POSTER: Not All Pixels are Born Equal: An Analysis of Evasion Attacks under Locality Constraints

Vikash Sehwal^{*1}, Chawin Sitawarin², Arjun Nitin Bhagoji¹

Arsalan Mosenia¹, Mung Chiang³, Prateek Mittal¹

¹Princeton University ²University of California, Berkeley ³Purdue University

ABSTRACT

Deep neural networks (DNNs) have enabled success in learning tasks such as image classification, semantic image segmentation and steering angle prediction which can be key components of the computer vision pipeline of safety-critical systems such as autonomous vehicles. However, previous work has demonstrated the feasibility of using physical adversarial examples to attack *image classification systems*.

In this work, we argue that the success of realistic adversarial examples is highly dependent on both the structure of the training data and the learning objective. In particular, realistic, physical-world attacks on semantic segmentation and steering angle prediction constrain the adversary to add localized perturbations, since it is very difficult to add perturbations in the entire field of view of input sensors such as cameras for applications like autonomous vehicles. We empirically study the effectiveness of adversarial examples generated under strict locality constraints imposed by the aforementioned applications. Even with image classification, we observe that the success of the adversary under locality constraints depends on the training dataset. With steering angle prediction, we observe that adversarial perturbations localized to an off-road patch are significantly less successful compared to those on-road. For semantic segmentation, we observe that perturbations localized to small patches are only effective at changing the label in and around those patches, making non-local attacks difficult for an adversary. We further provide a comparative evaluation of these localized attacks over various datasets and deep learning models for each task.

KEYWORDS

Adversarial examples; Computer vision; Deep learning

1 INTRODUCTION

DNNs have achieved state-of-the-art performance on computer vision tasks such as object detection, classification, and segmentation. However, a large amount of recent literature has demonstrated the vulnerability of DNNs to adversarial examples [5, 12]. In particular,

(c) Patch 1: single car window (d) Patch 2: single car

Figure 1: Comparison of adversarial examples generated for Deeplabv3+ [6]. Fig. 1a, 1b presents the original and corresponding adversarial image respectively. Fig. 1c highlights that the success of a smaller adversarial patch is limited to its local pixels. A significantly larger patch size, which will be hard to use in physical-realizable attacks, is required to propagate the effect of added patch to non-local pixels.

previous work has highlighted that these vulnerabilities exist in the physical-world as well [4, 8, 10, 11].

These attacks succeed by embedding the adversarial perturbations on a physical object in the targeted environment. Sharif et al. [10] add perturbations on eyeglass frames to fool facial recognition systems, whereas Brown et al. [4] print out adversarial patches which, if added to an object, can attack image recognition systems. Imperceptible perturbations are used to fool traffic sign classification networks in [8, 11]. It should be noted that these existing works has considered only image classification, a task which has been generally superseded by considerably more complex tasks such as image segmentation in applications like autonomous vehicles [7].

In this work, we evaluate the success of realistic adversarial examples with localized perturbations for image classification, steering angle prediction and semantic image segmentation. To our knowledge, this is the first work to evaluate the success of localized visual adversarial perturbation for multiple computer vision applications. Localized adversarial perturbations are of interest when carrying out physically realizable attacks since an adversary cannot, for example, cover an entire street with adversarial perturbations! The critical nature of this locality constraint is demonstrated in Figure 1. Our main findings are as follows:

1. Not all pixels are equally important: Using multiple datasets and targeted applications (Section 3.1, 3.2), we demonstrate that the attack success is highly correlated with pixel locations. For e.g., with the MNIST and Udacity self driving car datasets [2], we observe

^{*}vvikash@princeton.edu

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CCS '18, October 15–19, 2018, Toronto, ON, Canada

© 2018 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5693-0/18/10.

<https://doi.org/10.1145/3243734.3278515>

a significant difference in attack success when perturbations are localized to specific pixel locations such as center and border pixels.

2. Attack success rates are highly correlated with size of locality constraint: With each application, we observe a significant decrease in attack success with a decrease in size of the locality constraint. With semantic segmentation, we demonstrate that this correlation further depends on the choice of adversarial objective (Section 3.3). Under complex adversarial objectives, such as classification of each pixel as a random label, the attack success with localized perturbations decreases significantly as compared to adversaries able to add noise to the entire image.

3. Current attack methods do not produce reliable adversarial examples under locality constraints: We generated adversarial perturbations under locality constraints using the state-of-the-art attack method [5]. In each experiment, we observe a significant variance in the success of this attack with different input images and the size and location of added perturbations. This limitation indicates that current attack algorithm have to be improved to carry out realistic attacks.

2 SETUP AND METHODOLOGY

Our primary focus is to evaluate the success of evasion attacks with localized adversarial perturbations in multiple computer vision application such as image classification, segmentation, and steering angle prediction. To generate adversarial perturbations, we minimize Lagrangian relaxation of the adversarial objective [5] using projected gradient descent with the Adam algorithm over 200 iterations and a learning rate of 0.1. We evaluate attack success only with targeted attacks which are strictly harder than untargeted attacks. To be consistent with previous work [4], we use the terms localized perturbation and adversarial patch interchangeably. We measure the target success rate and accuracy for each adversarial attack. The former refers to the percentage of perturbed images classified as targeted labels while the latter represents the percentage of perturbed images classified as correct labels.

Initially, we experiment with image classification where the adversarial objective is the classification of perturbed images as the targeted labels. With steering angle prediction, the network takes an input or a sequence of inputs and learns to predict the correct steering angle for a perturbed image. In semantic segmentation, the network learns the pixel-wise classification of input images, which is considerably more difficult than predicting a single label for the whole image. The adversarial objective with image segmentation is the classification of each pixel of the perturbed image as its target label.

3 RESULTS

3.1 Image Classification

We experiment with MNIST and ImageNet datasets for image classification. For MNIST, we train a four-layer convolutional neural network, achieving 99.17% accuracy. For ImageNet, we use the pre-trained InceptionV3 model, which achieves 94.6% top-1 accuracy on ImageNet-1k subset [1]. We experiment with both center and border patch location, where the center patch refers to masking of

Figure 2: Limited success of border adversarial patches in comparison to center patches with MNIST (M) and ImageNet (I) dataset. In comparison to ImageNet, the difference is significantly more pronounced for MNIST. It highlights that the training dataset significantly affects the correlation of adversarial patch success with different patch locations.

$n \times n$ pixels at the center, while the border patch refers to masking n border pixels of an image (we choose n between 0-10).

Takeaway 1: Attack success is significantly influenced by patch location. For both accuracy and target success rate (Figure 2), we observe that the border patches lead to significantly poorer performance compared to the center patches on MNIST. For instance, an adversarial center patch with 8×8 pixels size degrades accuracy to 0.77% from original 99.17%, while a border patch of similar area reduces the accuracy only to 81% (Figure 2a). The target success rates for these patch sizes are 80% (close to maximum target success possible) and 0.2% (ineffective) respectively (Figure 2b).

Takeaway 2: Training dataset influences the correlation between attack success and patch location The poor success of border patches can be attributed to the consistent presence of *less informative features* at the border pixels across all images in the MNIST dataset. It can be due to the reason that the MNIST dataset was originally created by centering 20×20 pixels images from NIST dataset in a 28×28 pixels bounding box. To validate this hypothesis we further experiment with the ImageNet dataset, which is more diverse than MNIST. With ImageNet, even with small patch areas, both border and center patches lead to significant attack success. However, as the dataset is originally aggregated by filtering search results from the internet for images corresponding to WordNet synsets, it is plausible for it to exhibit a bias towards key objects being in the center of the image. This difference can be observed in Figure 2 as the center patch yields a higher target success rate and a lower accuracy compared to a border patch of the same size.

3.2 Steering Angle Prediction

For steering angle prediction, we consider two neural networks: PilotNet [3] and Chauffeur [2]. PilotNet is a convolutional neural network (CNN) proposed by NVIDIA [3], while Chauffeur, a CNN+LSTM model, achieved the third place in the Udacity steering angle prediction challenge [2]. We evaluate the attack success on this task using both off-road and on-road patches over 100 random images from the test set of Udacity self-driving car dataset [2]. With each patch, the adversarial objective is to predict a 90° right turn.

Takeaway 1: Attack success is significantly affected by patch location. The median change in output angle for an on-road patch is at least $2-3\times$ that of the change with an off-road patch of similar size (Figure 3). This observation supports the previous hypothesis that the attack success rate with adversarial patches depends on the patch location. In real-world driving datasets, such as [2], a significant number of training images capture straight roads. Due

Figure 3: Change in output steering angle due to the adversarial patch. In comparison to on-road patches, it highlights the limited success of off-road patches for both models. This difference between the success of both type of patches further depends on model architecture, where deeper networks (e.g., Chauffeur) achieves higher robustness to off-road patches as compared to shallower nets (e.g., PilotNet).

to this bias towards images of straight roads, where the key features such as lane markings, road edges, and curvature are missing at border pixels, the network learns a weaker correlation between output and border pixels. Thus with off-road patches, which tend to occupy border pixels, the attack success rate is very limited.

Takeaway 2: Network architecture significantly influences attack success. We observe a significant difference in the attack success rate of on-road and off-road patches for the Chauffeur model compared to PilotNet. It can be attributed to its very deep network, a combination of CNN and LSTM, which is expected to learn better features than a single-CNN based model.

Takeaway 3: Decrease in patch size leads to a decrease in attack success. Due to the regression nature of the steering angle prediction problem, the attack success decreases approximately linearly with patch area. This requires the adversary to use large patch sizes to achieve significant change in the output of the network.

3.3 Semantic Image Segmentation

We evaluate semantic image segmentation robustness on two state-of-the-art models: Deeplabv3+ [6] and PSPNet [13]. We evaluate the attack success of each $n \times n$ (we choose n between 0-713) pixels adversarial patch with random locations for 100 randomly selected images from test partition of Cityscapes [7] dataset.

Takeaway 1: Attack success is highly dependent on the learning objective. Figure 4 shows the limited success of adversarial patches which highlights that a single adversarial patch is unlikely to influence the classification of each pixel in the input image. We anticipate that this behaviour is most likely due to the limited effective receptive field of neural networks [9], which makes it very difficult to achieve non-local changes with smaller patch sizes.

Takeaway 2: Adversarial objective significantly affects attack success. We initially select the target adversarial label for each pixel as its least likely predicted label. However, we observe that the attack cannot achieve a 100% target success rate using this objective, even with perturbations added to the whole image. Thus, we also report another set of results with an easier objective where the target label for all pixels is identical but selected randomly from all possible categories (similar to Figure 1). The higher success of adversarial patches with this objective highlights that the adversarial objective can influence the success of adversarial patches.

Takeaway 3: Network architecture influences attack success. We observe approximately 1.5-2 $\times$ higher target success rate for PSPNet than Deeplabv3+. This difference can be attributed to the

Figure 4: Attack success of adversarial patches with different adversarial objectives for Deeplabv3+ (D) and PSPNet (P) models. With semantic segmentation, smaller patch sizes are ineffective in achieving significantly high attack success. The attack success is further dependent on adversarial objective, i.e., selecting the target label for each pixel, and the network architecture.

different architecture of these models. We are currently performing ablation studies to understand the key reasons for this difference.

4 CONCLUSION

This paper evaluates the success of localized adversarial perturbations for three different key applications of deep learning in computer vision. Our analysis indicates that in order to carry out realistic evasion attacks on these systems, better attack strategies taking their architecture and datasets into account are required. We are currently exploring these.

5 ACKNOWLEDGEMENT

This work was supported by the National Science Foundation under grants CNS-1553437 and CNS 1409415, by the Office of Naval Research (ONR) through Young Investigator Prize (YIP), by IBM through IBM Faculty award and by Intel through Intel Faculty Research Award.

REFERENCES

- [1] 2017. NIPS 2017: Non-targeted Adversarial Attack. (2017).
- [2] 2018. self-driving-car: The Udacity open source self-driving car project. (2018).
- [3] Mariusz Bojarski, et al. 2016. End to End Learning for Self-Driving Cars. *arXiv:1604.07316 [cs]* (April 2016).
- [4] Tom B. Brown, et al. 2017. Adversarial Patch. *arXiv:1712.09665 [cs]* (Dec. 2017).
- [5] Nicholas Carlini and David Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (SP)*. 39–57.
- [6] Liang-Chieh Chen, et al. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *arXiv:1802.02611 [cs]* (Feb. 2018).
- [7] Marius Cordts, et al. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. *arXiv:1604.01685 [cs]* (April 2016).
- [8] Kevin Eykholt, et al. 2018. Robust Physical-World Attacks on Deep Learning Visual Classification. In *IEEE CVPR '18*.
- [9] Wenjie Luo, et al. 2016. Understanding the effective receptive field in deep convolutional neural networks. In *NIPS*. 4898–4906.
- [10] Mahmood Sharif, et al. 2016. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In *ACM CCS '16*.
- [11] Chawin Sitawarin, et al. 2018. DARTS: Deceiving Autonomous Cars with Toxic Signs. *arXiv:1802.06430 [cs]* (Feb. 2018).
- [12] Christian Szegedy, et al. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).
- [13] Hengshuang Zhao, et al. 2016. Pyramid Scene Parsing Network. *arXiv:1612.01105 [cs]* (Dec. 2016).