ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Seeing is Deceiving: Fortifying reCAPTCHAv2 through Adversarial Machine Learning

The widespread reliance on Google's reCAPTCHAv2 as a primary defense[1] against automated web attacks is facing a critical challenge due to recent advances in computer vision technology. With an estimated market share of 99.93% [1], this Turing test is vulnerable to solvers using pre-trained object detection and segmentation models. These attacks have been shown to achieve a success rate of 92-98% [2, 8, 18, 9] and most recently 100% as demonstrated by Plesner et al. [16]. The low computational cost of these open pre-trained models makes them accessible to a wide range of adversaries and poses a significant threat to the security of online platforms.

We hypothesize that due to the transferability of adversarial perturbations across models [6, 3] and the gap between human and machine perception [5] perturbing CAPTCHA images can effectively mitigate vision-based attacks without compromising the user experience.
This approach offers a practical, cost-effective, and scalable solution for Google and other organizations that rely on reCAPTCHA to secure their online platforms. Our study aims to fill a critical gap in the field by developing and implementing a practical solution to fortify reCAPTCHAv2 against vision-based attacks, an approach that has been hypothesized for a subset of the CAPTCHA tasks [7] but not yet realized.

We expect this counter-offensive strategy to be a generalizable defense against adversarial attacks. We will evaluate the effectiveness of the perturbations by measuring the success rate of vision-based attacks on the perturbed CAPTCHAs and the usability of the perturbed images for human users.
This leaves us with the following research questions:

RQ1 How do the existing vision-based attacks against reCAPTCHAv2 compare, and what are the common patterns among them?

RQ2 How do perturbed CAPTCHAs perform against these attacks and how well do they transfer across models?

We will address these questions empirically by building a reCAPTCHAv2 clone, perturbing its images with adversarial noise such as FGSM [6] and PGD [14] and evaluating the effectiveness of the perturbations against vision-based attacks.
We will also assess the transferability of the perturbations across different models and the robustness of the perturbed CAPTCHAs against adversarial attacks. The results will provide insights into the generalizability of adversarial defenses against vision-based attacks and the potential of adversarial perturbations to fortify reCAPTCHAv2 against vision-based attacks.

[1] Google's reCAPTCHAv3 falls back on reCAPTCHAv2 when it detects suspicious traffic.

By addressing these research questions, our study will not only contribute to the fields of adversarial machine learning and cybersecurity but also provide practical insights for improving the security of widely used CAPTCHA systems. The results could have far-reaching implications for online security practices and the development of more robust human-AI differentiation techniques.

Detailed project outline

The detailed project outline is as follows [2]:

- Literature review (related work, previous approaches) (⋆⋆⋆)
- Building a reCAPTCHAv2 clone (as an open-source attack/defense benchmarking tool for the community) (⋆⋆⋆⋆)
- Generating a robust dataset for the CAPTCHA clone using adversarial examples (⋆⋆⋆)
- Evaluating the effectiveness of perturbations against vision based attacks (⋆⋆⋆)
- Writing the final report/thesis (⋆⋆⋆⋆⋆)
- Midterm and final presentations (⋆⋆⋆⋆)

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

Extension

Optional extensions to the project if time allows include researching the nature of adversarial examples:

- Assessing previous work on the dimpled manifold hypothesis
- Studying the transferability of adversarial perturbations to human vision systems. Thinking of the possibility of emulating human vision systems using deep learning models.
- Formalizing falsifiable hypotheses for the dimples paper and conducting experiments to test them (based on: [13, 12, 11])
- Exploring distillation learning to improve adversarial robustness (based on: [15, 14])
- ... (more ideas are welcome)

For some context: Since the adversarial vulnerability of deep neural networks was discovered in 2013 [6], there have been many attempts to explain why adversarial examples exist and how they work, each with their limitations and assumptions – some complementary, some

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

contradictory [3]. And there are still many open questions. One of the hypotheses is the "dimpled manifold hypothesis" [4] proposed by Shamir et al. [17], 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.

Additionally the paper "Adversarial Examples that Fool both Computer Vision and Time-Limited Humans" [5] showed that adversarial examples can fool time-limited humans, suggesting that there could be a connection between adversarial examples and human perception.

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] Björklund, Arvid and Uogele, Marius. Classifying Google reCAPTCHA v2 - A study using transfer learning models and evaluating their robustness against adversarial perturbations, 2023. Student Paper.

[3] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th USENIX security symposium (USENIX security 19)*, pages 321–338, 2019.

[4] 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.

[5] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, 31, 2018.

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

[7] 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.

[8] I Hossen, Yazhou Tu, F Rabby, Md Nazmul Islam, Hui Cao, and Xiali Hei. Bots work better than human beings: An online system to break google's image-based recaptcha v2. 2019.

[9] Md Imran Hossen, Yazhou Tu, Md Fazle Rabby, Md Nazmul Islam, Hui Cao, and Xiali Hei. An object detection based solver for {Google's} image {reCAPTCHA} v2. In *23rd international symposium on research in attacks, intrusions and defenses (RAID 2020)*, pages 269–284, 2020.

[3] For a comprehensive overview of the hypotheses, see the Addendum of Ilyas et al. [10]

[4] A similar idea was previously proposed by Elliot et al. [4]

[10] 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.

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

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

[13] 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].

[14] 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.

[15] 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.

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

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

[18] Krish Sukhani, Sahil Sawant, Sarthak Maniar, and Renuka Pawar. Automating the bypass of image-based captcha and assessing security. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 01–08. IEEE, 2021.