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

RQ1: What causes adversarial examples?

In 2013 Szegedy et al. [1] discovered that deep neural networks are vulnerable to adversarial examples – inputs with imperceptible perturbations that cause misclassification, revealing blind spots where the models' decision boundaries are brittle, despite appearing to generalize well on normal inputs. Since then we have made a lot of progress in both finding attack vectors [4] [11] [12] and adversarially training models to be more robust [15] [11] [13] to these attacks.
However, despite the progress, we still don't fully understand:

- What they are, why they exist and how they work.
- How to defend against them without sacrificing accuracy (robustness vs accuracy tradeoff).
- Why they transfer between different models, datasets, architectures, training procedures and even to the human visual system [3].

There have been many attempts at explaining adversarial examples, each with limitations and assumptions – some complementary, some contradictory.
For example, the *dimpled manifold hypothesis* [16] [1] suggests that the decision boundary of deep neural networks is close to the data manifold, making it easy to find adversarial examples, while the *non-robust features hypothesis* [6] suggests that models exploit non-robust features that are imperceptible to humans, leading to a vulnerability against small perturbations [2].

The *dimpled manifold hypothesis* is a particularly controversial one, as it was criticized by Yannik Kilcher [9] in 2021, who also provided a counterexample in less than 100 lines of code [8]. Despite the lack of generalizability, a master's student from the University of Vienna, Lukas Karner, successfully verified and replicated all experiments detailed in the dimples paper [7] in 2023. Lukas allowed me to use his results in my work. He mentioned that there is currently no paper or thesis based on his work and that he would be happy to see his results being used in a meaningful way.
This leaves us with the possibility that the experiments carried out are correct in themselves, but that the chain of reasoning is inconclusive and therefore doesn't generalize. Investigating this further would require more rigor by formalizing falsifiable hypotheses based on the paper and conducting experiments to test them.

[1] A similar idea was previously proposed by Elliot et al. [2]
[2] For a comprehensive overview of the hypotheses, see the Addendum of Ilyas et al. [6]

Another aspect of the *dimpled manifold hypothesis* worth exploring is the idea of projecting the perturbations on the data manifold before applying them to the input space. Reducing the dimensionality of the perturbations or compressing them makes them visible to the human eye and interpretable as demonstrated by Karner [7].

The *non-robust features hypothesis* on the other hand also has ideas worth exploring such as using distillation to improve adversarial robustness as also demonstrated by Papernot et al. [13].

RQ2: How can adversarial examples be used to design CAPTCHAs?

Google's reCAPTCHAv3 falls back to version 2 when it suspects a bot, which can be broken with a 98% success rate using object detection and segmentation models as demonstrated by Plesner et al. [14] – effectively rendering the entire system useless. This removes another layer of security from the internet, making it easier for bots to scrape data, launch DDoS attacks and perform other malicious activities.
One way to mitigate this vulnerability is to apply perturbations to the CAPTCHA images that are imperceptible to humans but cause misclassification in the object detection and segmentation models by leveraging adversarial examples. Although this has been previously hypothesized for 1/3 of the reCAPTCHAv2 tests [5], no practical implementation has been developed yet, based on our literature review.

Objectives

This leaves us with 2 possible directions for the project:

a) Understanding the causes of adversarial examples by:

- Formalizing falsifiable hypotheses for the dimples paper and conducting experiments to test them.
- Exploring dimensionality reduction and compression of adversarial perturbations to make them interpretable.
- Exploring distillation learning to improve adversarial robustness.

b) Improving CAPTCHAs by:

- Adding perturbations to CAPTCHA images to make them harder to solve for object detection (YOLOv5) and segmentation (SAM) models [10].

Both directions are promising and can be pursued in parallel by reducing the scope of the project. The directions can also complement each other, as the insights gained from the first direction can be used to design better adversarial examples for the second direction.

Project Plan

... (to be discussed)

References

[1] Ekin D Cubuk, Barret Zoph, Samuel S Schoenholz, and Quoc V Le. Intriguing properties of adversarial examples. *arXiv preprint arXiv:1711.02846*, 2017.

[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] 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.

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

[5] 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.

[6] 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.

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

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

[9] 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].

[10] Jiahao Lu, Xingyi Yang, and Xinchao Wang. Unsegment anything by simulating deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24294–24304, 2024.

[11] 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.

[12] 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.

[13] 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.

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

[15] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in neural information processing systems*, 32, 2019.

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