Adversarial Attacks

Requirement Specification Document

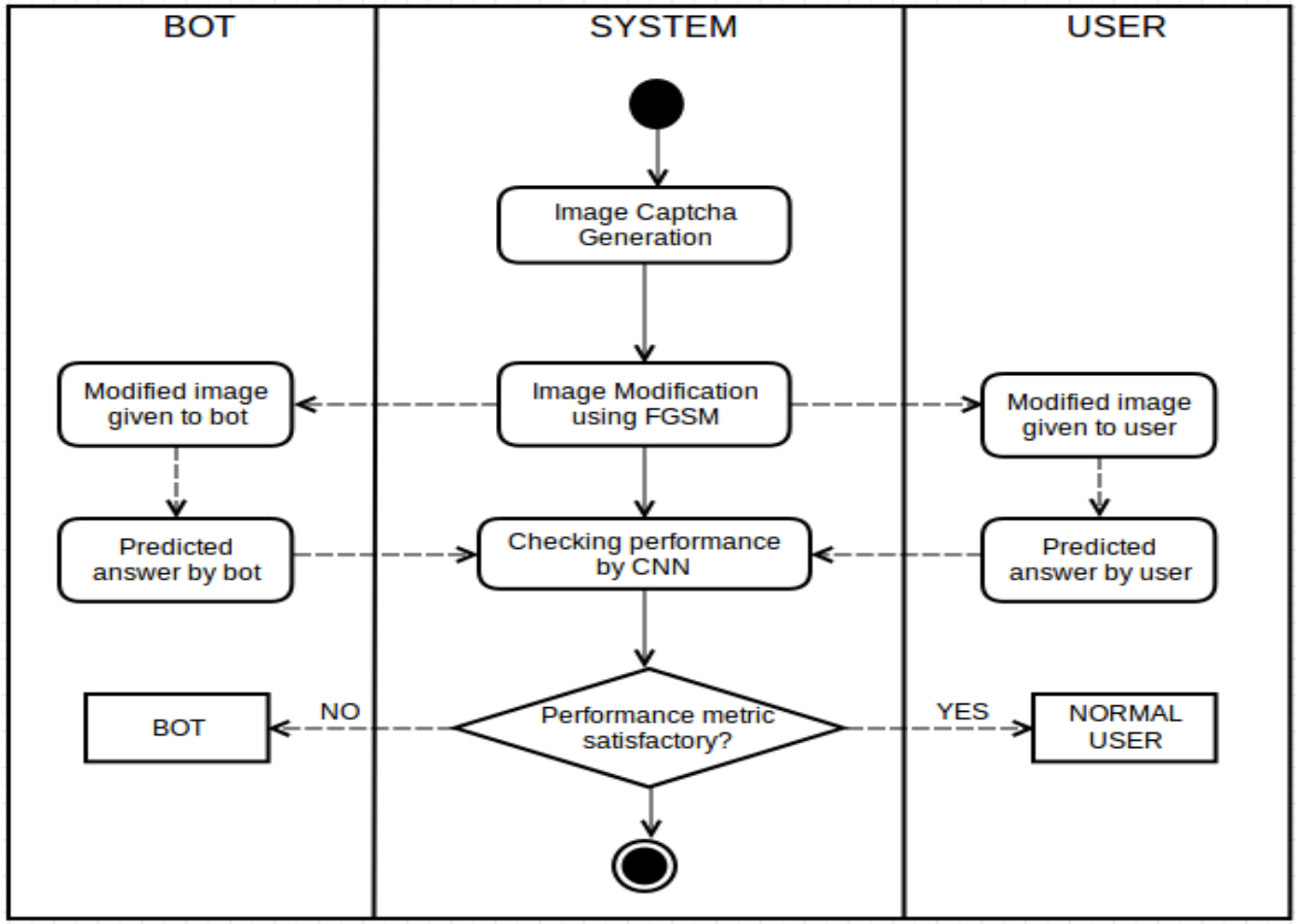
This **Requirement Specification Document** has been prepared by Jessal V. A., Adhithya S., Janaki Keerthi and Jayasoorya Jithendra under the guidance of **Prof. Rajasree R.**, Assistant Professor, College of Engineering Trivandrum.

Contents

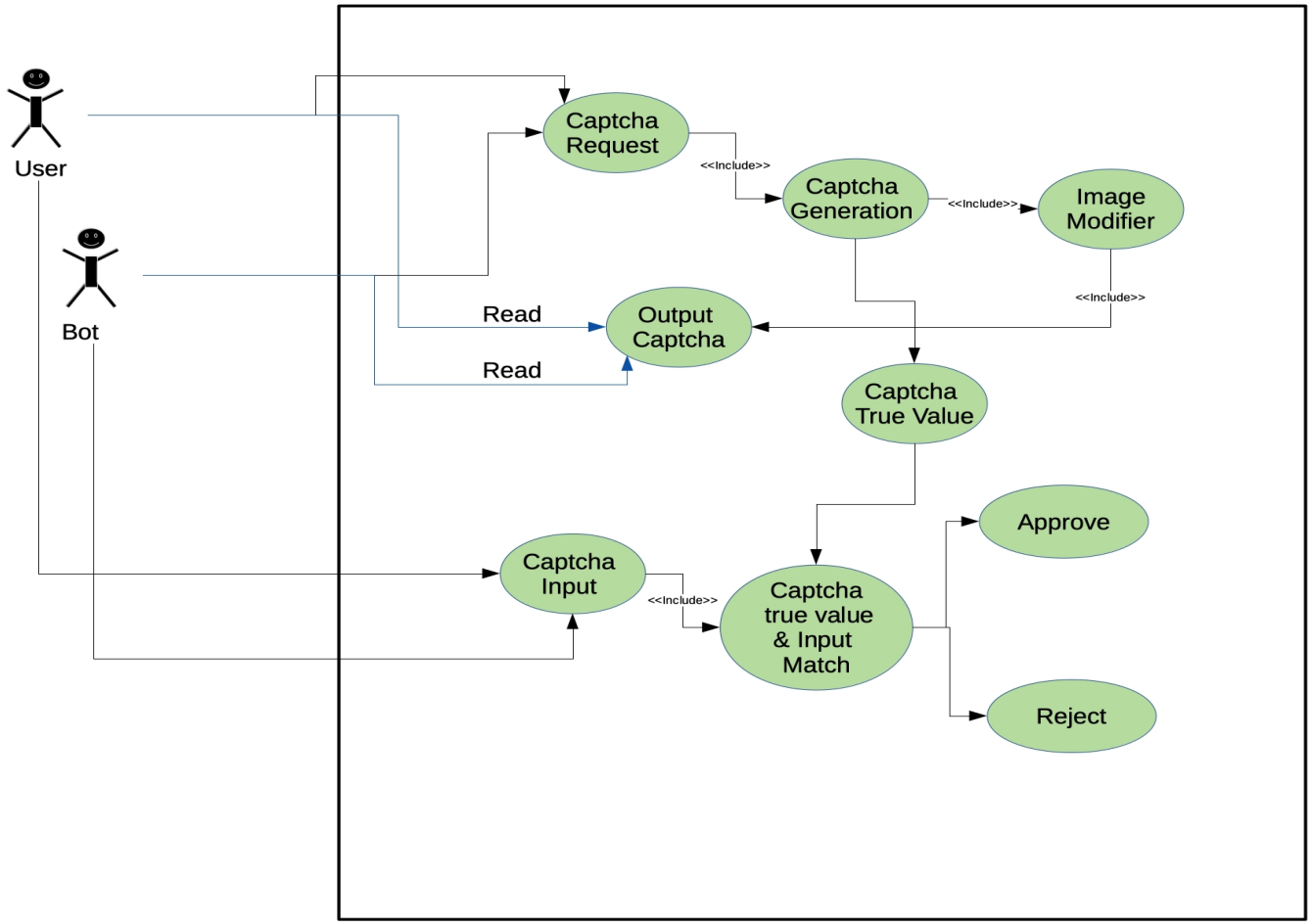
Introduction

- Adversarial attacks on images refers to applying small perturbations on them so that they are misclassified by the neural network
- The model predicts an incorrect answer with high confidence
- But the attack can be constructively used in image captchas for increased security

Activity Diagram



Model


User
Bot
Captcha Request
<<Include>>
Captcha Generation
<<Include>>
Image Modifier
<<Include>>
Read
Output Captcha
Read
Captcha True Value
Approve
Captcha Input
<<Include>>
Captcha true value & Input Match
Reject

Functional Requirements

- The adversarial captcha is given to the CNN model and it will predict a corresponding output.
- Performance assessment can be done by comparing the predicted value and original value of captcha.

Input

- Captcha generator is used to give an input captcha to the system.
- The captcha consist of six characters in image format

Image Modification

- Image modifier will add noise to the given captcha.
- Noise is generated using FGSM algorithm.
- The modified image will be indistinguishable to human eye.

CNN Model

- Modified image is given to a CNN 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.

Physical Environment

- We will train our model in cloud
- Require a programming environment which support TensorFlow framework and allows GPU training

Interfaces

- The user will be given a web application to interact with.
- The web application will provide the user with a random captcha generated real time.
- The user will be given a text box to give the response to the displayed captcha.

Users

- The system can be used by an individual or organization who wants to secure their image from image recognition.
- The user will only have to provide the data or dataset on which attack has to be performed on.

Process Constraints

- CNN model used for the process has to trained to recognize the characters from an image captcha close to optimal bayes error.
- Training a CNN model is a difficult task as it requires a lot of computational power which can be harnessed with the help of Cloud GPU .

Quality Constraints

- Training requires high computational power
- The probability of the randomly guessing each character of the captcha is $(\frac{1}{26})^6$.
- This comes out to be about 1% probability.
- The attack of will be successful if the model's accuracy can be reduced to atleast 10%.

References

Ian J. Goodfellow, Jonathon Shlens Christian Szegedy (2015)

"Explaining And Harnessing Adversarial Examples"

Gamaleldin F. Elsayed, Ian Goodfellow, Jascha Sohl-Dickstein (2018)

"Adversarial Reprogramming of Neural Networks"

F. Tramèr et al. (2017)

"Ensemble Adversarial Training : Attacks and Defenses"

Hassan Gomaa (2011)

"Software Modelling And Design-UML, Use Cases, Patterns, and Software Architectures"