Robust Assessment of Real-World Adversarial Examples

Brett Jefferson
brett.jefferson@pnnl.gov

Carlos Ortiz Marrero
carlos.ortizmarrero@pnnl.gov

Pacific Northwest National Laboratory

Abstract

We explore rigorous, systematic, and controlled experimental evaluation of adversarial examples in the real world and propose a testing regimen for evaluation of real-world adversarial objects. We show that for small scene/ environmental perturbations, large adversarial performance differences exist. Current state of adversarial reporting exists largely as a frequency count over a dynamic collections of scenes. Our work underscores the need for either a more complete report or a score that incorporates scene changes and baseline performance for models and environments tested by adversarial developers. We put forth a score that attempts to address the above issues in a straightforward exemplar application for multiple generated adversary examples. We contribute the following: 1. a testbed for adversarial assessment, 2. a score for adversarial examples, and 3. a collection of additional evaluations on testbed data.

1. Introduction

Now that Deep Learning is an established success [12], there is a rapidly expanding body of work assessing its limitations [17, 9, 2]. In particular, there has been a large number of papers published in recent years, interested in finding new ways to hack deep learning systems with a focus on manipulating convolutional neural networks into false and missed classifications [15, 14, 3]. Much of the early work with so-called adversarial attacks were only successful in virtual environments, *i.e.* the adversarial inputs were produced and evaluated digitally and without consideration of physical limitations. In the past year, researchers have expanded adversarial attacks to include physically created objects that can impact classifiers and detector models in real-world systems [10, 1, 8, 19]. The range and ability of physical attacks are improving at an impressive rate, accounting for a variety of real-world considerations including static scenes, robust angle, and distance changes. Although researchers have established some guidelines for evaluating the robustness of virtual adversarial attacks [4], the analysis recommendations do not map subjectively onto physical adversarial attacks where perturbations are difficult to measure and success varies on a frame by frame basis. The consistent computational measure of success for physical attacks is the percent of frames the attack accurately manipulated the classifier or detector [7, 1, 20]. This is separate from adversarial generation metrics that are often included in the optimization loss function to improve physical challenges like imperceptibility and printability. In our work we propose an evaluation experiment and post-generation, effectiveness score for testing the robustness of real-world adversarial examples in different environmental conditions. In particular, we tested our score on adversarial "patches", an idea clearly outlined in work done by Thys et al [19], but incorporating ideas from Athalye et al. [1], Chen et al. [5], and Eykholt et al. [7]. The key aspect we wish to address with our score is the importance of baseline performance across different environmental conditions when assessing the effectiveness of a given physical adversarial object.

2. Our Aim

We recognize that it is not always possible to recreate environments to evaluate adversarial versus non-adversarial scenarios. However, many of the existing adversarial objects have the ability to be evaluated in well-controlled scenarios that can be reproduced.

We believe that in conjunction with well-controlled experiments, a proper adversarial score must be included when evaluating adversarial objects. Our proposed score does so in simple to interpret terms and is aimed at providing a foundation that can be updated, revised, and enhanced by others studying the problem of adversarial scoring. The score is designed to be a relative measure of performance that considers only those environments studied by the researcher while taking into account the same scenario in a non-adversarial condition. This is, in one sense, akin to something like a Bayesian information criterion score that only measures the model and parameters presented and cannot explicitly account for non-nested models.

Experimentally, we studied a single target object with three differently trained adversarial patches and an occlusion condition in several well-controlled scenarios. To the best of our knowledge, this is the most systematic assessment of adversarial attacks to date. For comparison, [20] studied the effectiveness of adversarial attacks in an indoor environment and outdoor environment. While studying the attacks under similar scenes (across multiple days and weather conditions), by making changes in angle, distance, and day, the authors did not control for confounding factors such as other objects being present in the scene (scene complexity) nor did they report a baseline for the model. Our work aims at providing a coarse view of adversarial effectiveness using a fine-grained paradigm.

The paradigm in our work can be expanded to include more scenes/ environments, but we found it important to have an initial study that avoided confounding factors (such as patch performance being modulated by distance of the patch from the camera), while still providing plausible scene perturbations. For example, using our approach we can begin to quantify scene appropriateness or complexity, although this question deserves its own dedicated exploration.

3. Experiments

Many researchers focus on the performance of an adversarial attack in native environmental conditions (e.g. a patch on the bumper of a vehicle in actual traffic, a patch attached to a stop sign, or a patch attached to a person's clothing in an office space). We assess performance in a well-controlled environment with little frame-to-frame variability due to moving objects, novel items entering the scene, changing lighting, etc. This tight control is necessary to accurately compare patch performance to a baseline condition where the model is allowed to detect objects without adversarial interference. In other words, we needed an environment that was reproducible and where happenstance occurrences were not a factor. For reproducibility, below we outline our experimental equipment and setup.

3.1. Equipment

Equipment for the experiment included scene setting items and camera hardware. Our scene setting items include a custom constructed mounting rail measuring roughly 7 feet long with attached platform (4 feet high) with plate for attaching camera devices and light source with one 40-watt, 390-lumens halogen bulb and one 40-watt, 450-lumens LED bulb. Camera and GPU devices include Jetson AGX Xavier, Jetson Xavier Developer Kit with an attached e-CAM130_CUXVR camera. We tested 3 pre-printed adversarial patches created via different algorithm methods.

3.2. Patch Generation

We generate our patches using the training technique outlined by Thys et al. [19]. The broad idea can be summarized as follows:

1. Curate a set of training images that your object detector can recognize.

2. For each image we superimposed a patch (300×300 pixels, then scaled accordingly to fit the size of the object bounding box) to the image. We use the Expectation over Transformation algorithm to produce our adversarial patch using the following transformations: change of location, rotation angle, scale, brightness, contrast, and noise level [1].

3. We extract a classification score from these altered images, back-propagate the gradients to the input layer, and only update the pixels inside the region of the patch.

We leverage and extend the codebase provided by Thys et al. [18] to generate adversarial patches for all classes contained in the output of our model. In our case, these are all classes in the COCO dataset [13].

We trained three patches for the vase class (our target object) using images from ImageNet [6] and OpenImages [11]. Our patches were obtained by minimizing two different objective functions: objectness score (O) and the product of class probability score and objectness score (CxO). For our ImageNet patches we triage images from the WordnetID *n04522168*, corresponding to the ImageNet "vase" class. For our Composite patch, we combined our extracted images from ImageNet with images extracted from OpenImages corresponding to the "vase" class name.

3.3. Procedure

During piloting we found our patches should be 2×2 inches. We manipulated three environmental conditions/ dimensions: target item distance (1 inch, 5 inches, 10 inches, 15 inches and 20 inches from a Plexiglas surface), patch placement location (center and slight right of center), and lighting (2 bulb types). This brought the total number of environmental conditions to $5 \times 2 \times 2 = 20$. There were five patch conditions including no-patch (baseline), ImageNet (O), ImageNet (CxO), Composite (CxO), and a white patch, which was a simple white 2 inch square cut from 8.5×11 inches, 92 bright copy paper. All patches were printed on the aforementioned office paper at 1200 dpi. Section 3.2 describes in detail how the patches where generated. Each patch type except for the no-patch condition was used in all environmental conditions. For the no-patch condition, the target item was always center-placed, but used with all light source and distance combinations.

Our camera was manually set so that white balance, exposure time, and focus was fixed throughout the experiment (rather than automatically adjusted by the camera). The focus was set so that the sticker and vase were in focus from the camera (focal length: 2.8 mm; F-number: 2.8; Field of View: 134° (D), 73° (V)). We were able to position the camera at closer positions with such a wide field of view and shorter focal length. All other camera settings were constant throughout the data collection.

Our camera was placed 5 inches from a Plexiglas surface affixed to a table. The Plexiglas served as a mounting structure to ensure consistent placement for each patch. There was a single light source used at a time and the source was positioned behind and above the camera, pointed toward the target placement region. The camera was placed on the rail platform and fixed for the duration of the experiment. A large black board was used as background for the experiment. Testing proceeded as follows:

1. For a given distance, the target item (a green vase, see Figure 2) was placed in one of two positions depending on the patch placement condition.

2. With the target fixed, the no-patch condition was recorded first. Then each patch (white patch, ImageNet (O), ImageNet (CxO), and Composite (CxO)) was placed on the Plexiglas. Each time a patch was placed, it remained untouched through both lighting conditions. This was to minimize object and patch shifts across the conditions.

3. Once a scene was established, we allowed 30 seconds for the bulb to warm-up.

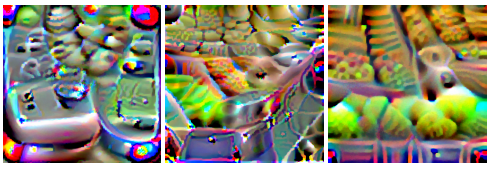
4. We ran a script that captured 500 frames and used each frame as an independent input to YOLOv2 [16]. We wanted enough frames for a single scene for a robust evaluation of that condition in lieu of natural image variation produced by the camera or by nature.

5. We recorded bounding boxes, confidences, objectness scores for each frame.

4. Results and Effect Score

4.1. Classifications

Figure 1: Patches Generated for Experiment



Each of the above patches was designed to hide a target item (i.e. a green vase) from detection for the YOLOv2 classifier. In addition to these patches a simple white square patch was also used to compare performance with an obstruction case. A first evaluation of each patch's ability to hide the target was to simply count the number of frames the classifier was able to detect the target class in each scene (See Table 1). This is a standard measure. Higher values in the table indicate the patch was not effective at disrupting detection of the target.

There are a few conditions that stand out when looking at only frequencies. When the target item was placed very close to the camera (1 inch condition) and **no patch was present**, the classifier had difficulty detecting it. In all frames with LED lighting, the target was missed, while in the halogen bulb lighting, the target was detected in less than 40% of frames. Another stand-out is the LED, 1-inch, ImageNet (CxO) condition where the number of detections is higher than 500. In this case, the target is detected twice, once as a lower identification of the target and an upper identification of the target (see Figure 2).

Judging from frequency alone, one might conclude that the Composite (CxO) patch and ImageNet (CxO) patch are *better* than the other two patches. This conclusion would match intuition since one patch had a larger training set and both patches were trained using an objective function accounting for both class score and objectness. But there is more to be discovered. For instance, suppose one desired an all-around effective patch for a variety of physical environments. Is the Composite (CxO) patch better than the ImageNet (CxO) patch? A problem with error frequency is that is does not account for how well the model performs without adversarial interference. In the LED, 1-inch case, not having a patch at all is better than adding anything to the scene. We seek to develop a score that not only accounts for a variety of environmental changes, but also accounts for baseline performance in one summary.

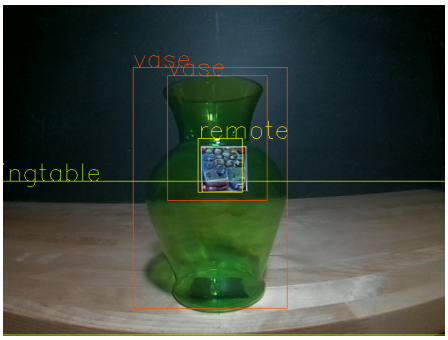


Figure 2: In this condition the vase was detected twice in nearly all frames.

Table 3 shows the classifications per patch throughout the experiment. Because YOLOv2 is an object detector,

Table 1: Number of Vase Detections (and Average Confidence) for Each Condition

Location	Bulb	Distance	None	Composite (CxO)	ImageNet (CxO)	ImageNet (O)	White Patch
center	**Hlgn**	**1 inch**	182 (.856)	7 (.609)	0	500 (.785)	500 (.822)
		5 inches	404 (.905)	20 (.828)	0	500 (.869)	500 (.887)
		10 inches	500 (.879)	0	0	30 (.795)	0
		15 inches	298 (.846)	0	0	1 (.771)	0
		20 inches	499 (.823)	0	0	32 (.670)	0
	LED	**1 inch**	0	419 (.818)	955 (.602)	500 (.846)	500 (.834)
		5 inches	500 (.928)	500 (.892)	391 (.691)	500 (.938)	500 (.915)
		10 inches	500 (.872)	27 (.870)	0	500 (.823)	456 (.766)
		15 inches	500 (.872)	0	0	500 (.763)	0
		20 inches	500 (.880)	0	0	28 (.694)	0
right	**Hlgn**	**1 inch**	182* (.856)	0	18 (.835)	499 (.817)	479 (.817)
		5 inches	404* (.905)	0	206 (.914)	500 (.904)	313 (.899)
		10 inches	500* (.879)	206 (.855)	0	302 (.819)	1 (.715)
		15 inches	298* (.846)	0	0	26 (.761)	0
		20 inches	499* (.823)	0	0	1 (.551)	0
	LED	**1 inch**	0*	2 (.820)	457 (.851)	500 (.867)	500 (.866)
		5 inches	500* (.928)	247 (.928)	500 (.947)	500 (.934)	500 (.931)
		10 inches	500* (.872)	500 (.893)	500 (.859)	500 (.840)	500 (.807)
		15 inches	500* (.872)	479 (.878)	500 (.860)	500 (.689)	490 (.855)
		20 inches	500* (.880)	0	0	0	120 (.727)

*When no patch was present, targets were center located. Only one run of the model at different distances and bulb types was completed without a patch.

multiple objects can be classified in a given frame. As a result, many classified objects are not misclassifications of the vase, but misclassifications of other scene objects. While, we attempted to minimize this effect, we often found that the table the vase was placed on was classified as a dining table and the background was classified as a refrigerator. More consistent misclassifications included classifying the vase as a bottle or a cup. 'Bottle' labels occurred frequently, even in the absence of an adversarial patch. The patches themselves were classified in some many instances. The patches trained to decrease class score and objectness were classified as a cell phone or remote, while the third patch (optimized for objectness only) was classified as a wider range of objects. Here we reemphasize that no patch hid the vase 100% of the time, but that there are some scenes where patches performed well and others that it simply did not work. We did a parameter sweep to pick penalties for the non-printability and total variation term in the loss function and trained all patches until we saw no improvement in the loss function. The question of producing the "optimal" patch was outside the scope of our work, given that our main focus was assessing adversarial patches.

4.2. Effectiveness Score

To gain a better understanding of patch performance, we make the straightforward adjustment of comparing patch performance for a given scene with model performance in the absence of an adversary. As noted above, there were several misclassifications of the vase in the no patch condition, leading to potential misunderstanding of the effectiveness of a given adversary. Our proposed score is derived

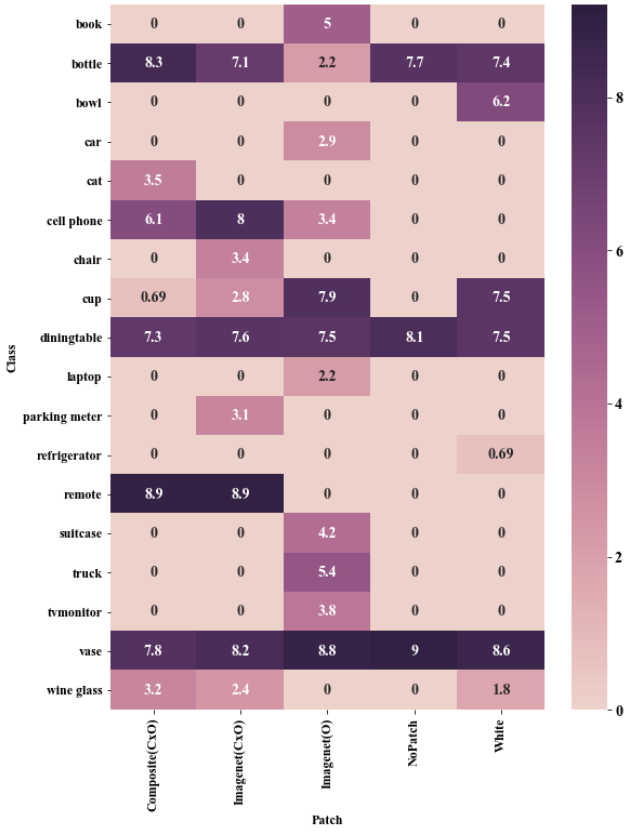


Figure 3: Depiction of all classifications per patch. These are summed over all environmental conditions and logged.

for each patch in a given scene/ condition (or conditions). For a given patch P and a single environmental condition e (e.g. lighting, distance, patch location), we compute the frequency of target detection in both an adversary condition (patch present) and baseline (adversary not present). Let the $f_{P,e}$ be the frequency for a given patch and environmental condition (out of the set of tested environmental conditions) $e \in E$. We let $f_{\emptyset,e}$ denote the frequency of target detection of baseline. Let n denote the total number of frames captured. We define the score for a patch conditionally over the set of environments to be

$$S(P, E) = \frac{1}{|E|} \sum_{e \in E} \frac{\left(f_{\emptyset,e} - f_{P,e}\right)}{n}$$

This is the simply the average difference of detection probabilities between the no-patch and patch conditions. For this score function, when the target is detected in all 500 frames of our experiment and the patch successfully hides the target in all 500 frames, the score will be 1. When the baseline model is unsuccessful at detecting the target, the score is lower. Negative scores can be interpreted as the adversary having the highly undesired effect of helping the model to detect a target more often, instead of less. Lastly, the score is averaged across all conditions run to provide a simple summary of performance. The scores for our patches are given in Table 2. A current limitation is that the score cannot distinguish between poor baseline versus poor adversary.

Note: It is interesting to note that the white patch outscores the patch trained using only the ImageNet corpus with objective function only utilizing the *objectness* score from YOLOv2. A further investigation is needed to rule out or help validate the causes of the observation. Nonetheless, our goal is to highlight issues like this that would only show up when doing a systematic robust evaluation. Furthermore, we are not making generalized claims about the specific algorithms used in this investigation but are advocating for more systematic studies as exemplified through our methods.

Table 2: Patch Scores

Patch Condition	Score
No-Patch	0
Composite (CxO)	**0.536**
ImageNet (CxO)	0.424
ImageNet (O)	0.135
White Patch	0.241

We also computed scores for each condition independent of the other conditions (each E is a singleton). The analysis in the next section dives into the pattern of results on these dimensions.

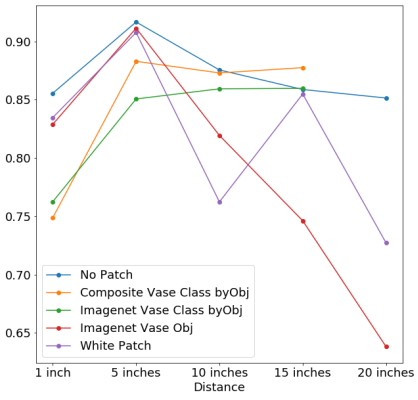
4.3. Dimension Impacts

Recall that for a 1 inch distance, baseline model performance was particularly poor. However in the ImageNet and white patch conditions model performance **increased** significantly, regardless of lighting or center/ right patch placement. This may indicate that pre-trained YOLOv2 is not robust to large-scaled objects. Somehow, more *generic* occlusions provide YOLOv2 with enough context to make an accurate identification.

Figure 4: Image capture of YOLOv2 detection at 5 distances with white patch under LED lighting.

In addition to poor model/ patch performance for close distances, the scores dip when the target item is 15 inches from the Plexiglas. The dip could be driven by either low baseline model performance at 15 inches, or by poor patch performance. A quick look at Table 1 reveals that baseline model performance also decreases at 15 inches (number of correct detections without the patch is 298 out of the possible 500 for halogen bulbs). When this occurs, the model score decreases since the score is relative to baseline.

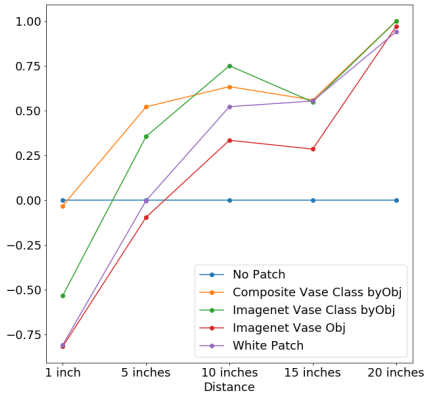
Figure 5: Mean Confidence per Distance



We also recorded confidence scores for each condition to study if there was a relationship with patch performance. Confidence scores did not influence patch performance across the distances. Figure 5 displays the average confidence for each distance and each patch condition. Baseline and ImageNet are the only two patch conditions

that decrease from 5 inches to 20 inches. The two other simulated patches are consistently high for 5 to 15 inches and then effectively hide the target at 20 inches, while the white patch telescopes in performance. One might expect lower confidence scores leading to fewer detection. However, confidence scores are constructed independently of probability of detection within YOLOv2. The model detects a target when the objectness score is above 0.5 and to prevent multiple detections of the same object, the NMS threshold is set to 0.4. The two patch conditions with consistently high confidences when the target was detected are also the two models that are unable to detect the target the most. At 15 inches (where all simulated patches have a sudden decrease in score), the highest scoring patch, Composite (CxO), has a higher confidence when the target is detected than the baseline model. This provides some experimental evidence that confidence score alone and without context to other class scores are not a clear predictor of success for patches designed to hide a target.

Figure 6: [Score by Distance for Each Patch] Points are averaged values for each distance.



An alternative explanation for model performance beyond 10 inches is that the target itself is mostly occluded from camera view at further distances. More than 97% of all frames at this distance were successfully either misclassified or not identified considering all patches. The sub-condition with the highest detection frequency occurred with the white patch placed slightly right of the target under LED lighting. Further testing is required to confirm occlusion is the main reason for poor performance. A counter condition is the ImageNet patch. When that patch was used there were 60 correct detections across lighting at center location, but only a single correct detection when the target was slightly right of the patch. The white patch, LED, right-position condition led to 120 of the 500 frames having correct detections at 20 inches.

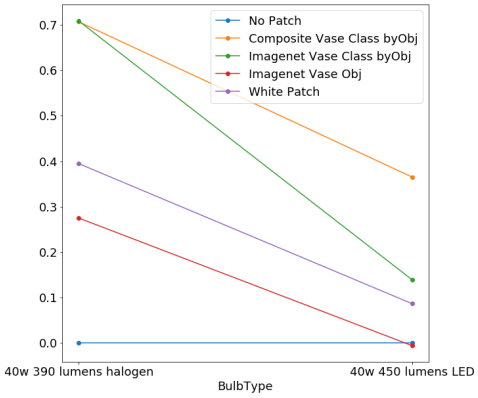
Lighting was also a significant factor for patch scores. Figure 7 displays the difference in lighting used for this experiment. We predicted that for the LED condition, the patches would be more effective at hiding the target. However, the LED condition resulted in fewer disappearances (more correct detections) than the halogen condition. This is surprising given the LED is more luminescent.

Figure 7: Image capture of YOLOv2 detection at 2 light configurations with white patch at 10 inches.

(a) LED Lighting **(b)** Halogen Lighting

Considering only changes in this factor, we find that across the two lighting conditions, the Composite (CxO) and ImageNet (CxO) patches outperform the other two patches. In addition, the white patch has higher scores than the ImageNet (O) patch in both lighting conditions. Performance in the halogen bulb condition is limited to 0.7 by the fact that we are averaging scores across the other dimensions (location and distance) and there are cases in which the baseline model had missed detection occurrences in these conditions.

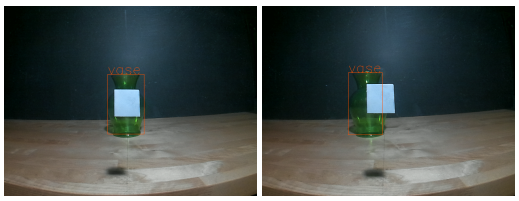
Figure 8: [Score by Light for Each Patch] Points are averaged values for each lighting source.



The same trend, although on a smaller scale, occurs when marginalizing over location values. We computed 68% (roughly two standard deviations) confidence intervals by a bootstrapping procedure. Across the two target locations, the confidence intervals have high overlap which is an indicator that performance is likely equal regardless of target location. The Composite (CxO) and ImageNet (CxO) patches are the only two that had confidence intervals that did not overlap with a score of 0, indicating that for both location conditions, these patches had some effect on

YOLOv2. However, after running one-way ANOVA's, even these patch scores were not significantly different from 0 (p =[0.323, 0.219, 0.969, 0.595] for Composite (CxO), ImageNet (CxO), ImageNet, and white patch respectively).

Figure 9: Image capture of YOLOv2 detection at 2 location configurations (center and right) with white patch at 10 inches under LED lighting.



5. Conclusion and Future Work

This paper makes two contributions. First, we propose a score for adversarial attacks in the physical world. The score compares attack performance to a baseline. We compute the score in a controlled environment with reproducible environmental conditions. We include two potential use cases for the proposed score. The second contribution is that, to the best of our knowledge, this is the most systematic assessment of adversarial attacks to date. Chen et al. [5] had an indoor systematic assessment of sticker attacks but did not investigate varying controlled lighting and placement, rather distance and angle. While many of the current papers highlight that their methods and patch generation methods work well in the real world, it is of importance to account for the weaknesses in any method to not only prevent other researchers from making the same mistakes, but to advance scientific understanding of deep learning models in general. Moreover, we can empirically conclude that camera aspects and model training are interacting with environmental conditions to produce odd model results (such as the baseline model not detecting a vase five and fifteen inches in front of the camera). Our aim with the approach taken was to strive towards a full report of both the model and the adversarial object in the real-world and to highlight in detail the challenges researchers face when evaluating adversarial objects.

References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017. 1, 2

[2] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 1

[3] Nicholas Carlini. A complete list of all (arxiv) adversarial example papers. https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html. 1

[4] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 1

[5] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer, 2018. 1, 7

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[7] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramér, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Physical adversarial examples for object detectors. *arXiv preprint arXiv:1807.07769v2*, 2018. 1

[8] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning models. *arXiv preprint arXiv:1707.08945*, 2017. 1

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

[10] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 1

[11] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018. 2

[12] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015. 1

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[14] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 1

[15] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387. IEEE, 2016. 1

[16] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016. 3

[17] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[18] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Adverserial yolo. https://gitlab.com/EAVISE/adversarial-yolo. 2

[19] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 1, 2

[20] Yue Zhao, Qintao Shen, Ruigang Liang, Kai Chen, and Shengzhi Zhang. Practical adversarial attack against object detector. *arXiv preprint arXiv:1812.10217*, 2018. 1, 2