class
    Back-propagation
    Update optimizer
end for
```

Evaluation. Top1 and Top5 accuracies measure how well the patch fooled the model into predicting the target class. Top1 and Top5 error measure the error the model generated to predict the true class.

Results

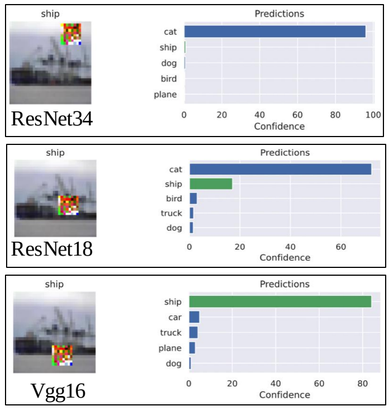

Figure 1. Model predictions with "cat" patch

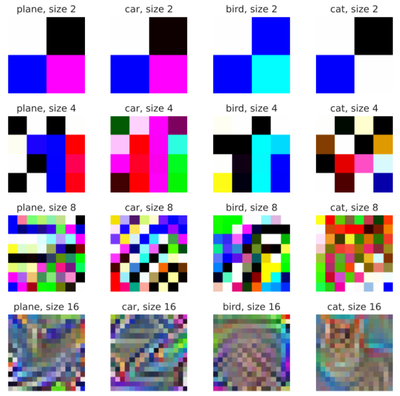

Figure 2. Trained Patches

Targeted Attack

Goal: Fool model into predicting the patch's target class

Target class: Car

Metric: Top1/Top5 accuracy

Result:

- Larger patches yield higher accuracies

Figure 3. Patch Accuracy vs. Patch Size

Model	Transfer Attack	Patch Size: 2x2	Patch Size: 4x4	Patch Size: 8x8	Patch Size: 16x16
ResNet34	No	(0.99%/19.27%)	(5.14%/47.34%)	(83.73%/98.85%)	(99.99%/100%)
ResNet18	Yes	(0.89%/24.11%)	(10.04%/66.90%)	(59.69%/92.75%)	(99.15%/100%)
Vgg16	Yes	(0.88%/31.89%)	(5.70%/50.08%)	(10.98%/35.75%)	(96.11%/99.72%)

Table 1. Top1 & Top 5 Accuracy for Different Models

Untargeted Attack

Goal: Fool model into generating erroneous predictions

Target class: NA

Metric: Top1/Top5 error

Result:

- Larger patches yield higher errors.

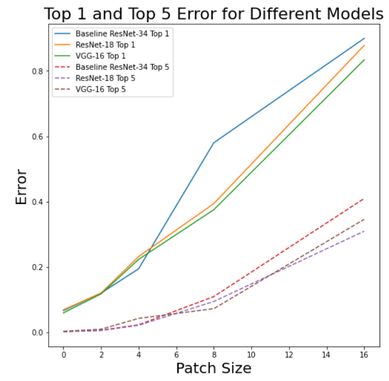

Figure 4. Model Error vs. Patch Size

Model	Transfer Attack	Baseline Error	Patch Size: 2x2	Patch Size: 4x4	Patch Size: 8x8	Patch Size: 16x16
ResNet34	No	(6.67%/0.25%)	(12.04%/0.70%)	(19.48%/2.35%)	(58.01%/10.98%)	(89.90%/40.17%)
ResNet18	Yes	(6.93%/0.26%)	(12.05%/0.57%)	(23.22%/2.21%)	(39.40%/9.48%)	(87.77%/30.96%)
Vgg16	Yes	(6.00%/0.36%)	(11.79%/1.00%)	(22.36%/4.31%)	(37.54%/7.29%)	(83.34%/34.60%)

Table 2. Top1 & Top 5 Error for Different Models

Conclusion

We implemented a simple, straightforward algorithm that generates patches that can easily attack deep learning models trained on the CIFAR-10 dataset. In particular, we found that:

For targeted attacks:

- Within the same model, ***larger*** patch sizes result in higher misled accuracies
- With the same patch size, ***more similarity*** between the transferred model and the baseline model also results in higher misled accuracies

For untargeted attacks:

- Within the same model, ***larger*** patch sizes result in higher errors
- With the same patch size, ***more similarity*** between the transferred model and the baseline model also results in higher errors

The increasingly widespread use of deep learning algorithms in diverse fields from natural language processing to autonomous vehicles to fraud detection to healthcare raises concerns over the severe security and privacy threats from adversarial attack techniques. Notably, pictures of simple adversarial patches can fool state-of-the-art facial recognition system, which exposes the underlying vulnerability of security systems using deep learning techniques [7]. Given the transferability, effectiveness, and ease of use of adversarial patch attacks, further research is required to improve the robustness of deep neural networks against these types of attacks.

References

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[2] Aleksander Madry, Aleksander Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial examples. arXiv preprint arXiv:1706.06083, 2017.

[3] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. arXiv:1712.09665, 2018.

[4] Chong Xiang and Prateek Mittal. Patchguard++: Efficient provable attack detection against adversarial patches. arXiv:2104.12609, 2021.

[5] H. Lin and B. Biggio. Adversarial machine learning: Attacks from laboratories to the real world. *Computer*, 54(05):56–60, May 2021.

[6] Huy Phan. (2021). huyvnphan/PyTorch_CIFAR10 (v3.0.1). Zenodo. https://doi.org/10.5281/zenodo.4431043

[7] M. Pautov, G. Melnikov, E. Kaziakhmedov, K. Kireev and A. Petiushko, "On Adversarial Patches: Real-World Attack on ArcFace-100 Face Recognition System," *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, 2019, pp. 0391-0396, doi: 10.1109/SIBIRCON48586.2019.8958134.