INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 12, Issue 4, April-2021
ISSN 2229-5518

Review of various CAPTCHA generating systems and vulnerabilities

Pratik Kahar

M.E. Cyber Security, Graduate School of Engineering and Technology,
Gujarat Technological University, Ahmedabad, Gujarat, India.

Abstract— CAPTCHA, a mechanism to check human as a user over internet. Currently available in many forms, text, image, audio etc. But attackers are able to break this mechanism using ML methods. ML instead can be use for protectection. Vulnerabilties assement and model development can be done using ML.

Keywords- Machine Learning, CAPTCHA, reCAPTCHA,Attacks on CAPTCHA, CNN

1 INTRODUCTION

Completely Automated Public Turing Tests to tell Computers and Humans Apart (CAPTCHA) are one of the common things when browsing the internet today, and these techniques is frequently used as a security measure by the websites to defend against the robots and automated bots to reduce unwanted traffic and unauthorised access. CAPTCHAs are tests to identify potential autonomous bots/robots from humans. The general usage of a CAPTCHA is that it should be hard for computers programs or bot and easy for humans to solve. The most popular CAPTCHA in use today is Google's reCAPTCHA v2 and v3 which are mainly using image-based CAPTCHA challenges [2][4].

We're regularly do login on internet over various web pages for signing in process. We're using applications like g-mail, yahoo mail, and many web sites which ask for registration, and during the registration process we come across CAPTCHA for security reason. The CAPTCHA is a test we need to clear this test to move further. Thetest is not difficult for us but it is difficult for bots. The CAPTCHA must be simple not confusing for us to use. On the other side it should be almost impossible for machineto break. The CAPTCHA which is mostly in use is an image. To solve CAPTCHA is ours job. To passed the test we have to type the text shown in image or to selectthe item ask in question give above in image. Why to use CPATCHA? Because of current risk of hacking, the hackers are continuously trying to get unauthorised access into asystem to gain the data inside it. Because of the hackers who are doing illegal activity on Internet causing damage to other user over internet by stealing their data, identity, and also damaging their computer software's by mean of malware. For example, trojan horse virus installation in victim computer. This is used to steal password credential etc from victim's computer. So, the identification is necessary fordistinguishing a bot and a human. Hackers are continuously building malicious scripts/code/programs for breaking into a system. So robust CAPTCHAs and system have to bedeveloped so that the efforts of the hackers can be made useless.

1.1 WHY USE CAPTCHAs

As discussed above the risk of hacking is getting on its peek and the security threats are increasing at a high pace, so the problem is getting big. What to do in this state, we have to create a secure and safe environment which should be free from mainly vulnerabilities in the applications because the attackers are always in the search of vulnerabilities. Why we should focus more on vulnerabilities? Because the vulnerability is the week point from where an attacker can get through in, and breaks into the system, and then he/she can damage the system. In our case of web security, we are making the login system the authentication process secure from automated machine attacks and bot attacks who are trying to get unauthorised access. For this one very well-known method is CAPTCHA test which make computer and human apart.

Some examples:

1. Accounts which are being made free over free online sites are storing the credentials and these are use to create fake accounts for forgery and to stole copyrighted material without being identified by security.
2. In rating systems, the fake accounts with real identity can be use to make the high rating of product, and show the fake product as original and good.
3. Attackers are creating programmes for changing the product rate up and down, also changing the user's ideas about the product.
4. Spammers uses free services for register themselves on mail service and use their automated programmes/ scripts to send fake e-mails to other users, generally spam e-mails.
5. With consideration of above points there occurs a need for valid user verification for grant access of product service. In our case only human can access the service. The CAPTCHA is the way for this.

From when the CAPTCHAs are created, various types are also derived/ generated. Today we classify CAPTCHAs in three main categories, (1) text-based, (2) image and (3) audio/video. Developers are finding CAPTCHA weaknesses to identify its robustness and then check that which propertiesare good for its

INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 12, Issue 4, April-2021
ISSN 2229-5518

security. Since the date many versions have been created and also, they were attacked and test for their weaknesses and strength. The CAPTCHA methods of generation and breaking are developed continuously, and the literature related to this is now is in abundant form [1][3].

1.2 CAPTCHAs: Types and Attacks

Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA). To prevent automated bot and robots to get registered themselves on internet, for the sake of security mechanism it's been used [1].

If a human is solving a CPATCHA with a great accuracy of more than 90 percent and the same CAPTCHA solving rate of a bot is less than a 1 percent, than the CAPTCHA is considered as a good one [1]. So many types of CAPTCHA's are created, Mainly classified in three categories (1) text,
(2) image and (3) audio/video-based. Among various types the successful one is alphanumerical CAPTCHAs.Thus, the attacking machines can be distinguished.

CAPTCHA technology is constantly developing, and their vulnerabilities too. Developers are finding CAPTCHA weaknesses to identify its robustness and then check that which properties are good for its security. Today, Machine learning and deep learning are taking charge in modern technologies almost all the application are being developed using this technology, as these technologies are self-driven which have the ability to learn without being explicitly programmed.

Machine Learning methods are well known for attacks. Various Machine learning algorithms are used for both generation and breaking of CAPTCHA systems.

1.3 Various types of CAPTCHAs

Text, Image, Maths equation, reCAPTCHA, SEMAGE-Semantically Matching image, Audio, Video, Word problem, Honeypot, Media sign, Multi-lingual

Attacks methods used on CAPTCHAs

1. CNN - Convolutional Neural Network
2. GAN - Generative Adversarial Network
3. SVM – Support Vector Machine
4. KNN - k-nearest neighbours
5. RCN - Recursive Cortical Network
6. SIMGAN - semantic image manipulation using gener-ative adversarial networks
7. OCR – Optical Character Recognition
8. Segmentation, Recognition
9. Transfer Learning,Computer Vision.

And observing these papers I found that the mostly used and with high success rate method is the CNN

Fig. 1. Types of CAPTCHAs

(Convolutional Neural Network). And in comparison, of CNN the GAN (Generative Adversarial Network) is alsodoing good [1][3].

1.4 ML Method Comparative Analysis

Here are three most used ML methods for CAPTCHA recognition. This method have been prove very effective for text and image CAPTCHAs and the attackers are intendedto use these methods for automatic CAPTCHA text, pattern, and design recognition. As ML allows no human interaction, the task become free from human interaction their will be no need of further human contact, the ML algorithm perform the task alon.

1.4.1 Generative Adversarial Network [1][2][4][8][10] GANs are uncontrolled learning methods GANs do not require labeled data; can be trained using unlabeled data as they read internal data presentations.GANS train model by an adversarial process the two models are trained simultaneously the generator creates image which looks real, and a discriminator learns to differentiate between real and fake images. This process ends when the discriminator no longer differentiates the real and fake images.
Key Characteristics

1. GANs that Produce photorealistic images can be used to visualize interior design, small details from image.
2. Great for OCR
3. Good for image and video CAPTCHAs.

Scope of Improvement

- 1.constructing a large dataset need to reduce overconfidence

INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 12, Issue 4, April-2021
ISSN 2229-5518

and overfitting

1.4.2 Convolutional Neural Network [1][2][3][4][8][10].

CNNis very powerfull machine learning algorithm which works onfeed forward neural network and generally used for analyze image processing, detect and classify the objects in it. CNN works in flow of a) Convolution layer, b) Max pool,c) Relu activation function, and d) Fully connected layer. everyimage is an array of pixels.

Here's how CNN processes :

1. The pixels from the image are given a convolutional layer that performs the function of convolution
2. It results in a merged map The map shown is used in the ReLU function to generate a modified feature map
3. The image is processed through multiple convolutions and ReLU layers for features
4. Different layers of mixing with different filters are used to identify specific parts of the image
5. The integrated feature map is flat and is inserted into a fully integrated layer to get the final result

Key Characteristics

1. CNN automatically detects the important features without any human supervision
2. CNN reduces the post processing of input.
3. Highly used in OCR method
4. Best for Image and Average for video CAPTCHAs.

Scope of Improvement

1. Need to increase levels of structure: Neurons, and Layers to generate less matrix for image pooling.

1.4.3 RNN - Recurrent Neural Network [1][2][3][4][9].

The Recurrent Neural Network works on the goal of saving the output of a particular layer and feeds this after the input to predict the release of the layer. Recurrent neural networks created because of a few problems with the feed-forward neural network:

1 Unable to manage consecutive data
2 Views only current input
3 We cannot memorize previous entries
4 The solution to these problems is the Recurrent Neural Network (RNN). RNN can manage consecutive data, accept current inclusion data, and previous acquiredentries. RNNs can memorize previous inputs because of their internal memory.

Key Characteristics

1 RNN is suitable for temporal data, also called sequential data.
2 Plain text images recognition is easy
3 RNNs reuse activation functions from other data points in the sequence to generate the next output in a series.
4 Best for text image CAPTCHAs.

Scope of Improvement

1. Training of RNN models can be difficult.
2. Training time for image data can be improve.

2 Identified Issues

Almost all types of existing CAPTCHAs are vulnerable [2]. Whether we use text based or image based CAPTCHA both can be broken some take less time and some take more time to break.

Bypassing of CAPTCHAs for text, image, and maths equations are easy [1][2][11].

As text CAPTCHAs are alphabet and numbers they are very easy to recognise by the use of some algorithms. and the machine learning methods for image processingare the way to break image.

Not all type CAPTCHA generators are efficient in terms of time complexity, space complexity andvisualization [1][2][10].

The CAPTCHAs generated by some generators are not good in visualization many times it is difficult to recognise the text and image in it.

CAPTCHAs vulnerable to Methods

Abbreviation use are G - GAN, C - CNN, R - RNN, O - OCR[11], CV - Cloud Vision[5], YOLO - (You-only-look-once)[5].

Variants	Type	Methods	Rate
Simple text	Text	G, C, R, O	99.8
Cursive text	Text	G, C, R, O	98.1
Cut and Split	Text	G, C, R, O	98.04
Overlap	Text	G, C	95.4
Segmented	Text	G, C	97
Python	Text	G, C	94
reCAPTCHA	Image	YOLO, CV	68.3
Object	Image	G, C, CV	91
CaptchaStar	Image	G ,C ,CV	79
Audio	Audio	Voice Recognition	98
Video	Video	Image Processing	90
Animated	Animated	Image Processing	93.5

TABLE I
RESULT OF CNN

From the above table it is clear that most of the CAPTCHAs are vulnerable towards ML methods even googles re-CAPTCHA is broken. Though many simple CAPTCHAs are not that much vulnerable but because of complex structure which are difficult for humans to recognise are not applicable, and are substance of wast product. The CAPTCHA should be more and more easy for human to recognise and use. It is my own experience that many text CAPTCHAs are tedious to solve. Which take more time at the time of verification. One CAPTCHA is based on pediatric human motion [6]. It is only applicable in mobile device and are not applicable in PC, laptop and Desktop computers. Today we

INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 12, Issue 4, April-2021
ISSN 2229-5518

are using image CAPTCHAs but ML and DL methodsof image processing and recognition making it vulnerableas they work on data sets and self evolving algorithms, there's the threat of simple bypass of CAPTCHA. Audio

CAPTCHAs are good development in this technology but voice recognition is done for cracking it. One novel approach CaptchaStar[13] is made to make this technology vulnerabil- ity free but the developer had broken it by himself with 70 percent accuracy. and their is the issue of time occurs for human, on an average more than 30 seconds for recognising the CAPTCHA. So how to make a CAPTCHA that will be robust and free from vulnerabilities and also very easy for human to recognise and solve in minimun time.

2 PROPOSED WORK

4.1 Data Description

Our work is to generate a robust CAPTCHA system. So, the first step is to know about the existing CAPTCHA systems.

Points to Follow:

1. CAPTCHA types
2. Which of these are mostly used
3. Are they vulnerable
4. Attacks on CAPTCHA system
5. Success rate of attack
6. Prevention from attack
7. Measures for prevention

If the prevention measures available, check they're for its success.

Second step is to make our own Design of prevention.

Points to Follow:

8. Performing practical on existing system
9. tries for any vulnerability.
10. try to remove that vulnerability
11. checking it usability
12. checking again its strength

Third is to propose the new system

And last assessment of proposed system

Result of CNN Algorithm

B. TABLE II

C. RESULT OF CNN

D.

Parameter	Performance
No. of Epoch	100
Accuracy obtain	0.9895
Loss during processing	0.007 percent
Speed of training	263 ms/step

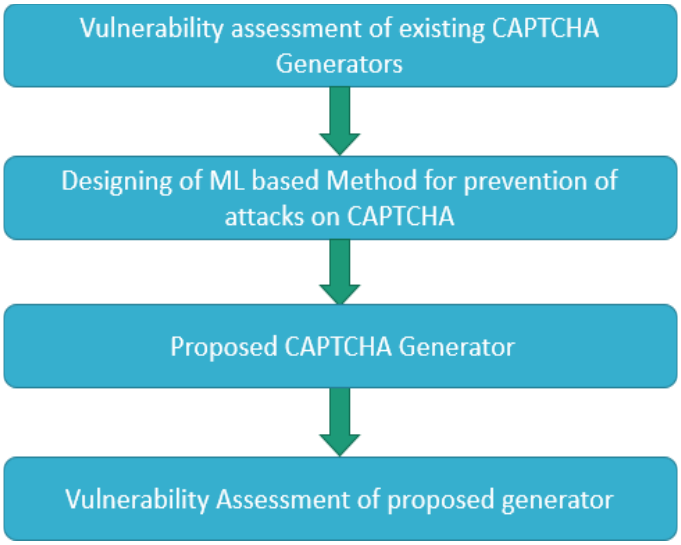

Fig. 2. Proposed flow of work

Figure shows the Steps of proposed flow of work.

Our design will be simple for humans to use . For that the design will not be different from text embedded in image , just the difference will be a Strong random combination of alphabets both uppers and lower case plus numbers , this is what we humans are familiar . For strengthening and makingit robust the image will be having several noise factors that make the text appear like untidy handwriting of humans . Now to prevent from robots and bot programs we will use a virtual keyboard for entering text . This Will also prevent key logging activity and robots will not be able to use this system .

SCOPE OF PROPOSED WORK

The proposed system will be very difficult (nearly impossible) to break for automated systems or bots.

Now a days their is big problem arising of automated systems. Due to increase in IOT, AI, and ML technologies for the ease of application usage and time reduction, the attackers are tacking advantage of these technologies too for their greed. They are now using automated programs for performing attacks on machine and bypassing the security systems. In termsof CAPTCHA systems the use of ML methods are used for bypassing the system.

Using ML technology attackers are Generating automated bots for bypassing of system. So the system should be so difficult so that not any bot can bypass it.

Making a robust and random CAPTCHA Generator.

FUTURE SCOPE

The Google's reCAPTCHA can be used restrictedly or it can be made more robust and more tedious so the time for

machine to crack it will be high, so that with time restriction the machine will fail [2][4]. Future of CAPTCHAs may be in the direction of human motion or may can be virtual CAPTCHAs.

CONCLUTION

By analysing in deep, this work reviews the security mechanisms of three kinds of CAPTCHA, and the attack methods. Whether is a text or image CAPTCHA the best attack method used so far is CNN and the most widely used CAPTCHA system today that is Google reCAPTCHA v2,v3 are good but yes, they can also can be broken.

ACKNOWLEDGMENT

I would ld like to express a deep sense of gratitude and whole hearted thank to my parents and research guide prof. Mahesh Panchal in order to guide me throughout the way. it was his keen interest, encouragement and full cooperation that have made it possible for me to move ahead with the work. i would also like to thank my freinds whose guidance had inspired me for this work.

REFERENCES

[1] A. Dionysiou and E. Athanasopoulos, "SoK: Machine vs. machine – A systematic classification of automated machine learning-based CAPTCHA solvers," *Comput. Secur.*, vol. 97, 2020, doi: 10.1016/j.cose.2020.101947.

[2] Y. Zhang, H. Gao, G. Pei, S. Luo, G. Chang, and N. Cheng, "A survey of research on CAPTCHA designing and breaking techniques," Proc. - 2019 *18th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. IEEE Int. Conf. Big Data Sci. Eng. Trust. 2019*, pp. 75–84, 2019, doi: 10.1109/TrustCom/BigDataSE.2019.00020.

[3] Z. Nouri and M. Rezaei, "Deep-CAPTCHA: A Deep Learning Based CAPTCHA Solver for Vulnerability Assessment," SSRN Electron. J., 2020, doi: 10.2139/ssrn.3633354.

[4] P. Wang, H. Gao, Z. Shi, Z. Yuan, and J. Hu, "Simple and Easy: Transfer Learning-Based Attacks to Text CAPTCHA," IEEE Access, vol. 8, pp. 59044–59058, 2020, doi: 10.1109/ACCESS.2020.2982945.

[5] D. Wang and M. Moh, "Using Deep Learning to Solve Google reCAPTCHA v2 ' s Image Challenges," pp. 4–8, 2020.

[6] S. Kulkarni and H. S. Fadewar, "Pedometric CAPTCHA for mobile Internet users," RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018-Janua, pp. 600–604, 2017, doi: 10.1109/RTEICT.2017.8256667.

[7] R. Omid, S. Gary, J. Large, and B. David, "In-Depth Study of CAPTCHA," no. April, pp. 0–21, 2017, doi: 10.13140/RG.2.2.17533.97768.

[8] N. Yu and K. Darling, "A low-cost approach to crack python CAPTCHAs using AI-based chosen-plaintext attack," Appl. Sci., vol. 9, no. 10, 2019, doi: 10.3390/app9102010.

[9] T. Zhang, H. Zheng, and L. Zhang, "Verification CAPTCHA based on deep learning," Chinese Control Conf. CCC, vol. 2018-July, pp. 9056–9060, 2018, doi: 10.23919/ChiCC.2018.8482847.

[10] J. Zhang et al., "Robust CAPTCHAs towards Malicious OCR," IEEE Trans. Multimed., vol. 9210, no. c, pp. 1–1, 2020, doi: 10.1109/tmm.2020.3013376.

[11] M. Conti, C. Guarisco, and R. Spolaor, "CAPTCHaStar ! A Novel CAPTCHA Based on Interactive Shape Discovery," pp. 1–15.

INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 12, Issue 4, April-2021
ISSN 2229-5518

INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 12, Issue 4, April-2021
ISSN 2229-5518

INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 12, Issue 4, April-2021
ISSN 2229-5518

INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 12, Issue 4, April-2021
ISSN 2229-5518

5