

Capture the Bot: Using Adversarial Examples to Improve CAPTCHA Robustness to Bot Attacks

Dorjan Hitaj*, Briland Hitaj†, Sushil Jajodia‡ and Luigi V. Mancini*

*Dipartimento di Informatica, Sapienza University of Rome, Italy

Email: {hitaj.d, mancini}@di.uniroma1.it

†Computer Science Laboratory, SRI International

Email: briland.hitaj@sri.com

‡Center for Secure Information Systems, George Mason University

Email: jajodia@gmu.edu

Abstract—To date, CAPTCHAs have served as the first line of defense to prevent unauthorized access by (malicious) bots to web-based services, while at the same time maintaining a trouble-free experience for human visitors. However, recent work in the literature has shown that sophisticated bots using advancements in Machine Learning (ML) can easily bypass existing CAPTCHA-based defenses. This work introduces CAPTURE, a novel CAPTCHA scheme based on adversarial examples. Typically adversarial examples are used to lead an ML model astray. With CAPTURE, we attempt to make a “good use” of such mechanisms in order to increase the robustness and security of existing CAPTCHA schemes. Our empirical evaluations show that CAPTURE can produce CAPTCHA challenges that are easy for humans to solve, while at the same time, CAPTURE can effectively thwart sophisticated ML-based bot solvers.

Index Terms—Security and Privacy, CAPTCHA, Machine Learning, Adversarial Examples

1 INTRODUCTION

Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) is a type of challenge-response test in computing applications which is used to distinguish between humans and machines.

Usually CAPTCHAs are generated by distorting an image that contains text and numbers in such a way that only a human can properly read or understand the content of that image. However, current advancements in Machine Learning (ML), particularly in the domain of Deep Learning (DL), pose a significant threat to existing CAPTCHA mechanisms. Deep Neural Networks (DNNs), the core component of deep learning, have even surpassed human capabilities in certain tasks, such as image and speech recognition. Recent research also shows that DNNs are able to bypass existing CAPTCHA-based defenses, opening the doors to automated attacks on virtually any online service that relies on CAPTCHA as a defense mechanism.

To counteract this phenomenon, CAPTCHA designers are trying to make the CAPTCHA task harder to solve by introducing additional security features. However, most of these additional security features end up impacting the user experience as well. Some types of CAPTCHAs, such as the Google image-selection CAPTCHA, are becoming increasingly hard to solve even for humans due to the difficulty of recognizing the small details shown in the challenge images. In turn, this additional burden leads to frustration whenever users are presented with a complex CAPTCHA that requires a significant concentration to be solved, just to be able to access the desired online service.

Our paper introduces a novel technique, called CAP-

TURE - CAPtcha Technique Uniquely REsistant, based on adversarial machine learning to improve the resilience of image-selection CAPTCHAs against automated solvers, while at the same time maintaining a positive user experience in the CAPTCHA solving process. Specifically, we investigate the “good use” of unrecognizable images [1] and adversarial patch [2] to design an improved image-selection CAPTCHA. *We show that unrecognizable images and adversarial patches can be used to alter images without affecting human-based recognition and user experience, while at the same time leading DNN-based automated solvers to misclassify the images with high confidence.* Differently from traditional adversarial perturbations [3], unrecognizable images and adversarial patches are highly transferable between various DNN model architectures [4], which makes these two techniques good building blocks to design a general countermeasure against machine learning based CAPTCHA solvers.

This paper focuses on DNN-based automatic solvers, since DNN models have proven to be the go-to machine learning technique to handle complex tasks such as image recognition, thus being among the most widespread techniques employed by automatic solvers to solve a CAPTCHA challenge.

We conduct security and usability experiments for our novel CAPTURE technique. Our experiments show that the CAPTURE challenges are hard for bots to solve, but are easily solved by humans in a short time.

2 BACKGROUND

2.1 (Deep) Neural Networks

Neural networks are a set of algorithms that focus on determining underlying relationships existing in the input data in a relatively similar way to the human brain. Deep neural networks (DNN) differ from traditional single-layer neural networks by their number of hidden layers. Typically, a DNN consists of two or more hidden layers. Stacking more layers in a DNN enables the later to identify and extract more complex features from the input data. This hierarchy-based structure enables DNNs to process tremendous amounts of high-dimensional data, such as images, with impressive results both in classification and clustering tasks. Despite the superhuman capabilities that DNNs have shown on performing certain tasks, they are susceptible to adversarial attacks that can deteriorate their performance significantly. This vulnerability towards adversarial attacks makes entities approach the domain with care even though the benefits of using DNNs seem undoubtedly large. Even though this property of DNNs seems rather negative, we attempt to make a good use out of it, by leveraging it to introduce security features that make current CAPTCHA schemes harder to solve by automated CAPTCHA solvers that leverage DNNs in the CAPTCHA solving process.

2.2 Adversarial Examples

Adversarial examples are maliciously crafted inputs whose purpose is to lead a machine learning model into misclassifying the input. They are intended to be stealthy in a such a way that the perturbation that will be added to a legitimate input instance will not lead a human into misclassifying it. These malicious inputs were first observed by Szegedy et al. [3]. Since that time, extensive research is done on both directions, introducing new adversarial example generation techniques and new protection techniques against them. Moreover, extensive work has been done into creating models that are robust to adversarial perturbations, such as adversarial training [5]. Adversarial training is a training procedure that makes ML models more robust to adversarial examples, but not completely safe from them. To this date, there is no universal defense against adversarial examples. All the published defenses have been taken down shortly after by new adversarial perturbation generation methods that are stealthier and stronger.

Typically, adversarial examples are generated by exploiting the internals of a machine learning model, which makes them more effective and stealthier. Moreover, Papernot et al. [4] observed that some adversarial examples are also transferable, meaning that adversarial examples that affect one ML model often affect another ML model, even if the two models have different architectures or were trained on different training sets, as long as both models are trained to perform the same task. This means that an attacker may train their own substitute model, craft adversarial examples against the substitute, and transfer them to a victim model, with very little information about the victim.

Inspired by the transferability property, Shi et al. [6] present a methodology to harden CAPTCHAs by using adversarial perturbations in order to make ML-based Bots unable to solve the presented CAPTCHA challenge. They

present a new adversarial perturbation technique that adds perturbations in the frequency domain rather than in the space domain. According to their analysis, the perturbations added in the space domain are frail to image pre-processing and are considered as local change to images while the perturbations added in the frequency domain are a global change, thus more difficult to remove. Even though this technique is resilient to various perturbation removal methods, it still suffers from the inconsistency of the transferability property among various ML models. To this end, we aim at employing methods that allow us to create highly transferable instances in order to provide a novel CAPTCHA scheme that is more resilient to Bots. Moreover the methods introduced in this paper, enable us to also improve the user experience in the CAPTCHA solving process.

2.3 Unrecognizable Images

Nguyen et al. [1] show a related result to adversarial perturbations [3]. Instead of introducing human-imperceptible perturbations that will make a machine learning model misclassify an instance, the authors of [1] efficiently produce images that are completely unrecognizable to humans, but that DNNs believe to be recognizable images with 99% confidence. Such unrecognizable images are constructed by using Evolutionary Algorithms (EAs) [7] or gradient ascent.

EAs are optimization algorithms inspired by Darwinian evolution. In biological terms, EAs consist of a population of “organisms” that alternately face selection in terms of keeping the best. New organisms are produced by the mutation and crossover amongst best organisms. In this case, the organisms are images and the perturbations, added to them step by step to reach the target, represent the process of mutation and crossover. The selected organisms depend on the fitness function, which, in the case of creation of images that are able to fool a machine learning classifier, is the highest confidence score the target ML model gives to that image for belonging to a specific class. Specifically, Nguyen et al. [1] use two types of encoding to produce the adversarial images, called: *direct* [8] and *indirect* [9] encoding.

EAs with direct encoding: Each pixel value is initialized with uniform random noise within the [0, 255] range. Those pixel values are independently mutated. The amount of pixel values to be mutated starts at 10% and halves after every 1000 generations. The pixel values chosen to be mutated are then altered via the polynomial mutation operator [10] with a fixed mutation strength of 15.

EAs with indirect encoding: By using indirect encoding, elements in the genome can affect multiple parts of the image [11]. Specifically, the indirect encoding used by Nguyen et al. [1] is a Compositional Pattern-Producing Network (CPPN), which can evolve complex, regular images that resemble natural and man-made objects [11], [12]. Unrecognizable images generated using indirect encoding are more regular as opposed to the ones produced with direct encoding, which look a lot like random noise.

Another method used to generate unrecognizable images is by performing *gradient ascent*, which involves updating an image according to the information taken by

computing the gradients of the neural network the image is intended to fool.

In this paper, we modify the above-mentioned image generation approaches in order to create images that simultaneously fool a *multitude* of diverse and high quality DNNs trained to solve a complex image recognition task such as ImageNet [13] which consists of images belonging to 1,000 different classes. By creating images that fool a wide array of high quality ML models, we believe that the fooling image is able to fool even unseen classifiers that are trained to solve a similar task as the ones used to generate the image. We show that our method allows us to create unrecognizable images that are able to transfer, therefore to deceive, even unseen classifiers, thus being a promising approach to build a stronger CAPTCHA challenge.

2.4 Adversarial Patch

Adversarial Patch [2] is a work closely related to [1] and presents another method to fool DNN classifiers. Similar to [1], the authors do not limit themselves into producing human-imperceptible perturbations like adversarial perturbations commonly do. Instead, the Adversarial Patch approach generates an image-independent patch that is extremely salient to a neural network. This patch can then be placed anywhere within the field of view of the classifier, and causes the classifier to output the target class. The generated patch is scene-independent, and allows the authors to deceive the ML image classifiers without prior knowledge of the lighting conditions, camera angle, type of classifier being attacked, or even the other items within the scene.

The adversarial patch solution works by completely replacing a part of the image with a generated patch. An input patch is masked so that it takes any shape, and then it is modified via gradient descent over a variety of images, applying random translation, scaling, and rotation. The generated patch after these operations will be used as an adversarial patch to trick the ML classifiers. In particular, for a given patch p , image x , patch location l , and patch transformations t (e.g. rotations or scaling) the authors define a patch application operator $A(p, x, l, t)$ which first applies the transformations t to the patch p , and then applies the transformed patch p to the image x at location l . To obtain the trained patch \hat{p} they use a variant of the Expectation over Transformation (EOT) framework of Athalye et al. [14]. In particular, the patch is trained to optimize the objective function:

$$\hat{p} = \underset{p}{\operatorname{argmax}} E_{x \sim X, t \sim T, l \sim L} [\log \Pr(\hat{y} | A(p, x, l, t))]$$

where X is a training set of images, T is a distribution over transformations of the patch, and L is a distribution over locations in the image. This expectation is over images, thus encouraging the trained patch to work regardless of what is in the background. The Adversarial Patch solution exploits the way image classification tasks are constructed. While images may contain several items, only one target label is considered true, and thus the network must learn to detect the most "salient" item in the frame.

The ability of the adversarial patch to impact the prediction a classifier gives to an image containing it, is highly related to the portion of the image the patch occupies. Considering that, we use adversarial patches of various sizes in our CAPTCHA scheme in order to keep both the human usability and CAPTCHA security high. As we show in the coming sections, the use of adversarial patches to strengthen image-selection CAPTCHAs is one of the strongest techniques in fooling machine learning based CAPTCHA solving bots, and also one of the most cost-efficient for the organization that uses the CAPTCHA against bots.

3 CAPTURE: A NOVEL CAPTCHA SCHEME

Our novel CAPTURE solution is based on unconstrained adversarial perturbation techniques. These techniques allow us to build CAPTCHA schemes that are able to fool ML-based bots without affecting humans ability to solve the challenge. We identify two similar techniques to generate unconstrained adversarial examples and show how they can be used individually to build a strong CAPTCHA. We further present a more complex approach that allows us to incorporate these techniques together. Our proposed CAPTURE technique heavily increases the number of security features that a bot must evade in order to solve the CAPTCHA challenge, increasing the cost required to achieve a successful attack, while keeping the same level of effort for the human users.

3.1 Threat Model

In our threat model, the adversary employs an automatic solvers that is incorporating machine learning classifiers to break image-selection CAPTCHA schemes. The adversary has no knowledge (black-box) on the CAPTCHA challenge generation procedure.

The malicious Bot may employ several high quality ML-based classifiers that are trained to distinguish among the categories of the images which are part of the CAPTCHA challenge.

The adversary is presented with a set of images in the CAPTCHA challenge and, according to the answers its classifiers gives, it decides whether to click on one or more specific images which it believes to be the answer to the challenge.

Finally, we assume that the adversary has extra tools to understand the textual description of the CAPTCHA challenge that a human should respond in order to pass the test.

3.2 Unrecognizable Image CAPTCHA

The first adversarial perturbation technique we apply is unrecognizable images [1]. Unrecognizable images are images that can lead machine learning classifiers into recognizing actual objects, such as a flag or curtain, when in fact the image itself only contains some peculiar visual patterns that do not resemble any real-life object. There are three main methods that can be used to construct unrecognizable images: Evolutionary Algorithms (EAs) with *direct* encoding, Evolutionary Algorithms with *indirect* encoding, and gradient ascent. In the following, we attempt to generate



Fig. 1. The challenge presented to the user is: *Select all the choices that show a real image of THEATER CURTAINS*. CAPTCHA with indirectly-encoded unrecognizable images.

images that can fool with high confidence *multiple* high quality DNNs, instead of generating images tailored against *one* specific DNN classifier.

EAs with direct encoding: Unfortunately, our results showed that EAs with direct encoding cannot successfully create unrecognizable images for every target class considered. Moreover, the instances that we created using EAs with direct encoding were vulnerable to simple image processing techniques, such as resizing and blurring. Due to these results, we conclude that unrecognizable images generated via direct encoding are not good candidates to build a strong CAPTCHA scheme.

EAs with indirect encoding: By using Compositional Pattern Producing Network (CPPN) [11] encodings, it is possible to generate unrecognizable images that fool many neural networks with high confidence. A CAPTCHA instance using unrecognizable images generated with indirect encoding is shown on Figure 1. In the CAPTCHA challenge illustrated in Figure 1, the task is to select all the images showing a theatre curtain. Out of 9 images there is only one image of a real theatre curtain (third row-third column image from the left). Among the images, we added two unrecognizable images produced via EAs with indirect encoding (first row-second image and second row-third image from the left) that successfully fool multiple, high-quality models trained on the ImageNet [13] dataset.

While this approach is effective in generating unrecognizable images that can fool many different classifiers, it is also very computationally intensive, and the computational cost of the task vary widely based on the specific target class to which you want to assign the unrecognizable image. Nevertheless, such image generation process

can be performed offline. The unrecognizable images can be pre-computed and afterwards positioned to create the CAPTCHA challenges when needed.

Note that, unlike directly encoded unrecognizable images, indirectly encoded images are resilient to common image processing techniques. Moreover, as an extra security feature for the CAPTCHA scheme, we can use adversarial perturbation on the clean images [6] to make it harder for the bot to distinguish between the unrecognizable images and the actual clean images. Indeed, a smart bot could use two classifiers: one to distinguish between unrecognizable and clean images, and then a second one to identify the images required by the CAPTCHA challenge among the clean images only, effectively bypassing our scheme. By applying adversarial perturbation to the clean images we can avoid this attack.

Gradient Ascent: We modify the gradient ascent approach presented in [1], to create images that are able to fool an ensemble of models instead of only one. We successfully crafted unrecognizable images that look like random noise and appear meaningless to humans, but are classified with high confidence in a predefined target class by a neural network. Unrecognizable images generated with gradient ascent are extremely effective, fooling most classifiers with a confidence above 99%. However, these type of images are not easily transferable to neural networks that were not part of the generation procedure. Moreover, they are extremely sensitive to simple image transformations such as resizing. For example, resizing an unrecognizable image from 224×224 to 299×299 does not preserve the properties required to fool the classifier. Considering that the CAPTCHA is displayed in different sizes when viewed with different devices, this heavily reduces the practical applicability of the *Gradient Ascent* approach.

3.3 Adversarially Patched CAPTCHA

This section studies the applicability of adversarial patches [2] to create a CAPTCHA challenge. This approach results to be more computationally efficient than EA with indirect encoding which is the most effective technique that we have identified in the previous section.

Adversarial patches are small patches that can be added in a region of a clean image, triggering the ML model to misclassify the image into a different specific class. Adversarial patches can be generated efficiently and cover only a small portion of the original image, making it an appealing approach to design a resilient CAPTCHA challenge. Moreover, images modified with adversarial patches are still very easily recognized by humans, while DNNs misclassify them with high confidence. To further strengthen our adversarial patch-based CAPTCHA scheme, we specifically design adversarial patches that fool not only a single classifier, but an ensemble of diverse, high quality classifiers such as VGG16 & VGG19 [15], InceptionV3 [16], Xception [17], Resnet50 [18] and MobileNet [19]). Targeting an ensemble of deep neural models, rather than a specific model, greatly increases the chances that the adversarial patch produced will fool other classifiers that might be used by the CAPTCHA-solving bot.

An instance of the CAPTCHA scheme with adversarial patches is illustrated in Figure 2. The images in Figure 2



Fig. 2. Example of an adversarial patched CAPTCHA. The challenge presented to the user is: *Select all the choices that show an image of a FLAGPOLE.*

show various adversarial patches superimposed on pictures of several objects, such as a flagpole, a hammer and a camera. Users are asked to select all the images that include a flagpole. As we can see from the example in Figure 2, this task is fairly straightforward for a human, as the adversarial patches do not cover any important feature of the original images. However, a state-of-the-art DNN would misclassify all these images with high confidence, and would be unable to identify images containing the actual flagpoles. In this example, on the actual images of the flagpole, we superimpose patches that the model classifies as a camera or a hammer, while on all other images we superimpose patches representing a flagpole, thus increasing the chance of the bot to fail in the solution of the challenge presented.

We further integrate unrecognizable images and adversarial patches in a single CAPTCHA challenge, as illustrated in Figure 3. The resulting scheme is far more robust against automated ML-based bots than any single individual approach because the bot has to simultaneously evade multiple security features to solve the challenge. Moreover, combining adversarial patches and unrecognizable images provides better user experience compared to each individual scheme, as illustrated in our usability evaluation in Table 2.

4 EXPERIMENTS

A good CAPTCHA scheme is one that is secure while not compromising the human ability to solve it. The following sections present our experiments both on security and on user experience aspects of our proposed CAPTCHA scheme.

4.1 Security Evaluation

Our security evaluation models the image-selection CAPTCHA as a challenge where images belong to the



Fig. 3. Example of combined unrecognizable images and adversarial patching. The challenge presented to the user is: *Select all the choices that show an image of a FLAGPOLE.*

categories of the ImageNet [13] image recognition dataset. Malicious ML-based Bots could use state-of-the-art ML models to identify the CAPTCHA image categories thus solving the challenge. Our evaluation assumes that the malicious Bots utilize the best performing ML models for image recognition available today, such as VGG16 & VGG19 [15], ResNet50 [18], InceptionV3 [16], Xception [17] and MobileNets [19]. All these ML models have a published accuracy of above 80% in correctly classifying an image to its correct category. Thus, a CAPTCHA challenge based on real images (e.g. without the presence of unrecognizable images or adversarial patches) is vulnerable, and can be solved with 80% accuracy by Bots utilizing the above ML models. Our experiments show that introducing the proposed security features, i.e. unrecognizable images and adversarial patch, we can decrease the accuracy of these advanced ML models almost to zero percent, and significantly reduce the ability of the ML-based Bots to bypass our CAPTCHA challenge.

- *Unrecognizable Images:* We generate unrecognizable images with indirect encoding that are able to fool an ensemble of diverse and high quality DNNs. The neural networks employed on our experiments are: VGG16 & VGG19 [15], ResNet50 [18], InceptionV3 [16], Xception [17] and MobileNets [19]. We generate unrecognizable images that are able to deceive 5 of the above neural networks, and we also assess whether that ability to deceive is transferred to the sixth held out DNN. We generate 1200 unrecognizable images (200 per each ensemble combination). On average the unrecognizable images were able to deceive the held-out DNN around 70% of the time with over 95% confidence in the prediction.

The achieved fooling rates on the unrecognizable im-

ages highlight that the image generation procedure needs to be improved to achieve higher transferability rates.

- *Adversarial Patch*: in the case of adversarial patches we used for the experiments the same set of high quality DNNs as in the case of unrecognizable images. We generate 1200 adversarial patches (200 per each ensemble combination). Then, we evaluate their fooling capability by randomly putting them on 500 different images starting from 10% up to 100% of the image size and evaluate the ability of the patch to fool the remaining neural network to check how the generated adversarial patch transfers to that ML model. Generating an adversarial patch that is able to simultaneously fool multiple diverse and high quality ML models makes the patch capabilities more general and possibly effective even on other unseen ML models, like the one that the malicious Bot might be using.

Figure 4 displays the success rate on each of the unseen models. Figure 4 shows that an adversarial patch scaled

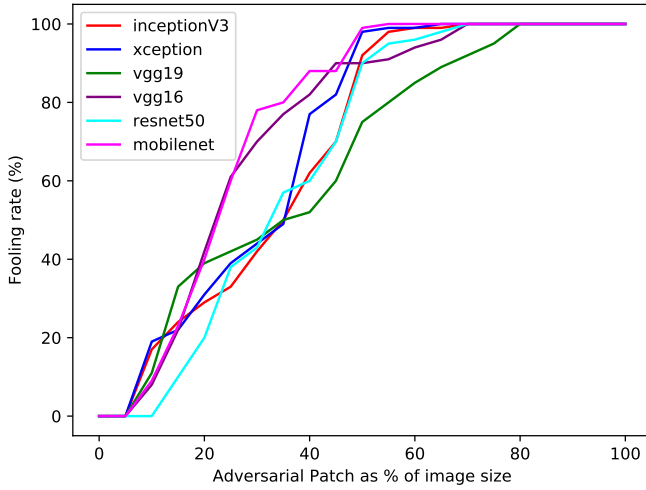


Fig. 4. Adversarial Patch success rates. The lines represent the ability of the adversarial patch of different size (generated on the white-box access models), to fool the black-box model. In the x-axis, the percentage 100% means that the patch is as large as the whole image. With 50% the patch is one quarter of the entire image

to a factor close to 60% is able to fool high quality classifiers almost 100% of the time. In addition, having a patch that covers slightly more than a quarter of the image, provides a better user experience in the solving process also. Since the patch covers a small portion of the image, we can make the object(s) that the user should choose to solve the CAPTCHA challenge, simpler to identify to human users.

4.2 Usability Evaluation

We tested our various configurations of CAPTCHA challenge with a group of users of different age groups. Each person was requested to solve 10 CAPTCHA challenges similar to the examples displayed on Figures 1, 2, and 3. In total, we surveyed 113 people. Table 1 shows the number of participants per age group. In Table 2, we report the success rate of the participants in solving the CAPTCHA

challenges presented. Note that almost all users have been able to successfully solve the CAPTCHA challenges with a success rate of above 85%.

Age	[16-20]	[21-30]	[31-40]	[41-50]	[51-60]	[61+]
Surveyed	1	72	5	12	19	4

TABLE 1
User Statistics

Challenge	Unrec. Images	Adv. Patch	Unrec. + Adv. Patch
Success Rate	96.5%	86%	92.3%

TABLE 2
Usability Statistics. (The reported success rate is the weighted average with respect to the percentage of participants per age group.)

When asked to compare their experience with other image selection CAPTCHAs, the users said our CAPTCHA task was easier to solve because the images the user should click were easy to find. The presence of weird-looking images did not prevent them from solving the task, and most of them said that it was even easier because they could discard whole unrecognizable images, and thus answer the challenge faster.

After carefully investigating the survey results and carrying out post interviews with the surveyed persons, we found out that some mistakes were accountable to people not paying attention to the challenge description. However, this does not impede the applicability of our scheme, as in real-life applications, if human users make a mistake, they are typically presented with another CAPTCHA challenge in order to access the required web-service.

5 CONCLUSIONS

This paper presents a new enhancement of the CAPTCHA challenge called: CAPTCHA - CAPTCHA Technique Uniquely RESistant. The new CAPTCHA scheme is based on a "good use" of the recent advancements in adversarial machine learning. In particular, this paper proposes a novel CAPTCHA scheme that employs two approaches to create human-unrecognizable images that are recognizable by machine learning models as a real object with high confidence. The core advantages compared to traditional image CAPTCHAs include:

- *ML-based Bots cannot easily solve the CAPTCHA challenge*: Using unrecognizable images and adversarial patches, the CAPTCHA scheme can fool the Bot into clicking on images that the human user would never click to solve the challenge. Also, incorporating multiple fooling strategies like unrecognizable images [1], adversarial patches [2], and adversarial perturbations [3] in the same CAPTCHA, we make it harder for the Bot to bypass all the security measures in each of the images that are part of the CAPTCHA challenge presented to it.
- *Improve the user experience while solving the CAPTCHA challenge*: Using unrecognizable images and adversarial patches throughout the images that the user should not click to solve the challenge (but the bots will be prone to click) allows us to put clear images of the objects the user should click thus making their life easier. Current

image CAPTCHAs employ images that are hard to distinguish even to human users which negatively impact the user experience. Our CAPTURE scheme increases the security while also keeping a positive level in terms of user experience in the solving process.


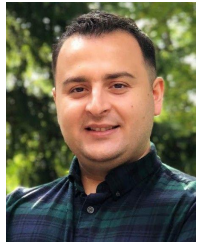


As future work, we intend to make the generation of unrecognizable image more computationally feasible. We also intend to focus on increasing the ability of our mechanism to mislead multiple ML models simultaneously. Based on our users feedback, this would lead to better-looking CAPTCHAs for humans, while also increasing the robustness against ML-based Bot attacks.

6 ACKNOWLEDGMENTS

The work of Dorjan Hitaj and Luigi V. Mancini was supported by Gen4olive, a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101000427, and in part by the Italian MIUR through the Dipartimento di Informatica, Sapienza University of Rome, under Grant Dipartimenti di eccellenza 2018–2022. The work of Sushil Jajodia was supported by the Office of Naval Research grant N00014-18-1-2670 and by the Army Research Office grant W911NF-13-1-0421.

REFERENCES

- [1] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015.
- [2] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *CoRR*, abs/1712.09665, 2017.
- [3] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.
- [4] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.
- [5] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. *CoRR*, abs/1705.07204, 2017.
- [6] Chenghui Shi, Xiaogang Xu, Shouling Ji, Kai Bu, Jianhai Chen, Raheem A. Beyah, and Ting Wang. Adversarial captchas. *CoRR*, abs/1901.01107, 2019.
- [7] Dario Floreano and Claudio Mattiussi. *Bio-Inspired Artificial Intelligence: Theories, Methods, and Technologies*. The MIT Press, 2008.
- [8] Kenneth O. Stanley and Risto Miikkulainen. A taxonomy for artificial embryogeny. *Artificial Life*, 9:93–130, 2003.
- [9] Jeff Clune, Kenneth O. Stanley, Robert T. Pennock, and Charles Ofria. On the performance of indirect encoding across the continuum of regularity. *IEEE Transactions on Evolutionary Computation*, 15:346–367, 2011.
- [10] Kalyanmoy Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Inc., USA, 2001.
- [11] Kenneth O. Stanley. Compositional pattern producing networks: A novel abstraction of development. *Genetic Programming and Evolvable Machines*, 8:131–162, 2007.
- [12] Joshua E. Auerbach. Automated evolution of interesting images. *Artificial Life 13*, 2012. The Humanities and ALife – Best Presentation Award Winner.
- [13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- [14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.
- [15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.
- [16] François Chollet. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017.
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [18] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

- 
- Dorjan Hitaj** is a Ph.D. student in Computer Science with Sapienza University of Rome. His research focus is on machine learning and security. He obtained his MSc. degree in Computer Science with high honors from Sapienza University of Rome in 2018.
- 
- Briland Hitaj** is an Advanced Computer Scientist at SRI International. His expertise is in security and privacy in deep learning systems. Prior to joining SRI International, he was a Visiting Research Scholar at Stevens Institute of Technology. Dr. Hitaj obtained his Ph.D. in Computer Science from Sapienza University of Rome, Italy.
- 
- Sushil Jajodia** is a University professor, BDM international professor and director of Center for Secure Information Systems, George Mason University. Prior to joining Mason, he held permanent positions with NSF, NRL, and University of Missouri-Columbia. He has sustained a highly active research agenda spanning database and cybersecurity for more than 30 years.
- 
- Luigi V. Mancini** is a Full Professor with Sapienza University of Rome, Italy, where he is the chairman of the Master degree in Cybersecurity. He has authored over 130 scientific papers in international conferences and journals. He obtained his Ph.D. degree in Computer Science from the University of Newcastle, U.K., in 1989.