1 Related Work

CAPTCHA, short for "Completely Automated Public Turing Test to Tell Computers and Humans Apart", is used by lots of websites to defense against attacks like malicious programs, automated registrations, email spam, dictionary attacks, and search engine bots [6] [15] . During the evolution of CAPTCHA, four different types of CAPTCHA have come into sight, which are text-based, image-based, video-based [7] and game-based. While video-based and game-based are rarely used, text-based and image-based CAPTCHAs are still widely used even today. However, with the rapid advancement of machine learning (ML) techniques, most of these CAPTCHAs have been broken [16].

Text-based CAPTCHA is the first type of CAPTCHA being used to tell that the client is a real human being instead of an automated machine. Typical text-based CAPTCHA requires you to identify characters on an image and submit your answer [1]. Attacks target at text-based CAPTCHA can be divided into segmentation-based and recognition-based [2]. To defense the attacks, text-based CAPTCHA has evolved into multiple variants, including overlapping characters, using a two-layer structure, rotating and waving characters or using a large character set [17]. However, even though some of the CAPTCHAs are difficult for humans to recognize after mutation, machines equipped with ML models like SVM, KNN [2] and back-propagation neural network [8] can easily recognize the characters after training. The vulnerability of text-based CAPTCHA leads to the development of image-based CAPTCHA.

Comparing with text-based CAPTCHA which only requires character recognition and segmentation, image-based CAPTCHA seems to be a better choice since it is built on semantics information of images before the thriving of ML based image classification techniques [4] [3]. The most common type of image-based CAPTCHA is the selection-based CAPTCHA, where a user is required to select all images that match a given image described by a context. To attack an image-based CAPTCHA, various methodologies have been proposed. Golle [5] builds a SVM classifier that attacks Asirra [4], which requires the user to find cats among 12 images of cats and dogs, with the highest accuracy of 82.74%. More recently, Sivakorn et al [13] propose a four-step model: get hint from the given image by GRIS, tag all images with labels with online services and libraries, build Word2Vec word vectors between tags and hint, select the images with highest cosine similarities. The highest accuracy of applying this model on the "no captcha reCaptcha" deployed by Google is about 70.78%, and is 83.5% on Facebook image CAPTCHAs. Current image-based CAPTCHA can not satisfy the security requirement anymore.

While ML attacks seem to be too strong for CAPTCHA to defend against, recently, researchers are inspired by adversarial machine learning in computer vision domain, which uses adversarial examples to misguide models [14]. Adversarial examples are created via adding some perturbations to the original legitimate image, such that these perturbations make this image cross the decision boundary and be recognized as another class [9]. Apart from this misguidance ability, adversarial attack also shows its transfer-ability, meaning that an attack that affect one model will also affect other models [11]. Osadchy et al [10] introduce DeepCAPTCHA with immutable adversarial noise (IAN), which is not only resistant to removal attempts like image pre-processing, but also can lead to the misclassification of deep learning networks. This IAN is achieved by iteratively calling the Fast Gradient Sign Method (FGSM). Shi et al [12] propose aCAPTCHA, an adversarial CAPTCHA generation and evaluation system. To decrease the recognition ability, aCAPTCHA adds as many as perturbations to images as long as it is still user tolerable. ACAPTCHA basically use Carlini-Wagner Attacks(CW) and Jacobian-based Saliency Map Attack (JSMA) with a few minor changes. A main problem of these models is that they only apply one adversarial attack each time. CAPTCHA attackers can use modified image processing methods to mitigate perturbations.

Considering that adversarial examples do improve the robustness of image CAPTCHA, we could consider going further in this direction. What if these attacks are integrated in one model? Will it improve the robustness of adversarial image-based CAPTCHAs?

References

[1] Fatmah H Alqahtani and Fawaz A Alsulaiman. "Is image-based CAPTCHA secure against attacks based on machine learning? An experimental study". In: *Computers & Security* 88 (2020), p. 101635.

[2] Elie Bursztein, Matthieu Martin, and John Mitchell. "Text-based CAPTCHA strengths and weaknesses". In: *Proceedings of the 18th ACM conference on Computer and communications security*. 2011, pp. 125–138.

[3] Monica Chew and J Doug Tygar. "Image recognition captchas". In: *International Conference on Information Security*. Springer. 2004, pp. 268–279.

[4] Jeremy Elson et al. "Asirra: a CAPTCHA that exploits interest-aligned manual image categorization." In: *CCS* 7 (2007), pp. 366–374.

[5] Philippe Golle. "Machine learning attacks against the Asirra CAPTCHA". In: *Proceedings of the 15th ACM conference on Computer and communications security*. 2008, pp. 535–542.

[6] Walid Khalifa Abdullah Hasan. "A survey of current research on captcha". In: *Int. J. Comput. Sci. Eng. Surv.(IJCSES)* 7.3 (2016), pp. 141–157.

[7] Kurt Alfred Kluever and Richard Zanibbi. "Balancing usability and security in a video CAPTCHA". In: *Proceedings of the 5th Symposium on Usable Privacy and Security*. 2009, pp. 1–11.

[8] Xiao Ling-Zi and Zhang Yi-Chun. "A case study of text-based CAPTCHA attacks". In: *2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*. IEEE. 2012, pp. 121–124.

[9] Gabriel Resende Machado, Eugênio Silva, and Ronaldo Ribeiro Goldschmidt. "Adversarial machine learning in image classification: A survey toward the defender's perspective". In: *ACM Computing Surveys (CSUR)* 55.1 (2021), pp. 1–38.

[10] Margarita Osadchy et al. "No bot expects the DeepCAPTCHA! Introducing immutable adversarial examples, with applications to CAPTCHA generation". In: *IEEE Transactions on Information Forensics and Security* 12.11 (2017), pp. 2640–2653.

[11] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples". In: *arXiv preprint arXiv:1605.07277* (2016).

[12] Chenghui Shi et al. "Adversarial captchas". In: *IEEE transactions on cybernetics* (2021).

[13] Suphannee Sivakorn, Iasonas Polakis, and Angelos D Keromytis. "I am robot:(deep) learning to break semantic image captchas". In: *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE. 2016, pp. 388–403.

[14] Christian Szegedy et al. "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199* (2013).

[15] Luis Von Ahn, Manuel Blum, and John Langford. "Telling humans and computers apart automatically". In: *Communications of the ACM* 47.2 (2004), pp. 56–60.

[16] Xin Xu, Lei Liu, and Bo Li. "A survey of CAPTCHA technologies to distinguish between human and computer". In: *Neurocomputing* 408 (2020), pp. 292–307.

[17] Yang Zhang et al. "A survey of research on captcha designing and breaking techniques". In: *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE. 2019, pp. 75–84.