Adversarial Attacks

First Assessment

Adhithya S.
Janaki Keerthi
Jayasoorya Jithendra
Jessal V. A.

under the guidance of **Prof. Rajasree R.**

Table of contents

1. Introduction
2. Phases
3. Works completed till date
4. Tools or Frameworks
5. Next Step

Introduction

Introduction

- An adversarial attack consists of subtly modifying an image such that the changes are almost undetectable to the human eye.
- When submitted to a classifier the adversarial image is misclassified, while the original one is correctly classified.

Introduction

- The aim of our project is to make use of this vulnerability of neural networks in a constructive manner to improve computer security.
- Adversarial examples for captcha applications are very difficult for Deep Learning algorithms while easy for humans (adversarial noise tends to be small and does not affect human perception of image content).

Phases

Phases

- Captcha Generation : A captcha generator is used to give an input captcha to the system. The captcha consists of 6 letters in text format.
- Image Modification : This is where adversarial noise is added to the image using FGSM. The modified image will be indistinguishable to human eye.

Phases

- CNN Model : Modified image is given to a Convolutional Neural Network model and prediction is made. Predicted value is compared with the original value and performance of system is evaluated.
- Output : Output is a modified captcha image, which is indistinguishable to human eye, but wrongly predicted by the CNN model.

Works completed till date

Captcha Dataset Generation

- A python script has been used to generate captchas.
- The script can be used to generate captchas consisting of lowercase alphabets or uppercase alphabets or digits. It also allows us to generate captchas of different lengths.
- For our project, we have decided to restrict the captcha length to 6 lowercase characters.
- We have successfully generated 10 million captchas.

Generated Captchas

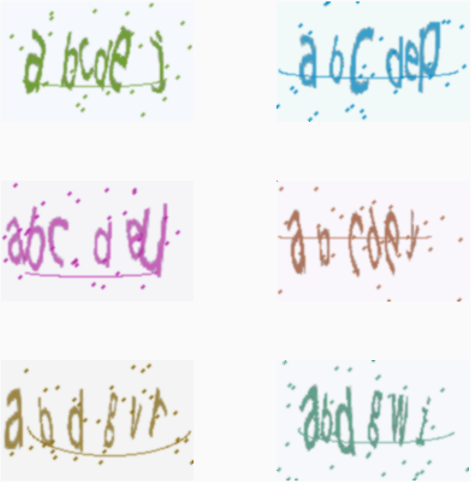


Figure 1: Examples

Tools or Frameworks

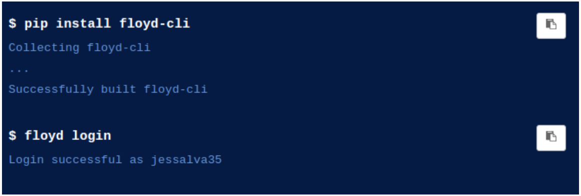
Tools or Frameworks

- **Jupyter Notebook :** For creating and executing our python code
- **TensorFlow :** Software library for high performance numerical computation
- **FloydHub :** Cloud training platform for training our model
- **floyd-cli :** Command-line for FloydHub

FloydHub

- The model on which the attack is based, will be trained on the cloud.
- Paperspace, FloydHub and AWS were the targeted cloud training platform.
- At the moment, FloydHub provides a good free platform for students on which we will try to train our model.
- If the GPU support and training is not going according to plan. Alternatives will be explored.

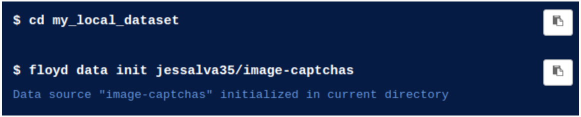
floyd-cli



```
$ pip install floyd-cli
Collecting floyd-cli
...
Successfully built floyd-cli

$ floyd login
Login successful as jessalva35
```

- floyd-cli was successfully installed on the system.



```
$ cd my_local_dataset

$ floyd data init jessalva35/image-captchas
Data source "image-captchas" initialized in current directory
```

Next Step

Next Step

- The dataset will be uploaded.
- Training will be done in the cloud GPU.
- Once we have attained the required accuracy, we can download the weights and use them in our system.
- We can extract the gradients using the trained model’s weights to perform attack on the dataset.

Thank You