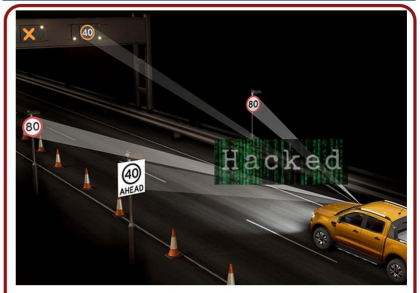
Attacking Autonomous Driving Machine Learning Algorithms with Adversarial Examples

Stephanie Tietz, Koosha Nassiri Nazif {stietz, koosha}@stanford.edu
Department of Electrical Engineering, Stanford University

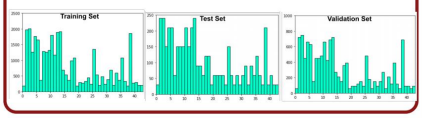
Motivation/Summary



Using GAN, we generate inconspicuous adversarial perturbations to traffic sign images and feed the perturbed images to a traffic sign classifier in an attempt to cause targeted misclassification and a reduction in the accuracy of the classifier.

Dataset

- German Traffic Sign (GTS) Dataset
  - More than 50,000 images in 43 unbalanced classes
  - Image sizes vary between 15x15 to 250x250 pixels (RGB). All data were converted to 32x32 RGB images.
  - 67% training, 8% validation, and 25% testing datasets with similar relative class size

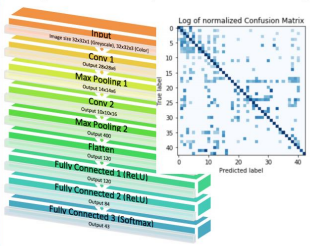


References

1. https://github.com/mohamedameen93/German-Traffic-Sign-Classification-Using-TensorFlow
2. Xiao, C. *et al.* "Generating Adversarial Examples with Adversarial Networks" *IJCAI International Joint Conference on Artificial Intelligence*, 2018–July, 3905–3911.
3. http://benchmark.ini.rub.de/?section=gtsrb&subsection=dataset
4. CIFAR10 Dataset: Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009.
5. Sermanet, P. and LeCun, Y. "Traffic Sign Recognition with Multi-Scale Convolutional Networks." *2011 International Joint Conference on Neural Networks, IEEE*, 2011.
6. Papernot, N. et al. "Distillation as a defense to adversarial perturbations against deep neural networks" *IEEE Symposium on Security and Privacy*, 2016

Models

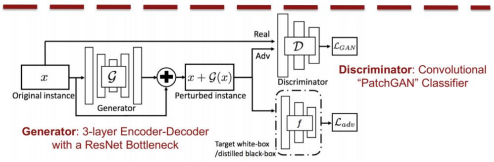
Target Classifier (Based on LeNet-5)



Traffic Sign Classifier Model Description	Validation Accuracy	Test Accuracy
Base model: Adam optimizer, learning rate = 0.001, 50% dropout on FC layers, 30% dropout on Conv layers	91.8%	89.9%
Adadelta optimizer, learning rate = 1.0, 50% dropout on FC layers, 30% dropout on Conv layers	88.6%	87.9%
Base model + 40% dropout on FC layers, 20% on Conv layers, batch size of 64	93.8%	91.2%
Base model + 50% dropout on all layers, batch size of 64 **(chosen model)**	94.5%	92.1%

TSR Model Description	Test Accuracy	Test Recall	Test Precision	Test F1 score
Optimized Model	92.1%	92.94%	93.21%	92.91%
Optimized Model w/o FC layer 1	91.1%	91.79%	92.37%	91.71%
Optimized Model w/o FC layer 2	91.0%	91.69%	92.38%	91.72%

AdvGAN

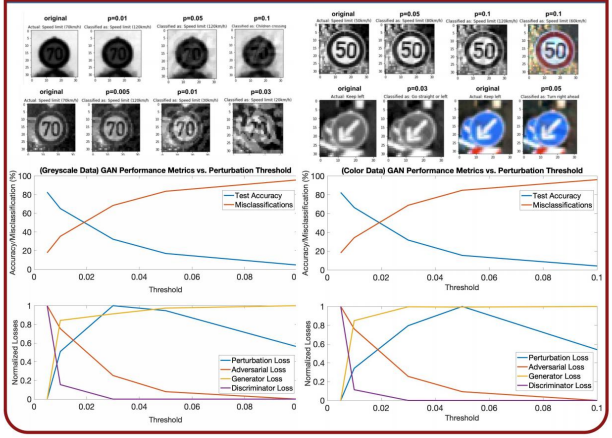


Discriminator: Convolutional "PatchGAN" Classifier

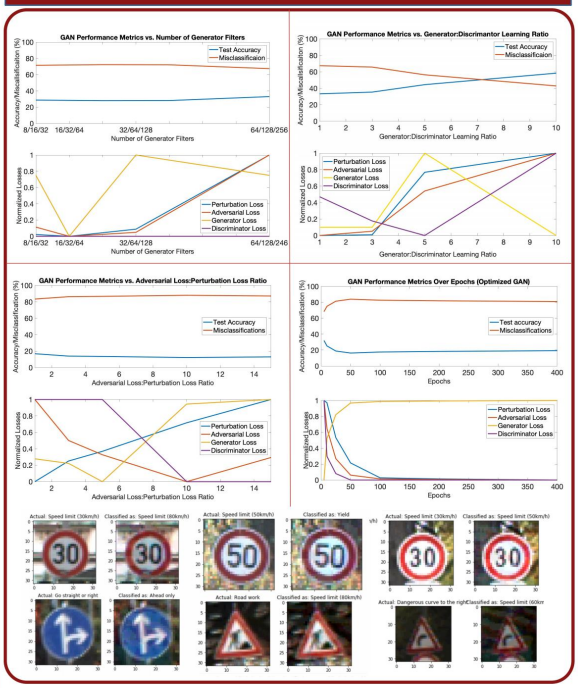
Generator: 3-layer Encoder-Decoder with a ResNet Bottleneck

Both generator and discriminator use modules of the form convolution-BatchNorm-ReLu.

Perturbation Analysis (Color & Greyscale)



Model Optimization



Conclusions and Future Work

By developing an optimized GAN (with perturbation threshold of 0.05, 10:1 adversarial:perturbation weighted-loss ratio, 16/32/64 filters in the generator, and a 1:1 generator:discriminator learning ratio), we generated inconspicuous adversarial perturbations to traffic sign images and fed the perturbed images to a TSR classifier, reducing the classifier test accuracy from 92.1% to 17.4% and resulting in a significantly high misclassification percentage of 82.9%.

To mitigate the effect of these attacks, autonomous vehicle software should include perturbed images in the training set for their traffic sign recognition classifier (future work suggestion). Another solution is *defensive distillation*.