CS230

Evaluating the Effectiveness of Adversarial Attacks on Traffic Light Color Classification Deep Neural Networks

Eunji Lee
Department of Computer Science
Stanford University
elee0324@stanford.edu

Connor Toups
Department of Symbolic Systems
Stanford University
ctoups22@stanford.edu

Spencer Paul
Department of Computer Science
Stanford University
spaul2@stanford.edu

Project Category: Computer Vision

Abstract

Traffic Light Recognition is a critical task in autonomous and assisted driving, but the literature on adversarial attacks against these models is woefully lacking. We adopt a two-stage traffic light recognition model with two CNNs working to localize and classify traffic lights. We first evaluate adversarial attacks against the classification model. We then attempt to fool the localization model by perturbing a small region of the original input region. Our adversarial attacks were unsuccesful in attacking either model, a major contrast to the one other paper on adversarial attacks against traffic light recognition models. We suggest possible reasons for our contrasting results, and we offer suggestions for additional work to be to assess the robustness of traffic light recognition models.

1 Problem Description

Detection of the color of traffic lights is a very important task in autonomous driving, as well as in assisted driving tools. Convolutional neural networks (CNNs) have been shown to be very effective in many image classification tasks, including traffic light color detection [1]. However, CNNs are very susceptible to adversarial attacks (i.e. small perturbations to the input images that can be imperceptible to the human eye that can cause a neural network to misclassify the image) [2]. The consequences of an adversarial attack that causes a light detection algorithm in an autonomous vehicle to misclassify a red light as a green light could potentially be devastating for passengers and pedestrians [1]. Given the importance of traffic light color detection, we wanted to further investigate which aspects of traffic light detection models are susceptible to attack.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related work

Some existing research has suggested that adversarial attacks are effective against traffic light color classification CNNs; furthermore, traditional adversarial attack defenses, such as adversarial training and defensive distillation, have shown to be ineffective at reducing the success rate of the attack [1].

Although some research on the effectiveness of adversarial attacks and defenses on the classification of traffic *signs* has been done [3], there has been less research looking at traffic *lights*. A 2020 study from Wan et al. on the effectiveness of the spatial, one-pixel, Carlini & Wagner, and boundary attacks against a deep convolutional network trained on a CARLA traffic light dataset found that the classification model in traffic light recognition was susceptible to both white box and black box attacks [1]. A limitation to this study is the fact that the data was hand-collected from the CARLA traffic simulator and consisted of only 477 manually localized/cropped images.

One aspect of Wan and others' research that we seek to explore further is which aspects of the recognition model are susceptible to adversarial attacks. Wan et al. only examines the classification model, and does not consider the localization model at all; however, localization – the process of finding a specific object, like a traffic light, in an image containing many different objects – is a critical element of traffic light recognition systems. Wan et al. only use manually cropped images that just include the traffic light [1]. We also orient our attacks around minimizing the amount of perturbation to the images, both in terms number of pixels edited and intensity of changes to those pixels; this draws upon a large literature in adversarial attacks that emphasize the importance of minimizing the visibility of the attack, especially since several defense algorithms work to recognize perturbations in the image [2, 4].

3 Dataset

We are using a dataset collected from CARLA, a driving simulator. The dataset was manually collected by researchers at Affinis Labs. The dataset consists of 1800 images of red and green lights. The split is roughly even between red and green lights. We did notice, however, that some photos that appeared "Yellow" to us were annotated by the simulator as "Green." The dataset can be found at this public Google Drive link `https://drive.google.com/drive/folders/1_RppuNf7LSBJlE2v9d_3iuIGkeWEs7Rn?usp`. We also used a dataset of manually localized traffic light images collected by Dr. Kyle Guan at Bell Labs. That dataset can be accessed at `https://github.com/kcg2015/traffic_light_detection_classification/tree/master/traffic_light_classification/training_images`.

Figure 1: An example image from our dataset

4 Model

We use an existing code base for traffic light localization and classification, which has been trained on the COCO dataset (for localization) and a separate dataset of 1,400 traffic light images (for classification). The GitHub link for the reference code base can be found at: `https://github.com/kcg2015/traffic_light_detection_classification`.

Our model, like many traffic light detection models, has two stages that run synchronously. In the first stage, our localization model – based on the SSD framework trained on the COCO dataset – takes in a 900x1600 image from the Carla Autonomous Vehicle Simulator dataset. It then attempts to localize the traffic in that image. If it finds a traffic light in the image with high enough confidence, we take that localized region and reshape it into a $32x32$ image. We then pass this $32x32$ to our classification model, which classifies the traffic light as either green, red, or yellow.

Pre-trained weights are loaded from the code base. The classification convolutional neural network model architecture can be found below in Figure 2.

Layer (type)	Output Shape	Param #
conv2d_4 (Conv2D)	(None, 32, 32, 16)	448
activation_6 (Activation)	(None, 32, 32, 16)	0
conv2d_5 (Conv2D)	(None, 30, 30, 16)	2320
activation_7 (Activation)	(None, 30, 30, 16)	0
conv2d_6 (Conv2D)	(None, 28, 28, 16)	2320
activation_8 (Activation)	(None, 28, 28, 16)	0
max_pooling2d_2 (MaxPooling2	(None, 14, 14, 16)	0
flatten_2 (Flatten)	(None, 3136)	0
dense_3 (Dense)	(None, 128)	401536
activation_9 (Activation)	(None, 128)	0
dense_4 (Dense)	(None, 3)	387
activation_10 (Activation)	(None, 3)	0

Total params: 407,011
Trainable params: 407,011
Non-trainable params: 0

Figure 2: Classifcation Convolutional Neural Network Model Architecture

5 Experiment Description

The only existing work we found on adversarial attacks on Traffic Light Detection models only examined the classification aspect of traffic light detection. They took manually localized images of traffic lights, trained a model to classify the color of the traffic light, and then ran adversarial attacks against that model. They found that numerous white-box adversarial attacks could be extremely effective against the classification model – some reducing the model's accuracy from 100% to 0%.

We were struck by this finding, and our first experiment was to replicate their finding on our own model. We took manually localized traffic light images, collected by Dr. Kyle Guan, and evaluated the accuracy of our classification model on these images; we then applied FGSM and Carlini Wagner distortions to these images and evaluated the accuracy of our model on the perturbed images. Additionally, we attempted adversarial attacks on our classification model running with our localization model. We ran our two-step model with localization on our CARLA dataset to get our baseline accuracy. Then, we ran our two-step model again; however, before we passed the cropped, localized image from the localization model to the classification model, we applied FGSM and Carlini

& Wagner attacks to the localized image, and then passed these perturbed images to our classification model.

For our second experiment, we wanted to move away from just attacking the simpler classification model and instead consider how one could attack the broader two-stage, localization-classification model. Instead of simply perturbing the entire 900x1600 pixel image given to the localization model, we only perturb the region that the localization model outputs. Our goal, consistent with the goal of adversarial attacks in general, is to minimize the perturbation of the input image while still yielding successful attacks. By perturbing only the region of the image that needs to be localized, we can avoid perturbing the entire image while hopefully still yielding an effective attack. In this experiment, we run the localization model, which returns the bounding coordinates of the localized region. We then apply attacks to this region, and replace the region in our original input image with our perturbed image. We then run the localization model yet again on this partially perturbed 900x1600 image; if the localization model outputs a different set of coordinates, we consider that a successful attack. We also consider true classification accuracy (whether or not the model outputs the correct color) as a measure of attack success. We choose to use the output of the localization model as our region to perturb as opposed to the true bounding region of the traffic light for two reasons: it allows us to assess the effectiveness of the attack on every single test image and because our localization model is taken from a public repository and therefore could be used by anyone – thereby better replicating how an adversarial attack in the real world might occur.

6 Results

6.1 Classification Model Attacks

We found drastically different results than those of Wan Et Al. While they found that adversarial attacks could be effective on the classifcation model, we found that both FGSM and Carlini & Wagner attacks were entirely ineffective. We created adversarial examples from the manually localized images as well as the model localized images. In both cases, we were entirely unable to fool the model; the accuracy on the adversarial examples was the exact same as the accuracy on the original examples.

Image Type	Attack	Accuracy without Attack	Accuracy with Attack
Manually Localized	FGSM	99%	99%
Manually Localized	C&W	99%	99%
Model Localized	FGSM	73%	73%
Model Localized	C&W	73%	73%

6.2 Modifying Original Image

Similarly, our attempts to fool the model by perturbing the region the model-localized region in the original input also failed to generate any successful attacks. Our localization model and classification model outputted the same exact results on unperturbed and partially-perturbed photos.

7 Discussion

7.1 Issues

In examining why our experiments failed to yield the same results as Wan Et Al. we came to two conclusions. Firstly, it's possible there is an issue with the Adversarial Robustness Toolbox in properly accessing the gradients of our Keras Model. We were unable to notice differences between our perturbed images and our original images (Appendix Figure 3). Admittedly, this is hard to do for our photos since they don't have many pixels, and they are generally dominated by two colors (black + traffic light color). However, we also applied boundary attacks, which only require final class prediction, to our images, and we were similarly unable to fool the model.

Secondly, our attempt to fool the localization model by only perturbing one region did not work; in hindsight, this is because we need to use the gradients of the localization model as opposed to the gradients of the classification model. Unfortunately, our frozen tensorflow graph, used to load in the localization model, was not compatible with the ART toolbox, and thus we were unable to develop adversarial attacks against it.

7.2 Takeaways and Suggestions

While we consider our results somewhat inconclusive, we hope that other researchers will be inspired to replicate these results yet again – given the discrepancy between our findings and others' findings. Furthermore, we encourage other researchers to consider the localization model, which – to date – has not yet been explored as a potential vulernability in traffic light recognition models. We believe that our idea of only perturbing the region that the localization model will output still stands as a viable way of generating successful attacks while minimzing the perturbation to the image space. We suggest that these perturbations be generated with respect to the loss from the localization model – not from the classification model.

8 Contributions

Eunji Lee and Connor Toups both worked on developing the baseline models and developing the adversarial attacks, with Eunji dealing more with the adversarial attacks and Connor dealing more with the baseline model. Spencer Paul experimented with defense algorithms. All members contributed equally to developing our project idea, analyzing the results, and writing the final report.

References

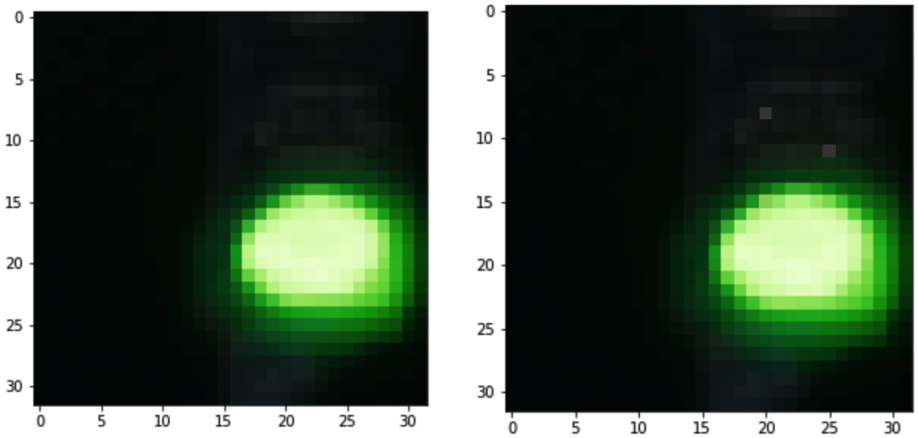
[1] Morris Wan, Meng Han, Lin Li, Zhigang Li, and Selena He. Effects of and defenses against adversarial attacks on a traffic light classification cnn. In *Proceedings of the 2020 ACM Southeast Conference*, pages 94–99, 2020.

[2] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 2020.

[3] Nir Morgulis, Alexander Kreines, Shachar Mendelowitz, and Yuval Weisglass. Fooling a real car with adversarial traffic signs. *arXiv preprint arXiv:1907.00374*, 2019.

[4] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346 – 360, 2020.

9 Appendix



(a) Unperturbed localized image (b) Perturbed (FGSM) localized image

Figure 3: Example of perturbation that seemed to have minimal effect