Impact of Adversarial Patches on Object Detection with YOLOv7

Darrien Hunt, Chutima Boonthum-Denecke, and Idongesit Mkpong-Ruffin#
Hampton University, Department of Computer Science
Florida A&M University, Department of Computer and Information Sciences

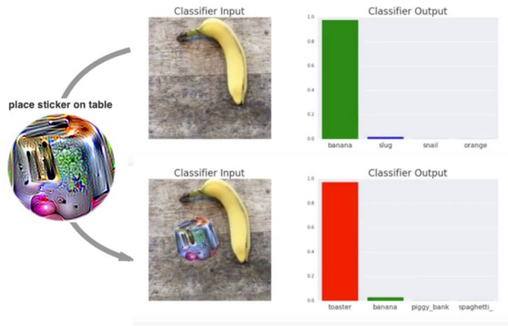
Abstract

With the increased use of machine learning models, there is a need to understand how machine learning models can be maliciously targeted. Understanding how these attacks are 'enacted' helps in being able to 'harden' models so that it is harder for attackers to evade detection. We want to better understand object detection, the underlying algorithms, different perturbation approaches that can be utilized to fool these models. To this end, we document our findings as a review of existing literature and open-source repositories related to Computer Vision and Object Detection. We also look at how Adversarial Patches impact object detection algorithms. Our objective was to replicate existing processes in order to reproduce results to further our research on adversarial patches.

Adversarial Attacks & Adversarial Patches

Adversarial attacks aim to fool machine learning models. Adversarial attacks can be done by making changes to a physical object so as to fool the machine learning model so that an image or object is mistaken for some other object or image or in some cases, not even detected. This ability to make a change to an object for the purpose of fooling a system is also known as **adversarial perturbation**. In this work, we investigated the usage of patches to cause the image classifier to misclassify, misidentify or be unable to identify given objects. **Adversarial patches** are images that once printed, added, or presented to the image classifier, can cause the classifier to ignore the other items or misidentify the items

Method



An attack on VGG16 using a physical patch.

Patch # 1

Patch #2

YOLOv7 Repository: https://github.com/wongkinyiu/yolov7 with Python Library.

Patches: Two (2) adversarial patches were selected for this experiment

Data Set: ***ImageNet*** consisted of images of people and objects/animals: bicycle, birds, cars, cell phones, and dogs.

Process: Run the detection algorithm on all the images to generate a **confidence value** with a border surrounding the classified objects and run the same algorithm on **altered images**, which an adversarial patch was placed over the face, over the body, or different part of the large items of the image. The confidence value represents how the model accurately defines an object. Customized script with parameters to set pre-trained model weights, a confidence value threshold, specify image size, and a target directory for our images to perform classification successively.

Car 1 and Car 2 with Patch # 1

While the original images average confidence value was 0.83496, Patch #1 had an overall negative effect on the detection rate of the classifier, producing a result of 0.71171. This drastic decrease was most sharply noticeable with the images of cars which had large patches applied to them and numerous detectable objects in the background of the images. Car1 and Car2 with Patch#1, background objects and objects within the patch itself served as potential impediments to the detection accuracy of the primary car. The size of the patch may also be an overwhelming factor given the algorithm's ability to attempt detection of the patch.

Patch # 2 on person's body vs face

Overall, the unaltered images possessed an average confidence value of 0.928238. When Patch #1 was applied over the body and the face, the confidence values generated were 0.9224 and 0.8647684 respectively. Patch #2 resulted in values of 0.93005 and 0.921353 when applied over the body and the face respectively, citing an increase in detection accuracy over the body and a miniscule decrease over the face.

Results

	Confidence Value		
Object	Non-people	Non-people Patch #1	Non-people Patch #2
Bicycle	0.947094	0.948981	0.941189
Bird 1	0.785761	0.75917	0.796503
Bird 2	0.687701	0.475139	0.616994
Car 1	0.90718	0.397678	0.907187
Car 2	0.784057	0.540132	0.77098
Dog 1	0.771059	0.916712	0.883784
Dog 2	0.961846	0.944186	0.956071
Average	**0.83496**	**0.71171**	**0.83896**

Table 1. Confidence values of YOLOv7 ran on subset of images representing non-people

	Confidence Value				
Object	People	People Patch #1 (Over Body)	People Patch #1 (Over Face)	People Patch #2 (Over Body)	People Patch #2 (Over Face)
Person 1	0.972414	0.965844	0.788071	0.973064	0.967817
Person 2	0.95275	0.949856	0.817727	0.956822	0.956297
Person 3	0.962644	0.960276	0.952369	0.961844	0.950804
Person 4	0.967071	0.953975	0.951217	0.961021	0.966046
Person 5	0.946961	0.934886	0.932216	0.943588	0.898543
Person 6	0.957467	0.949614	0.724247	0.959409	0.949607
….					
Person 13	0.916603	0.874622	0.699182	0.906344	0.920201
Person 14	0.821476	0.740749	0.83696	0.792484	0.85599
Person 15	0.966412	0.964674	0.961278	0.960658	0.969172
Average	**0.928238**	**0.9224**	**0.8647684**	**0.93005**	**0.921353**

Table 2. Confidence values of YOLOv7 ran on subset of images representing people

Conclusion

Overall, this project helped develop our understanding of computer vision and object detection. Exposure to rapidly improving object detection models such as YOLOv7. Reproducibility of this project was the first step in understanding how object detection works and to opening the door to working on improving the model itself. We were able to determine the effects that adversarial patches had on detection accuracy and analyze the results to create more hypotheses.

Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems, 30(11), 3212-3232.
Jaikumar, P., Vandaele, R., & Ojha, Varun. (2022). Transfer Learning for Instance Segmentation of Waste Bottles using Mask R-CNN Algorithm. In: Abraham, A., Piuri, V., Gandhi, N., Siarry, P., Kaklauskas, A., Madureira, A. (eds) Intelligent Systems Design and Applications. ISDA 2020. Advances in Intelligent Systems and Computing, vol 1351. Springer, Cham.
Thys, S., Van Ranst, W., & Goedemé, T. (2019). Fooling automated surveillance cameras: adversarial patches to attack person detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops.