Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Seeing is Deceiving: Fortifying reCAPTCHAv2 through Adversarial Machine Learning

1 Problem Statement and Motivation

CAPTCHAs are vulnerable Google's reCAPTCHA serves as a perimeter security measure that protects websites from credential stuffing, large-scale data scraping and distributed denial of service (DDoS) attacks, among others. Holding an estimated 99.93% market share [1], it is the most widely used CAPTCHA system. However, computer vision model-based attacks have been shown by Plesner et al. [11] to break reCAPTCHAv2 with a 100% success rate. The attack leverages a widely-used pre-trained object detection and segmentation model (YOLOv8) with minimal memory and compute resources, making it accessible to a wide range of adversaries.

We can do better To address this vulnerability, we propose applying perturbations to CAPTCHA images that are imperceptible to humans but cause misclassification in object detection and segmentation models. This proactive defense mechanism leverages adversarial examples, which exploit the gap between human and machine perception, to mitigate the risk of vision-based attacks. The main advantage of this approach is that it will neither require a complete overhaul of the reCAPTCHA system nor compromise the user experience.

We can make an impact This project is timely and relevant as it addresses a critical security vulnerability in the most widely used CAPTCHA system. The proposed approach is practical, cost-effective and scalable, making it an attractive solution for Google and other organizations that rely on reCAPTCHA to secure their online platforms. Based on our literature review, this approach has been hypothesized for a subset of reCAPTCHAv2 tests [4], but no practical implementation has been developed yet. This allows us to fill the gap and make a meaningful contribution to the field of adversarial machine learning and cybersecurity.

2 Aim of the Thesis and Expected Results

Building a better CAPTCHA The main objective of this thesis is to build a CAPTCHA prototype that can withstand vision-based attacks by applying adversarial perturbations to its images. Due to reCAPTCHAv2 being close-sourced, we will have to develop a system that closely mimics its functionality. The prototype will be designed to generate CAPTCHA images, apply imperceptible perturbations to them and verify their effectiveness against vision-based attacks. The perturbations will be generated using adversarial machine learning techniques, such as the Fast Gradient Sign Method (FGSM) [3] and the Projected Gradient Descent (PGD) [9]. The expected results are twofold: (1) the successful implementation of the CAPTCHA prototype and (2) the demonstration of its effectiveness in mitigating vision-based attacks. We will evaluate the prototype's performance based on the success rate of the vision-based attacks and the usability of the CAPTCHA images for human users.

Understanding the "dimpled manifold hypothesis" Since the adversarial vulnerability of deep neural networks was discovered in 2013 [3], there have been many attempts to explain why adversarial examples exist and how they work, each with their limitations and assumptions – some complementary, some contradictory [1]. And there are still many open questions. One of the hypotheses is the "dimpled manifold hypothesis" [2] proposed by Shamir et al. [12], suggests that the decision boundary of deep neural networks is close to the data manifold, making it easy to find adversarial examples. Additionally, the paper found that by reducing the dimensionality of the perturbations and projecting them on the data manifold before passing them to the model, they can be made perceptible and interpretable to humans. We aim to explore this hypothesis further and conduct experiments to test its validity building upon the work of Karner et al. [6]. We expect to gain a deeper understanding of the underlying mechanisms of adversarial examples that will help with the development of our CAPTCHA prototype – although not our primary focus and just an **optional extension** to the project.

[1] For a comprehensive overview of the hypotheses, see the Addendum of Ilyas et al. [5]

[2] A similar idea was previously proposed by Elliot et al. [2]

3 Methodology

The detailed project outline is as follows [3]:

- Literature review (related work, previous approaches) $(\star\star\star)$
- Building a reCAPTCHAv2 clone $(\star\star\star\star)$
- Building a second version of the reCAPTCHAv2 clone that is robustified with adversarial perturbations $(\star\star\star)$
- Evaluating the effectiveness of vision-based attacks on both versions of the reCAPTCHAv2 clone $(\star\star\star)$
- Writing the final report / thesis $(\star\star\star\star\star)$
- Midterm and final presentations $(\star\star\star\star)$

The student's duties include:

- One meeting per week with the advisors to discuss current matters
- A final report in English, presenting work and results
- A midterm and a final presentation (15 min) of the work and results obtained in the project

Optional extensions to the project if time allows include:

- Turning the CAPTCHA prototype into a reusable open-source attack and defense benchmarking tool
- Evaluating perturbed CAPTCHAs against humans (in addition to vision-based models)
- Assessing previous work on the dimpled manifold hypothesis
- Formalizing falsifiable hypotheses for the dimples paper and conducting experiments to test them (based on: [8], [7], [6])
- Suggesting a new information-theoretical hypothesis for adversarial examples based on the dimples paper (with the help of Alireza Furutanpey)
- Trying to make adversarial examples perceptible to humans by projecting them on the data manifold (based on: [12])
- Exploring distillation learning to improve adversarial robustness (based on: [10], [9])

[3] The stars indicate the estimated effort required for each task on a scale from 0 to 5 (0 = no effort, 5 = high effort)

References

[1] 6sense. Google Captcha Market Share. https://6sense.com/tech/captcha/recaptcha-market-share#:~:text=What%20is%20reCAPTCHA%20market%20share,of%2099.93%25%20in%20captcha%20market, 2023. [Online; accessed 17-July-2024].

[2] Andrew Elliott, Stephen Law, and Chris Russell. Explaining classifiers using adversarial perturbations on the perceptual ball. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10693–10702, 2021.

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

[4] Dorjan Hitaj, Briland Hitaj, Sushil Jajodia, and Luigi V Mancini. Capture the bot: Using adversarial examples to improve captcha robustness to bot attacks. *IEEE Intelligent Systems*, 36(5):104–112, 2020.

[5] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[6] Lukas Karner. The Dimpled Manifold Revisited. https://github.com/LukasKarner/dimpled-manifolds/, 2023. [Online; accessed 17-July-2024].

[7] Yannic Kilcher. dimple test. https://gist.github.com/yk/de8d987c4eb6a39b6d9c08f0744b1f64/, 2021. [Online; accessed 17-July-2024].

[8] Yannic Kilcher. The Dimpled Manifold Model of Adversarial Examples in Machine Learning (Research Paper Explained). https://www.youtube.com/watch?v=k_hUdZJNzkU/, 2021. [Online; accessed 17-July-2024].

[9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[10] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016.

[11] Andreas Plesner, Tobias Vontobel, and Roger Wattenhofer. Breaking recaptchav2. In *48th IEEE International Conference on Computers, Software, and Applications (COMPSAC 2024)*. IEEE, 2024.

[12] Adi Shamir, Odelia Melamed, and Oriel BenShmuel. The dimpled manifold model of adversarial examples in machine learning. *arXiv preprint arXiv:2106.10151*, 2021.