Shoplifting Smart Stores Using Adversarial Machine Learning

Preprint · July 2019

DOI: 10.13140/RG.2.2.14902.24644

5 authors, including:

Mohamed Nassar
American University of Beirut
68 PUBLICATIONS **328** CITATIONS
SEE PROFILE

Abdallah Itani
American University of Beirut
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Mahmoud Karout
American University of Beirut
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Mohamad El Baba
American University of Beirut
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Shoplifting Smart Stores Using Adversarial Machine Learning

Mohamed Nassar, Abdallah Itani, Mahmoud Karout,
Mohamad El Baba, Omar Al Samman Kaakaji
Department of Computer Science
Faculty of Arts and Sciences
American University of Beirut (AUB)
Beirut, Lebanon
Email: mn115@aub.edu.lb, [ami23—mtk09—mme111—oja05]@mail.aub.edu

***Abstract*—Smart stores cashier-less technology is partially based on camera-equipped object detection systems. Powerful machine learning algorithms are deployed in the back-end for classification. In this paper, we explore the usage of adversarial machine learning techniques to deceive the smart stores' classifiers. In particular, we experiment with printable adversarial patches and target making an expensive item classified as a cheaper one. By sticking patches to the objects and lifting them, a customer can make her customized discounts and alter the machine learning prediction. We discuss experiments, results, and possible countermeasures.**

***Index Terms*—Smart Stores, Adversarial Machine Learning, Deep Learning, Classification, Convolutional Neural Networks, Object Recognition**

I. Introduction

In today's world, everything is getting automated and led by Artificial Intelligence (AI), from self-driving cars to industrial pipelines and robotics. Grocery stores are following the trend. Smart stores, Amazon Go in the forefront, are on the rise. They offer excellent benefits and an unparalleled shopping experience. These stores aim at reducing human employees to 0% and making queuing to check out a thing of the past. However, with this new technology come new challenges. The research question that we address in this paper is: "Can smart stores' AI classification models be deceived by Adversarial Machine Learning (AML)?" We do not experiment with real smart stores, though, and our goal is undoubtedly not shoplifting by itself. We want to study AML scenarios and their countermeasures. From another side, Amazon policy seems to be tolerant with accidental or unintended shoplifting [1]. AML-guided shoplifting is not reported yet, but it is interesting to see such cases in the near future, and whether the countermeasures would be based on runtime detection or rather long-term customer profiling.

II. Background

A. Smart Stores Technology

The technology behind automated smart stores is not fully revealed yet [2], [3]. However, upon entering an Amazon Go store, one can notice hundreds of cameras and infrared sensors. These devices are used to track the movement of every customer. The shelves have weight scales that aid in determining whether a customer had picked up a product or was simply checking it out. The technology is still not perfect. In tens of recorded cases, the classification system failed to detect some objects or charged some items to the customer in error [1], [4].

B. Adversarial Machine Learning

Neural networks, deep convolutional ones in particular, are currently the best machine learning algorithms at object classification. However, these models have been shown to be vulnerable to many adversarial attacks. By slightly changing the pixel values of an image, one can lead a convolutional neural network to make wrong predictions [5]. Attacks can be divided into model-specific (white box) and model-agnostic (black box). Model-specific assumes access to the gradients of the learning model [6]. Model-agnostic techniques transact with classifiers as black boxes and do not require internal knowledge of the model or the training data [7]. Some attacks are very stealthy and unperceived to the human eye, such as changing only one pixel in the input image [8]. Other ones can be detected by humans but not by automated ML. Some attacks intervene at an advanced stage in ML pipelines, for example, after image acquisition. Other ones occur just at the beginning and are merely physical. For instance, adversarial 3D printed artifacts are shown to deceive a camera-equipped detection system [9]. For a summary of AML history in the last ten years, we refer the reader to [10].

III. Shoplifting using Adversarial Patches

A. Adversarial Patch

An adversarial printable patch that can be stuck on objects is designed in [11]. Basically, given an image $x \in \mathbb{R}^{w \times h \times c}$ of class y and a patch image p (initialized to 0 pixels), a modified patch $\hat{p}$ is sought such as rotating and scaling $\hat{p}$, and placing it at a location l in the image x would maximize the probability of classifying the composed image as a target class $\hat{y}$. The patch is made robust by randomizing over a set of images, a set of locations and a set of transformations as described in [9]. The patch is trained to optimize the following objective function:

$$\hat{p} = \arg\max_{p} \mathbb{E}_{x \sim X, t \sim T, l \sim L}[\log \Pr(\hat{y}|A(p, x, l, t)]$$

TABLE I
SHOPPING LIST

Item	Price	Target item	Target price
Vaseline Lotion	5$	Pomegranate	1$
Hair styling foam	3.75$	Orange	0.46$
Bottle of Wine	53.09$	Banana	0.20$

where X is the set of images, T is the set of transformations (scaling and rotation), and L is the set of locations. A is the patch application operator.

B. Shoplifting examples

In one scenario, one can simply wear a smartwatch displaying the patch and take out the object. As shown in Figure 1 the watch displaying a toaster patch is classified as a toaster, and it may trick the store ML.

In another scenario, the customer prepares a list of purchases and assigns a cheaper target for each item. An example is shown in Table I. In this example, the total price without shoplifting is:

$$5 + 3.75 + 53.09 = 61.84\$$$

with successful shoplifting, the price becomes :

$$1 + 0.46 + 0.20 = 1.66\$$$

and the total savings are $60.18\$$

The next step is to prepare the patches for the misclassification targets, as shown in Figure 2. Note that: (1) patches alone would not be classified correctly as the target classes. A banana patch would not be labeled as banana by the ML, but the composed image does, and (2) Patching with an authentic image, an orange image for example, does not lead to the desired outcome. However, a trained patch as small as 10% of the salient object size demonstrates an excellent deception. Figure 3 shows how ML can be deceived.

Our experiments are divided into two stages: (1) lab tests using 2D images, and (2) field tests by printing the patches and sticking them to the objects. We experimented with physical settings under the Demitasse application[1]. Demitasse is an IOS app that supports the VGG16 pre-trained classification model over the ImageNet dataset (1000 labels).

However the adversarial patch technique is not specific to this application, and it shows good transferability from one classifier to another. We targeted other DNN in model-specific and model-agnostic modes. A universal patch is one that works regardless of the recognition model, but also the background item, its location, rotation, or scaling. These latter attributes depend on the camera location and orientation, which make them mostly out of control for the tester.

The patches in Figure 2 are created in white box mode against the pre-trained VGG16 DNN. VGG16 is a deep convolutional network for large-scale image recognition [12].

[1]https://itunes.apple.com/us/app/demitasse-image-recognition-cam/id1138211169

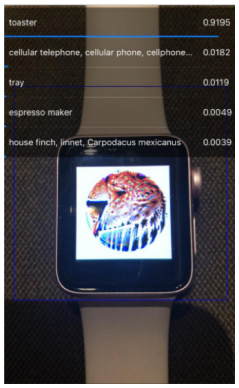
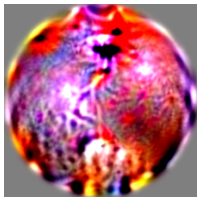
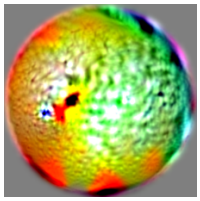


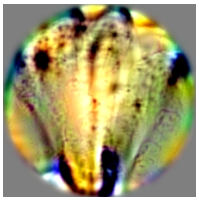
Fig. 1. Smart watch displaying a patch can deceive the classifier



(a) Pomegranate patch



(b) Orange patch

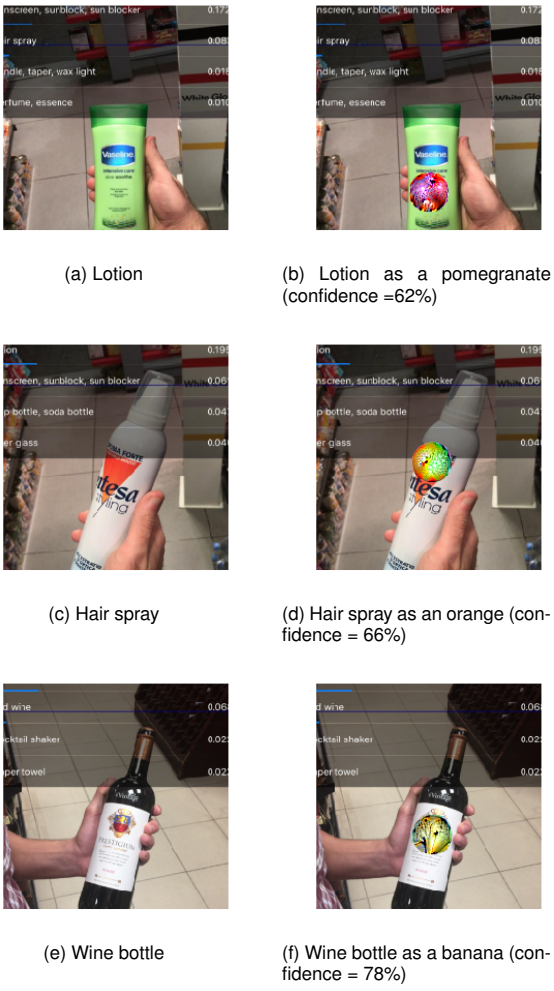


(c) Banana patch

Fig. 2. Preparing patches for shoplifting (In whitebox mode against VGG16)

In [11], experiments with a black box mode are performed, where the patch is trained using four ImageNet neural networks and tested against an unseen fifth one. The models are: inceptionv3, resnet50, xception, VGG16 and VGG19, and the dataset is ImageNet. However, in shoplifting, the attack is much more successful if the smart store uses a classification model known to the attacker, and if the final prediction is solely based on the camera-equipped classifier and without additional sensory inputs. We make the attack much more robust by training against apriori known backgrounds and salient objects as given by the purchase items list.

In our experiments, we find that a ratio of 10% of the patch size to the original image size is convenient. Printed patches with bright colors, such as the pomegranate patch, work better in physical world settings. We share two colab notebooks: the first one is to train adversarial patches and is available at https://colab.research.google.com/drive/1VOZIiNup3FkNwKSieqpDomEL7OZ4sWlB , the second one is to experiment with LIME explanations and is available at https://colab.research.google.com/drive/10o0ysXgYRet1ScXGBpkJz0BSOGcolLL9 . The two colabs are adapted from the notebooks of the original papers [11]



(a) Lotion

(b) Lotion as a pomegranate (confidence =62%)

(c) Hair spray

(d) Hair spray as an orange (confidence = 66%)

(e) Wine bottle

(f) Wine bottle as a banana (confidence = 78%)

Fig. 3. Shoplifting in action (Patch size = 10% of original image size), target classifier = vgg16

and [13], respectively. Slight modifications were required for compatibility with tensorflow v1.13.

We assessed the performance of the adversarial patch approach over a new image dataset, which is more suitable to our context. We have used the grocery store dataset from [14] (https://github.com/marcusklasson/GroceryStoreDataset.git), the part `/dataset/train/Packages/` in particular. Training is performed over batches of 16 random images at each learning step. A batch example is shown in Figure 4.

The performance of the single model attack rates for the pomegranate patches is shown in Figure 5a. Figure 5b shows the performance of the transfer attack where the target is the vgg16 model and the trained patch optimizes the expectation over the ensemble of the other four models. Figure 5c shows the performance of the ensemble white box patch against the different DNNs. We notice an improvement over the single model attack for some target DNNs. For a baseline true pomegranate, the attack won't be that successful as shown in Figure 5d.

IV. Countermeasures

In countermeasures, we estimate that an automated explanation of the prediction would play a key role in defeating the attacks. Local Interpretable Model-Agnostic Explanations (LIME) [13] is a system that generates such explanations.

LIME works by approximating a classifier f in the vicinity of a data instance x. By minimizing a loss function using a gradient descent approach, LIME finds a function g among a set of interpretable functions G. The function g must be of moderate complexity and very similar to the function f in the neighborhood of the predicted instance. The interpretability of g allows the selection of the features, which are the real cause of the prediction.

Lime support several data formats such as images, text, and relational data. For an image, the explanation features take the form of a set of megapixels that highlight a selected area in the original image.

Given a patched item, LIME can successfully locate the patch region as the explanation of the classifier inference. The attack detection becomes easily perceivable and possibly automated. Our intuition is that the explanation of a patched item would differ much from the explanations of clean, unpatched ones. The defender may store explanations for each class in the form of a one-class cluster and uses an outlier/anomaly detection with a simple distance metric to detect weird or inconsistent explanations. In the aforementioned purchase list example, the explanations of the original objects and the patched objects are shown in Figure 6. As shown in the figure, the patch explanations are centered around the patch locations and very different than the ones corresponding to the salient objects.

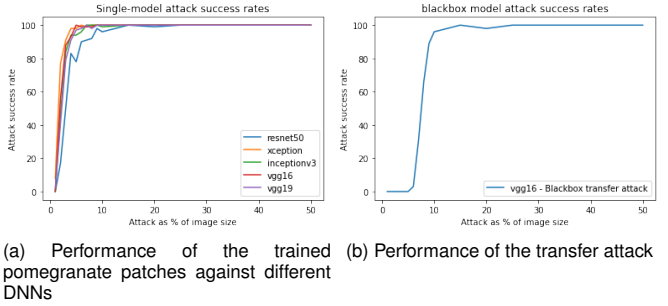
More recently, a defense technique based on detecting high-frequency changes at a particular image location is proposed in [15]. These changes are caused by introducing the patch at some location in the image. The proposed approach is to regularize gradients in the noisy region using Local Gradients Smoothing (LGS) before feeding the image to the deep neural network for inference. We intend to explore this technique in more details in future work.

V. Related Work

Recent studies show that the state-of-the-art deep neural networks (DNNs) for visual classification are vulnerable to physical-world attacks, which are caused by small-magnitude perturbations added to the input. In road sign classification, high targeted misclassification rates (84.4%) were achieved in the physical world under various environmental conditions and viewpoints [16]. With the help of trained white and black stickers, a real stop sign can be misclassified in the captured video frames of a moving vehicle.

Adversarial machine learning and lack of explanations are currently important challenges that hinder the deployment of ML in critical environments. An explainable AI initiative is launched by the United State Defense Advanced Research

Fig. 4. Example of a training batch



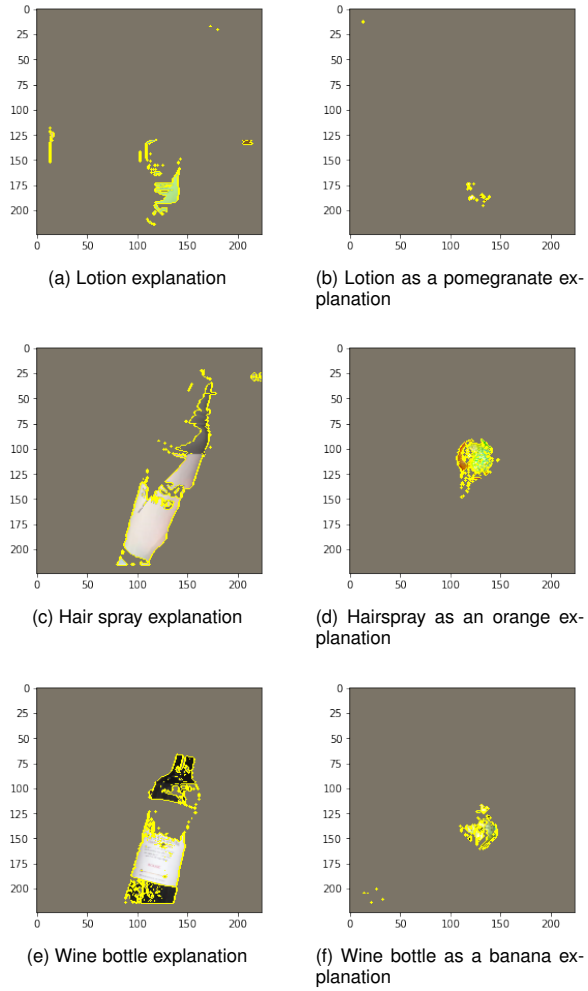
(a) Performance of the trained pomegranate patches against different DNNs

(b) Performance of the transfer attack

(c) Performance of the white box ensemble attack

(d) Performance of true baseline pomegranate

Fig. 5. Attack performance results



(a) Lotion explanation

(b) Lotion as a pomegranate explanation

(c) Hair spray explanation

(d) Hairspray as an orange explanation

(e) Wine bottle explanation

(f) Wine bottle as a banana explanation

Fig. 6. LIME explanations (5 features)

Projects Agency (DARPA) [17]. Regulations such as the European Union's new General Data Protection Regulation (GDPR) [18] require machine learning algorithms to be explainable.

In response, efforts are being made to make deep learning much more trustworthy and controllable by humans [19]. Interpretable machine learning algorithms and taxonomy are summarized in [20].

Our paper explores the relationship between versatility and explainability in the context of visual classification in smart stores. To our knowledge, this is the first attempt to explore AML for shoplifting smart stores and its possible countermeasures.

VI. Conclusion

In this paper, we have experimented with adversarial patches as a futurist possible way to shoplift smart stores. A patch as small as 10% of the camera capture can deceive the ML. We reported on experiments and discussed countermeasures. We showed that explainability of machine learning, and explainable artificial intelligence in general, may play a key role in the defense against adversarial machine learning.

REFERENCES

[1] N. Statt, "Amazon doesnt care if you accidentally shoplift from its cashier-less store," https://www.theverge.com/2018/1/22/16920784/amazon-go-cashier-less-grocery-store-seattle-shoplifting-punishment-detection.

[2] M. Tillman, "What is amazon go, where is it, and how does it work?" https://www.pocket-lint.com/phones/news/amazon/139650-what-is-amazon-go-where-is-it-and-how-does-it-work.

[3] B. McBeath, "Amazon go and the emergence of sentient buildings: How it works and what its impact will be," http://www.clresearch.com/research/detail.cfm?guid=6A608036-3048-78A9-2FB3-4E6295D65919.

[4] Linus Tech Tips, "We stole tampons from the cashier-less amazon go store," https://www.youtube.com/watch?v=vorkmWa7He8&t=601s.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

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

[7] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM, 2017, pp. 506–519.

[8] J. Su, D. V. Vargas, and S. Kouichi, "One pixel attack for fooling deep neural networks," *arXiv preprint arXiv:1710.08864*, 2017.

[9] A. Athalye and I. Sutskever, "Synthesizing robust adversarial examples," *arXiv preprint arXiv:1707.07397*, 2017.

[10] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[11] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[14] M. Klasson, C. Zhang, and H. Kjellström, "A hierarchical grocery store image dataset with visual and semantic labels," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[15] M. Naseer, S. Khan, and F. Porikli, "Local gradients smoothing: Defense against localized adversarial attacks," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1300–1307.

[16] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[17] D. Gunning, "Explainable artificial intelligence (xai)," *[Online]. Available: https://www.darpa.mil/program/explainable-artificial-intelligence*, 2016.

[18] G. D. P. Regulation, "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46," *Official Journal of the European Union (OJ)*, vol. 59, no. 1-88, p. 294, 2016.

[19] J. Choo and S. Liu, "Visual analytics for explainable deep learning," *arXiv preprint arXiv:1804.02527*, 2018.

[20] C. Molnar, *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/, 2018.