# Defending Convolutional Neural Network-Based Object Detectors Against Adversarial Attacks

Victor Hu[1], Jeffrey Cheng[2]
[1]Watchung Hills Regional High School [2]Bridgewater Raritan Regional High School

## Abstract

Convolutional neural networks are by nature susceptible to adversarial examples. In safety-critical systems, such as autonomous vehicles, it is paramount that object detection is resistant to adversarial attacks. We generated adversarial examples that successfully caused real-time object detectors to misclassify road signs as other objects, a scenario where misclassification could result in damage and loss of life. In addition, we proposed defenses to mitigate misclassification. First, to prove that CNN-based object detectors are capable of reliably classifying stop signs, we tested the YOLOv3 object detector with normal stop signs as well as stop signs with sticker graffiti. A Raspberry PI car with a front-facing camera was used to simulate a passing car, reproducing dynamic perspective and lighting conditions. The car successfully detected a normal stop sign in 100% of the video frames and a stop sign with graffiti in 89.02% of the video frames across three trials. We then tested YOLOv3 with our adversarial attack, which lowered "stop sign" detection rates to 58.74% and increased faulty "person" misdetection rates to 66.90%. Implementing defenses such as color thresholding and classification based on Haar features returned "stop sign" detection rates back up to over 99%. Our work shows that adversarial attacks are substantial threats to the safety of autonomous vehicles, but their effects can be mitigated by using a variety of defense methods.

## Acknowledgements

Thank you to our mentors, Elizabeth and Dennis Mabrey, for their support and advice on the presentation of our project.

## Background

Convolutional neural networks, known for their most common application of image classification, are currently being studied for their usage in autonomous vehicles to identify and react to objects in the roadside environment. In such cases, where a missed traffic sign could result in harm, damages, or death, their chances of failure must be minimized. Adversarial attacks pose a threat to the safety of autonomous vehicles that rely on convolutional neural network-based object detectors.

Object detectors, unlike standalone convolutional neural networks, have the additional task of detecting where objects are located in an image before classifying them. This makes their job exponentially more difficult, as they must locate and take into consideration thousands of possible regions-of-interest within a single frame. We focused our research around the YOLOv3 object detector, since it is capable of running in real time. YOLOv3, compared to its predecessors, saw great improvements in detecting small objects due to the use of multi-scale detection.

Adversarial examples are inputs which look more-or-less "ordinary" to a human, but actually cause a neural network to produce the wrong classification. By freezing the weights and biases of a neural network, adversarial examples can be trained by using backwards propagation and optimization functions to minimize a desired loss function.

**Figure 1** below shows an example of an adversarial attack. Creating adversarial attacks that succeed in a physical environment are a larger challenge due to the variety of distortions in the real world, such as changes to lighting, angle, scale, and rotation, as well as camera noise. Pixel-perfect adversarial noise, such as the one depicted above, would fail to translate into the real world.

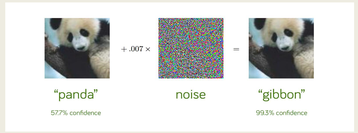
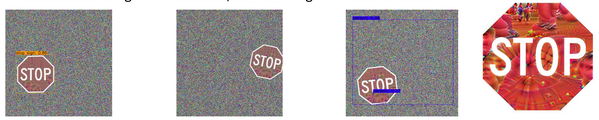



**Figure 1.** An example of an attack that generates adversarial pixel-noise atop an image to fool a classifier. The adversarial image looks seemingly identical to the original. These types of attacks are not robust enough to succeed in a physical environment. Image taken from [11].

## Procedure: Attack Generation Process

- Used the Expectation over Transformation method [9]
  - Generates random lighting, scalar, and rotational transformations of attack
- Modified the *ShapeShifter* attack [2] to target YOLOv3 trained on the MS COCO dataset
- For loss function of adversarial attack generator:
  - Classification probabilities across the 80 classes fed into a softmax function - creates mutually exclusive class probabilities
  - Cross-entropy loss computed with the target goal of 100% Person
  - To control the redness of the adversarial stop sign: L2 loss computed between pixels of a red stop sign mask and pixels of the adversarial stop sign
  - Losses were multiplied with constants to control the weight of the redness loss in relation to the "Person" classification loss - found by trial and error
  - Only medium and large detection scales for YOLOv3 included in loss function
- Adversarial stop sign generated trained for: 250 iterations, learning rate of 1, "Person" classification weight constant of 8, redness weight constant of 0.003




## Procedure: Suggested Defenses

- First proposed defense - **Color Thresholding**
  - Chromatic equivalent to image binarization - convolutional neural networks take color into account when making inferences
  - **Goal**: maximize the uniform redness of a stop sign while trying to avoid altering the rest of the image
  - **Strategy**: pixel values are "snapped" to red if the red-green or red-blue ratio is sufficiently high
    - Insert color thresholding step into the image preprocessing pipeline
    - Minimizes the human-like features that causes the object detector to misclassify the adversarial stop sign as human
- Second proposed defense - **Haar Features**
  - **Goal:** error-checking the inferences of the neural network using haar features
    - Haar features based on finding the average whiteness and blackness of different areas in the image - see depiction to right
    - E.g. facial detection
      - Bridge of a nose is lighter than the sides of a nose - area of an image that is more white in the middle and more black on the sides -> face
  - **Strategy:** Haar Cascade classifier uses a sliding window to test an area of an image for many of these haar features
    - Take positions of bounding boxes produced by YOLOv3 of "Person" classifications
    - Perform stop sign haar cascade classifier on an expanded region of interest - no need to search the entire image
    - If area passes all stages of the test -> object detected
    - Haar features are based off of average white and black areas - adversarial perturbations would have little effect upon them
  - **Limitation:** haar cascade classifiers are single class - features are exclusively trained for stop signs

- Raspberry PI mounted on an Arduino robot car used to mimic changing scale, lighting, and angle conditions of real driving
- Car was placed on a track one foot in width
- 5½' by 5½' stop signs printed out and placed one yard in front of car, one inch to side of track
- Raspberry PI Camera Module in front of car recorded trials
  - Post-processed with our object detector and proposed defenses

## Results

| Experiment Configuration | Average "Stop Sign" Frames | Average "Stop Sign" Conf. | Average "Person" Frames | Average "Person" Conf. |
|---|---|---|---|---|
| Regular Stop Sign | 100% (709) | 0.999 | 0% (0) | 0.000 |
| Stop Sign with Sticker Graffiti | 89.02% (605) | 0.746 | 0% (0) | 0.000 |
| Adversarial Stop Sign | 58.74% (358) | 0.545 | 66.90% (407) | 0.535 |

| | Average "Stop Sign" Frames | Average "Stop Sign" Conf. | Average "Person" Frames | Average "Person" Conf. |
|---|---|---|---|---|
| **Color Thresholding Defense** | 99.67% (607) | 0.983 | 0% (0) | 0.000 |

| | Average "Stop Sign" Frames After Defense Checking | Average "Person" Frames After Defense Checking |
|---|---|---|
| **Haar Classifier Defense** | 99.84% (608) | 0.16% (1) |

## Discussion

- Regular stop sign detected in 100% of frames at .999 average confidence level
  - YOLOv3 exceptionally capable of detecting stop signs under normal conditions
  - Handled sticker graffiti well - 90% success rate, no misdetections
  - Silhouette of person did not result in any "person" misdetections
- Adversarial attack effectively lowered success rate
  - More "person" misdetections than correct "stop sign" detections
  - Adversarial attack more effective as car approached sign - importance of successful detection increases with decrease of distance
- Color thresholding
  - 607 successful "stop sign" detections out of 609 frames
  - Not a single person misdetection
- Haar classifier
  - 608 successful "stop sign" detections out of 609 frames
  - One "Person" misdetection

## Conclusions

Ultimately, the results show that our adversarial attack poses a realistic threat in a safety-critical situation like riding in an autonomous car. A stop sign detection rate and confidence level of only around 50% is nowhere near reliable enough for use in real life, in addition to the fact that the object detector would be actively confused with the faulty "person" classifications. However, our research demonstrates that with our proposed defenses against adversarial attacks, stop sign detection and confidence rates return to near-optimal levels.

## Future Work

- Black-box transferability of adversarial attacks between different neural networks
  - Currently, we need access to the target neural network to train the adversarial attack
  - In the future, we wish to work toward using one neural network to train adversarial attacks that attack unknown, but similar neural networks.
- Improving adversarial defenses
  - Find defenses against adversarial attacks that are completely robust and efficient

## References

[1] Thyst, S., Ranst W. V., and Goedemé, T. 2019. Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection. arXiv preprint arXiv:1904.08653.
[2] Chen, S., Cornelius, C., Martin, J., and Chau, D. H. 2019. ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector. arXiv preprint arXiv:1804.05810.
[3] Liu, X., Yang, H., Liu, Z., Song, L., Li, H., and Chen, Y. 2019. DPatch: An Adversarial Patch Attack on Object Detectors. arXiv preprint arXiv:1806.02299.
[4] Zhao, Y., Zhu, H., Liang, R., Shen, Q., Zhang, S., and Chen, K. 2019. Seeing isn't Believing: Towards More Robust Adversarial Attack Towards Real World Object Detectors. arXiv preprint arXiv:1812.10217.
[5] Papernot, N., McDaniel, P., and Goodfellow, I. 2016. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. arXiv preprint arXiv:1605.07277.
[6] Redmond, J., Divvala, S., Girshick, R., Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. arXiv preprint arXiv:1506.02640.
[7] Ren, S., He, K., Girshick, R., Sun, J. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv preprint arXiv:1506.01497.
[8] Redmon, J., Farhadi, A. 2018. YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767.
[9] Athalye, A., Engstrom, L., Ilyas, A., Kwok, K. 2018. Synthesizing Robust Adversarial Examples. arXiv preprint arXiv:1707.07397.
[10] Yuan, X., He, P., Zhu, Q., Li, X. 2018. Adversarial Examples: Attacks and Defenses for Deep Learning. arXiv preprint arXiv:1712.07107.
[11] Goodfellow, I., Shlens, J., Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. arXiv preprint arXiv:1412.6572.
[12] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Tramèr, F., Prakash, A., Kohno, T., Song, D. 2018. Physical Adversarial Examples for Object Detectors. arXiv preprint arXiv:1807.07769.
[13] Eykholt, K., Gupta, S., Prakash, A., Rahmati, A., Vaishnavi, P., Zheng, H. 2019 Robust Classification using Robust Feature Augmentation. arXiv preprint arXiv:1905.10904.