ADVERSARIAL ATTACK ON IMAGE CAPTCHAS

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

Guided by Prof. Rajasree R.

ABSTRACT

An adversarial attack consists of subtly modifying an image such that the changes are almost undetectable to the human eye but the modified image when submitted to a classifier, is misclassified. At the same time, the original one is correctly classified. The project aims to make use of this susceptibility of Deep Neural Networks to adversarial noise, to prevent bots from automating captcha tests.

PROBLEM STATEMENT

Perform adversarial attack on Image Captchas to prevent bots from automating captcha tests.

BASIC MODEL ARCHITECTURE

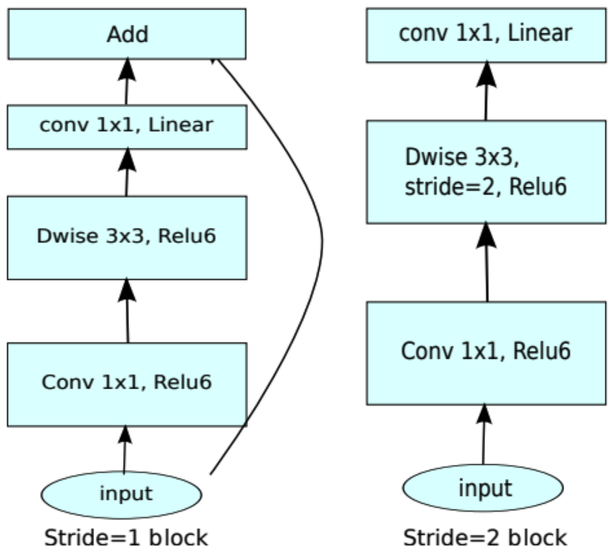


Fig. 1: MobileNet V2 architecture

METHOD AND WORKING

The adversarial noise is added using the Fast Gradient Sign Method (FGSM). It computes an adversarial image by adding a pixel-wide perturbation of magnitude in the direction of the gradient. This perturbation is computed with a single step, thus is very efficient in terms of computation time.

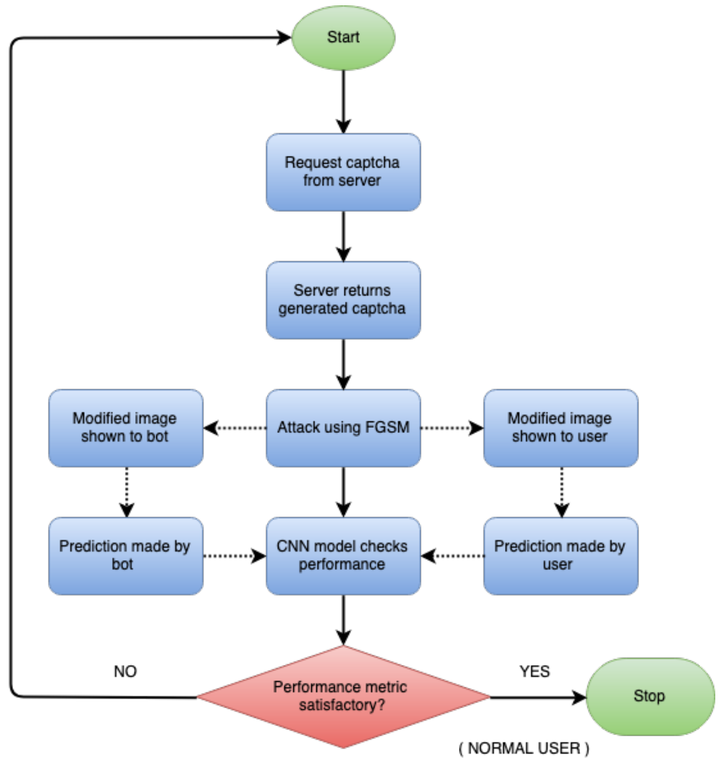


Fig. 2: Flowchart

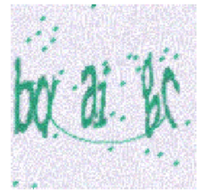
NOVELTIES AND CHALLENGES

- On the applied side, no one has yet designed a system capable of securing captchas by making use of adversarial attack algorithms.
- Noise added should be restricted such that the image is still intelligible to humans. Thus the noise limit is a hyperparameter that must be tuned.

APPLICATIONS

- The model can be used to secure captchas from being predicted by AI algorithms.
- The model can be used to secure social media pictures from being automatically tagged by bots.

RESULT



Original Label :bdaigc
Predicted Label :bdaigc

Predicted Label :aaaaaa

Fig. 3: (a) Before the attack (b) After the attack

REFERENCES

[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing Adversarial Examples”. In: *Conference paper at ICLR* (Mar. 2015).

[2] Shixiang Gu and Luca Rigazio. “Towards deep neural network architectures robust to adversarial examples”. In: *arXiv preprint arXiv:1412.5068* (2014).

[3] Florian Tramèr et al. “Ensemble Adversarial Training: Attacks and Defenses”. In: *Conference paper at ICLR* (July 2018).

Department of Computer Science and Engineering
College of Engineering Trivandrum