Security and Privacy of Captcha using Deep Learning- Black box Testing

-Pranay Bachu
12/7/2022

Advisor- Dr. Yingshu Li

Georgia State University

Goals

To enable enhanced security in Captcha with deep learning models and prevent attacks and issues related to security.

What is CAPTCHA?

A program or system intended to distinguish human from machine input, typically as a way of thwarting spam and automated extraction of data from websites – **WIKI**

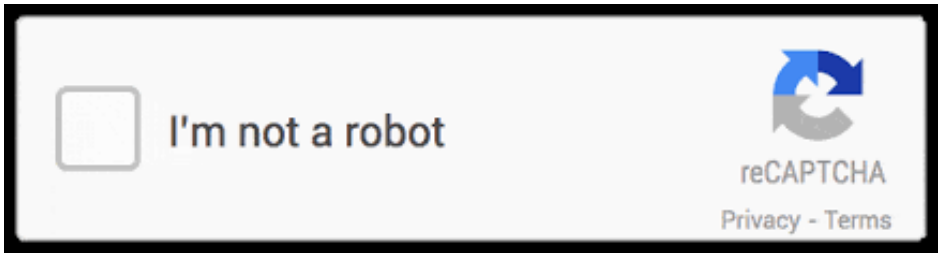
Completely Automated Public Turing test to tell Computers and Humans Apart(CAPTCHA)

Importance of Captcha

The moment a Captcha is seen people do not realize the importance of it and feel impatient to complete the captcha step.

But Captcha is very important for both the company and the users to have a seamless experience below are the reasons:

- Protect the integrity of online polls by stopping hackers using robots to send in repeated false responses.
- Stop brute force attacks on online accounts in which hackers repeatedly try to log-in using hundreds of different passwords.



More Applications of Captcha

- Prevent hackers from signing up for multiple email accounts that they'll then go on to use for nefarious purposes.
- Stop cyber criminals spamming blogs or news content pages with dodgy comments and links to other websites.
- Prevent ticket touts from using robots to bulk buy tickets for shows and gigs.
- To make online shopping more secure.
- To prevent spams and worms.
- Protecting websites registrations.

Different Types of Captchas

- Text Based Captcha
- Math Problem
- Word Problems
- Captcha Images
- Audio for Captchas
- Social Media Sign in
- No Capthca or ReCaptcha

Text Based Captcha

Text-based CAPTCHAs are the original way in which humans were verified. These CAPTCHAs can use known words or phrases, or random combinations of digits and letters. Some text-based CAPTCHAs also include variations in capitalization.

The CAPTCHA presents these characters in a way that is alienated and requires interpretation. Alienation can involve scaling, rotation, distorting characters. It can also involve overlapping characters with graphic elements such as color, background noise, lines, arcs, or dots.

Word Problem Captcha

- This popular type of captcha varies in different forms but they all come with two simple parts: a text box and a sequence of letters or numbers. To prove your human identity, it is important to follow the test's directions carefully.
- The test might ask you to retype the disordered sequence of letters, enter the last word among multiple ones or answer the color that words displayed in.
- This type of test gives a great option for users who have a visual impairment and have trouble with other types of captcha.

Math Problem Captcha

You may have seen this type of captcha regularly. A captcha form appears with a math problem, requiring you to solve and enter the answer. The questions which are quite simple, such as “1+2”, “8-3” can be difficult for a robot to solve.

The math problem is just a piece of cake for users so they can complete it quickly and move their tasks without much annoyance. However, this easy test can’t guarantee the security of a website because it is not as complicated as some of the other types of captchas

Audio Based

Audio CAPTCHAs were developed as an alternative that grants accessibility to visually impaired users. These CAPTCHAs are often used in combination with text or image-based CAPTCHAs. Audio CAPTCHAs present an audio recording of a series of letters or numbers which a user then enters.

These CAPTCHAs rely on bots not being able to distinguish relevant characters from background noise. Like text-based CAPTCHAs, these tools can be difficult for humans to interpret as well as for bots.

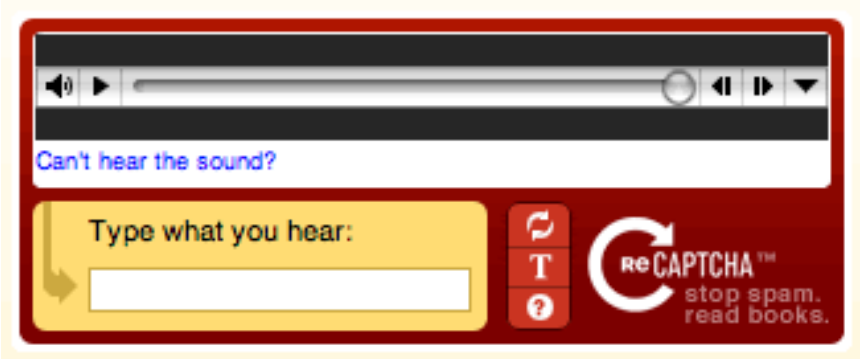
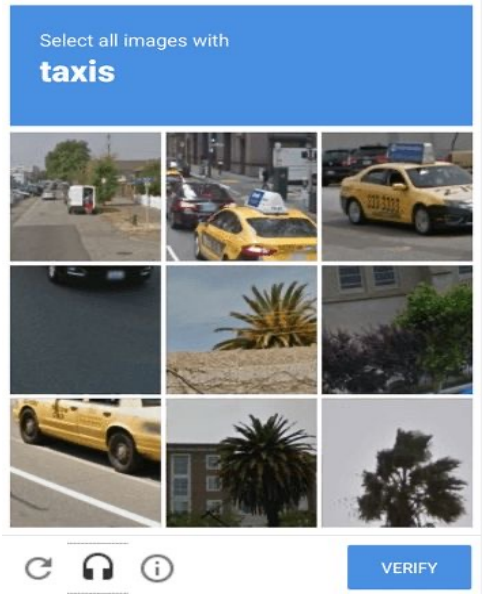
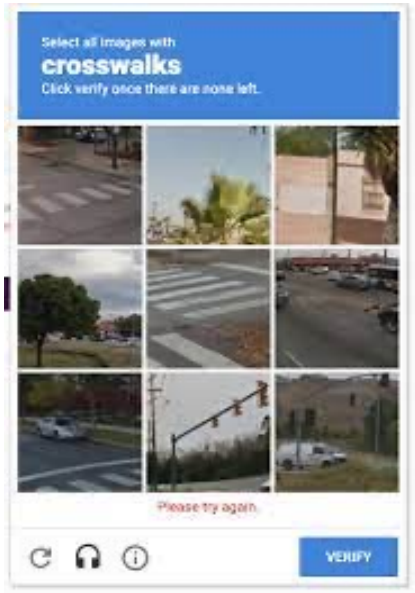


Image Based

Image-based CAPTCHAs were developed to replace text-based ones. These CAPTCHAs use recognizable graphical elements, such as photos of animals, shapes, or scenes. Typically, image-based CAPTCHAs require users to select images matching a theme or to identify images that don't fit.

Image-based CAPTCHAs are typically easier for humans to interpret than text-based. However, these tools present distinct accessibility issues for visually impaired users. For bots, image-based CAPTCHAs are more difficult than text to interpret because these tools require both image recognition and semantic classification.



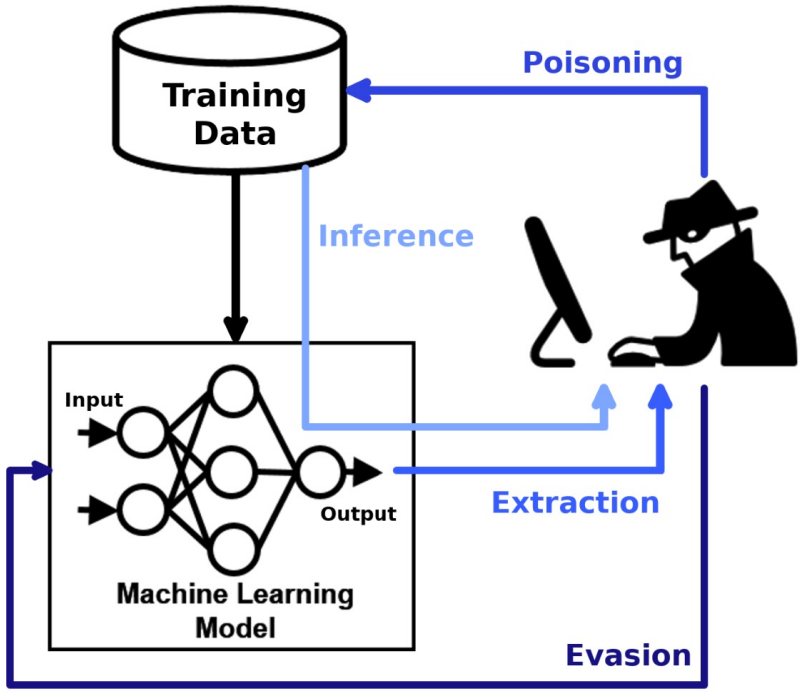


Data Set Generation

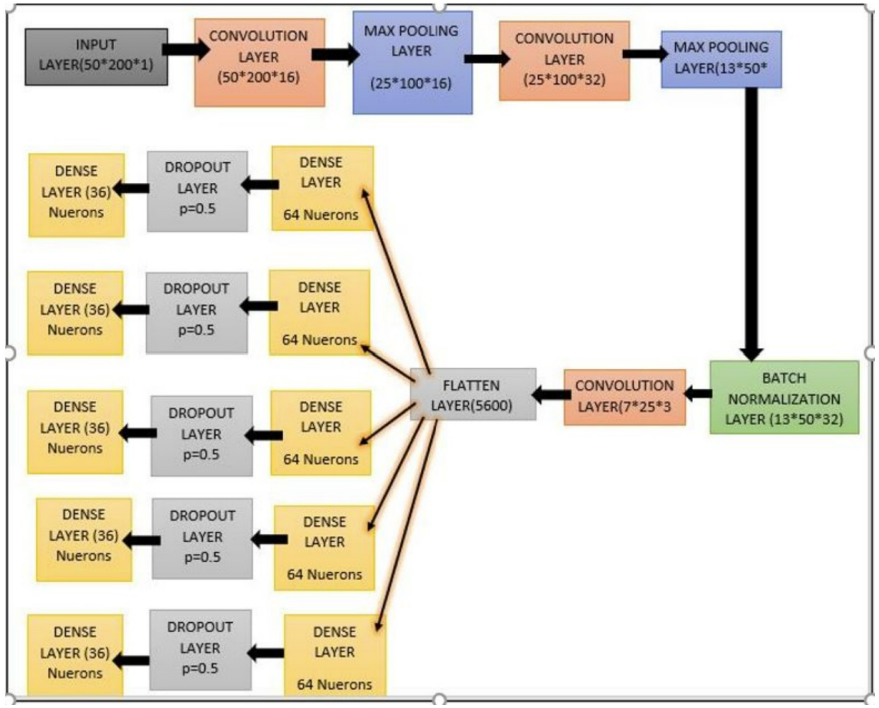
- Our first dataset, Dataset 1, is composed of all 26 English lower-case letters and 10 digits randomly without slant. However, the first dataset does contain character adhesions, blurry parts, distraction lines, and differently shaded parts in the background. The first dataset has an input size of 50*200. It was downloaded from Kaggle.

- The second data set is created in-house so that the model can be trained self sufficiently according to our Blackbox testing, we had a team of 4 people who were constantly creating new data sets of both Image and Audio. These data sets were decided to not be shared as an open source so that the attackers could not use these datasets to train their models

Data Poisoning

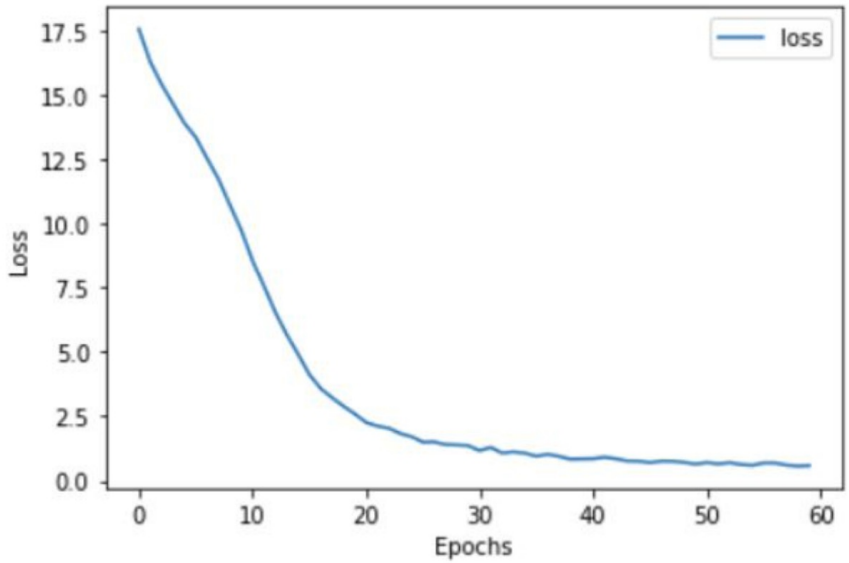


Architecture



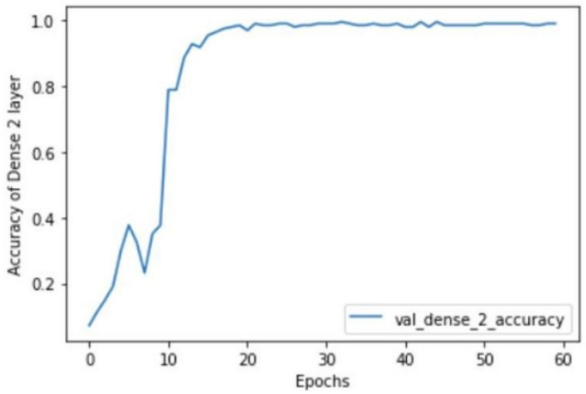
Experimental results

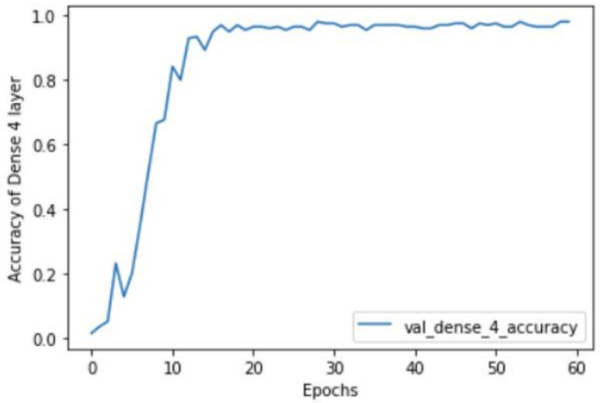
- After training the model for 60 epochs, the following graph was obtained for loss with respect to the number of epochs. We see that as the number of epoch's increases, the loss decreases exponentially. The loss at the end of 60 epochs is 0.5932. The loss obtained on the training set is 0.2391 while the loss on test set is 2.123.

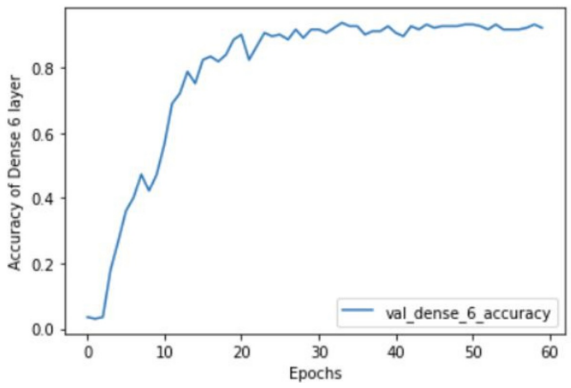


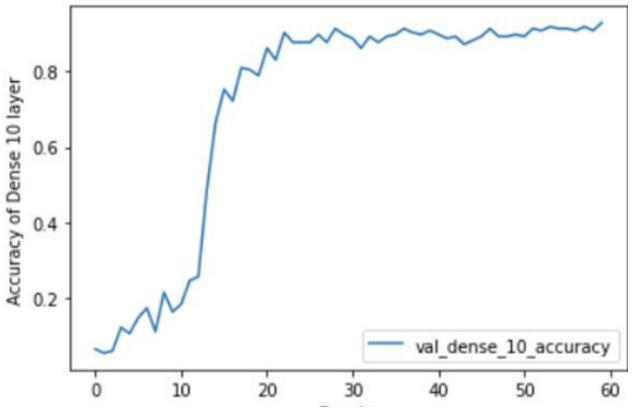
Experimental results

- We see that as the number of epochs increases, the accuracy of the layers improves and hence the system can predict the CAPTCHA more efficiently. The accuracy obtained after 60 epochs for dense layer 2 is 0.9897, dense layer 4 is 0.9794, dense layer 6 is 0.9227, dense layer 8 is 0.8969 and for dense layer 10 is 0.9278.



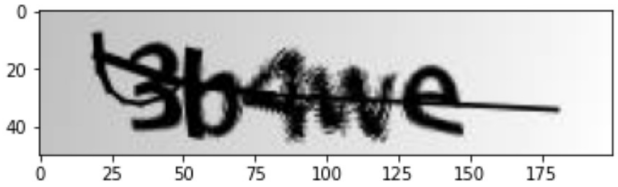






Captcha Prediction

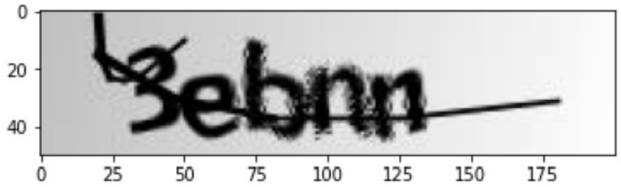
```
<matplotlib.image.AxesImage at 0x7f08a0788cf8>
```




```
print("Predicted Captcha =",predict('/content/drive/My Drive/Pranay/mnop2.png'))
```

```
Predicted Captcha = 3b4we
```

```
<matplotlib.image.AxesImage at 0x7f08a0733860>
```




```
print("Predicted Captcha =",predict('/content/drive/My Drive/Pranay/34d.png'))
```

```
Predicted Captcha = 3ebnn
```

Drawbacks of Using Captcha

- Disruptive and frustrating for users.
- May be difficult to understand or use for some audiences
- Some CAPTCHA types do not support all browsers
- Some CAPTCHA types are not accessible to users who view a website using screen readers or assistive devices

Advantages of Captcha

- Protect the integrity of online polls by stopping hackers using robots to send in repeated false responses.
- Stop brute force attacks on online accounts in which hackers repeatedly try to log-in using hundreds of different passwords.
- Prevent hackers from signing up for multiple email accounts that they'll then go on to use for nefarious purposes.
- Stop cyber criminals spamming blogs or news content pages with dodgy comments and links to other websites.
- Prevent ticket touts from using robots to bulk buy tickets for shows and gigs.
- To make online shopping more secure.

Conclusion

CAPTCHA was designed to improve the security of the systems but deep learning algorithms defeated its very purpose. Here, we used Convolutional Neural networks for CAPTCHA recognition. The model has been trained for CAPTCHA containing small English alphabets and digits, thus for a total of 36 characters.

The future scope of this work lies to expand this CAPTCHA recognition system for larger and more noisy CAPTCHA containing all the symbols possible.


SUGGESTION


ANY
ideas?