Evaluating Effectiveness of Adversarial Examples on State of Art License Plate Recognition Models

Kanishk Rana[§], Rahul Madaan[§]
Indraprastha Institute of Information Technology - Delhi
Email: {kanishk17241, rahul17179}@iiitd.ac.in

***Abstract*—Deep learning takes advantage of large datasets and computationally efficient training algorithms to outperform other approaches at various machine learning tasks. However,DNNs are vulnerable to adversarial examples due to imperfections in the training phase. In this work, we generate some adversarial examples to test their effectiveness against the state-of-the-art License Plate Recognition (LPR) models.**

***Index Terms*—Perturbation methods, Neural networks, Adversarial examples, Image denoising, Security**

I. Introduction

The increasing use of deep learning creates incentives for adversaries to manipulate DNNs to misclassify input. The discovery of adversarial examples has drawn attention and raised concerns among the domain researchers due to security issues it could create. In this work, we evaluate the robustness of the state-of-the-art LPR models - Chinese City Parking Dataset (CCPD) [1] and Automatic License Plate Recognition (ALPR) [2], against some of the adversarial examples. LPR has applications in parking automation and parking security, Road Tolling, Border Control, and Law Enforcement. With our work, we want to show that an adversary would be looking to tweak the images in a simplistic manner that would cause the maximum damage, and lead to a severe security breach. A large body of work has been developed on defensive models but a major drawback with them is that they are heuristic and fail to guarantee robustness. Recent works like [3], [4], [5] and [6] used gradient estimation and simple datasets like CIFAR10, MNIST, which had class labels. In the case of LPR, class labels are not present. Our paper focuses on creating adversarial examples for LPR models without prior knowledge of the model's inner working. The examples presented here could be divided into two types of attacks. In one case, the model cannot identify any License Plate, and in the second case, it identifies the License Plate but misreads. In the first case, systems present could raise the alarm if the License Plate is not recognized; therefore, such types of attacks could be usually avoided or captured. The second scenario is more dangerous since only human intervention would correct the mistake. In addition to this, we will pre-process the image using state of the art algorithms for denoising and test the effectiveness against the crafted adversarial examples. Also, in this paper, for one of the examples (Overlay), we will propose a probable defense mechanism. This attack also gives an insight into a future research area in the form of handling variable size images in DNN, which was a weakness exploited with this type of attack.

Our contributions with this paper are:

- Adversarial Examples for which any defense is not available/explored.
- Adversarial Examples that are not specific to a particular model with excellent results against two state of art models.
- Performance of models after using a state of art defense if available against the proposed adversarial attacks.

With the usage of Automatic Number Plate Recognition (ANPR) in critical places, especially in law enforcement, these systems' robustness is of utmost importance. It should be given due importance while developing. The work presented in our belief is the first one in crafting adversarial examples and evaluating the robustness of LPR models.

II. Proposed Approach

The approach of modification of images which leads to misclassification is based on the following weakness of DNN.

A. Optimization

The idea of optimization of the input is to extract some pattern in an image which was not observable earlier. This same idea can be used to fool the DNN. For example, let's say we have an object detector model at our disposal. Then, by maximizing the log likelihood of another class (other than the correct class), with minimal changes to the image (using backpropagation), we can get the model to predict that the input image is actually of another class. The change in the input is to be such that it is near invisible to the human eye. But the model predicts wrong output from the actual image.

B. Dimensionality

Images belong to a high dimensional space R^{nxn}. So even if we feed our model a million or 10 million images, there exists a very high number of samples which the model has not been trained on. Out of all the matrices that exist in this space, there is only a small number of matrices that represent a natural image which are normally fed to the model. And there remains a lot of random images which don't represent anything to humans.

When the model is fed the training images, it forms a decision boundary using the knowledge from these images only. This

[§] Equal contribution

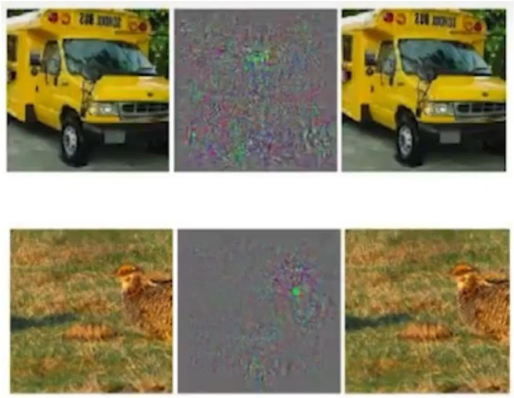


Fig. 1: Noise Addition Invisible to Human Eye

boundary can then be used to classify any other input fed to the model including but not limited to random images. Using the decision boundary, the model has decided upon a class for every possible image, i.e., every possible matrix existing in the bigger space, that can be received by the model. We can use this to our advantage and fool the model in predicting the wrong output by either optimizing an already existing example or creating new examples.

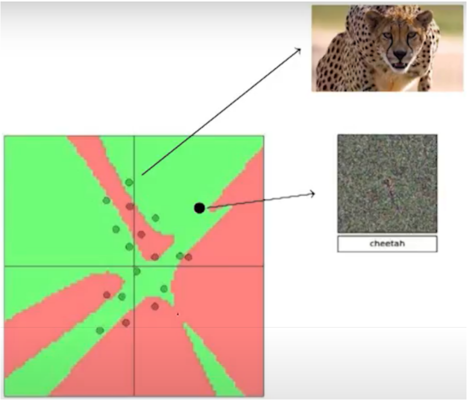


Fig. 2: Exploiting the fixed decision boundary to predict random image as Cheetah

Our adversarial examples are based on the following changes to an input image:

- Adding noise to a particular region
- Adding haze to the image
- Overlaying with adversarial image
- Rotation in images
- Blurring the images

Since two of our techniques, namely adding noise and haze, could be subjected to denoising methods, we employ state of the art denoising method, which originally were not part of the papers we tested. After denoising, the overall result is shown in Fig 3 and Fig 4. For another class of examples, we used blurring of the image. We can examine the distribution of the low and high frequencies using the Fast Fourier Transform of the image. If there is a low amount of high frequencies then image can be considered blurry, where this amount is a relative number designated by the threshold specified by the administrator of the LPR system. Instead, [7] reviews many methods to compute this "blurriness metric," some of the simple and straightforward using just basic grayscale pixel intensity statistics, others more advanced and feature-based, evaluating the Local Binary Patterns of an image. We estimate the blurriness of the image used in our examples using the Local Binary Patterns, and we find that a simple motion blur is enough to evade the methods of detecting method. The metric used in this method is interpreted as higher the score, the more precise the image.

III. Results and Inference

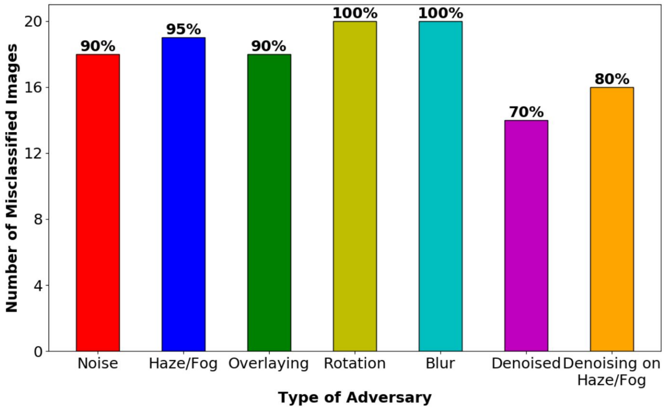


Fig. 3: Statistics on ALPR Dataset

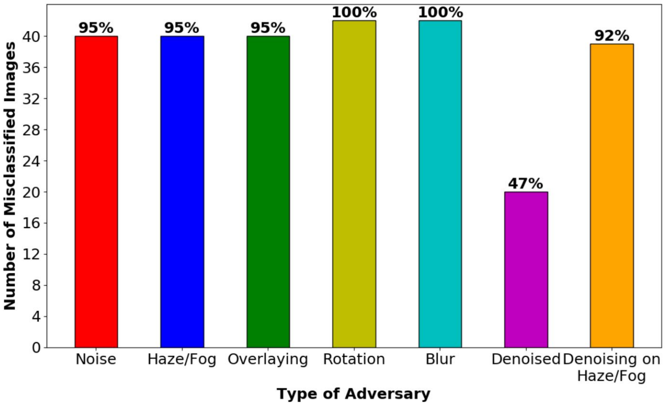
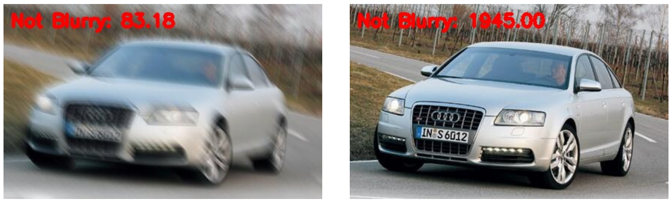


Fig. 4: Statistics on CCPD Dataset

We present the results in the form of a percentage of samples misclassified and these samples are the ones that were used originally in the research paper mentioned above. The percentage will give us insights into what type of attacks were highly successful, and what type of input is hard to modify minimally. Adding simple noise and haze to a particular area or an image as a whole is a practical yet outdated technique. There are several state of the art denoising methods that could be employed to defend against such attacks, which was evident from the decreased success after we used state of the art



(a) Blurred Image (b) Original Image

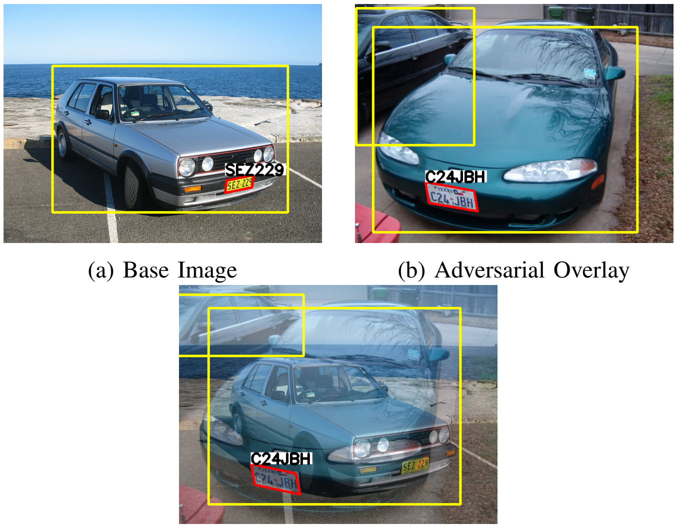
Fig. 5: Calculation of Blur

denoising model in addition to the LPR model. Even the defense against blur could be evaded, as is shown in the images. On the other hand, adversarial examples of a simple rotation and overlay being lightweight and straightforward, break entirely the underlying models, which indicates a reliable and straightforward method for an adversary to work on. For protecting against a successful overlay, a random image size could be employed, but the underlying model has to be robust to process different sized images.

IV. Conclusion and Future Work

There are several applications of LPR, which includes parking ticket generation, traffic rule violation, hit and run cases. A severe offense like these could go unnoticed if an adversary gains access to change the Deep Learning model's input. This is why we need to have a strong defense against adversarial examples. One essential future work could be in the domain of detecting rotation in an image and feeding of variable size images to DNN. In this paper, we investigated the effect of adversarial samples on the state of the art LPR models. We also presented the success rate of the adversarial samples when mitigation measures are employed.

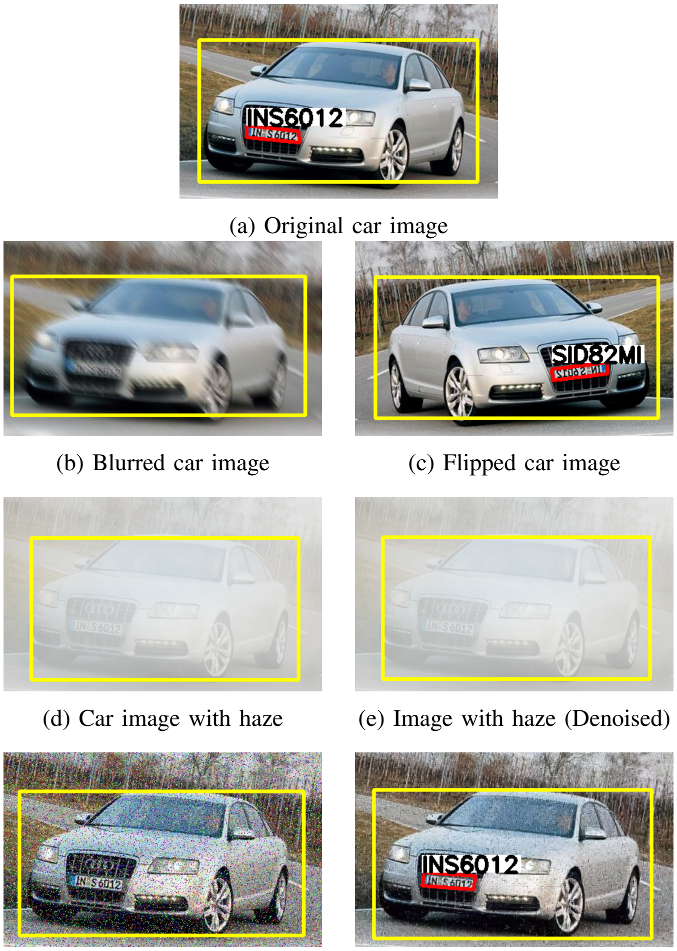
V. Images



(a) Base Image (b) Adversarial Overlay

(c) Result

Fig. 6: License Plate Recognition with Overlay



(a) Original car image

(b) Blurred car image (c) Flipped car image

(d) Car image with haze (e) Image with haze (Denoised)

(f) Car image with noise (g) Image with noise (Denoised)

Fig. 7: License Plate Recognition on an example

References

[1] Zhenbo Xu, Wei Yang, Ajin Meng, Nanxue Lu, Huan Huang, Changchun Ying and Liusheng Huang, 2018. Towards End-to-End License Plate Detection and Recognition: A Large Dataset and Baseline.

[2] S. M. Silva and C. R. Jung, 2018. License Plate Detection and Recognition in Unconstrained Scenarios.

[3] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh, 2017. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models.

[4] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh, 2018. Query-efficient hard-label black-box attack: An optimization-based approach.

[5] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song, 2018. Practical black-box attacks on deep neural networks using efficient query mechanisms.

[6] Yali Du, Meng Fang, Jinfeng Yi, Jun Cheng, and Dacheng Tao, 2018. Towards query efficient black-box attacks: An input-free perspective.

[7] Pertuz, Said, Domenec Puig and Miguel Ángel García, 2013. Analysis of focus measure operators for shape-from-focus.