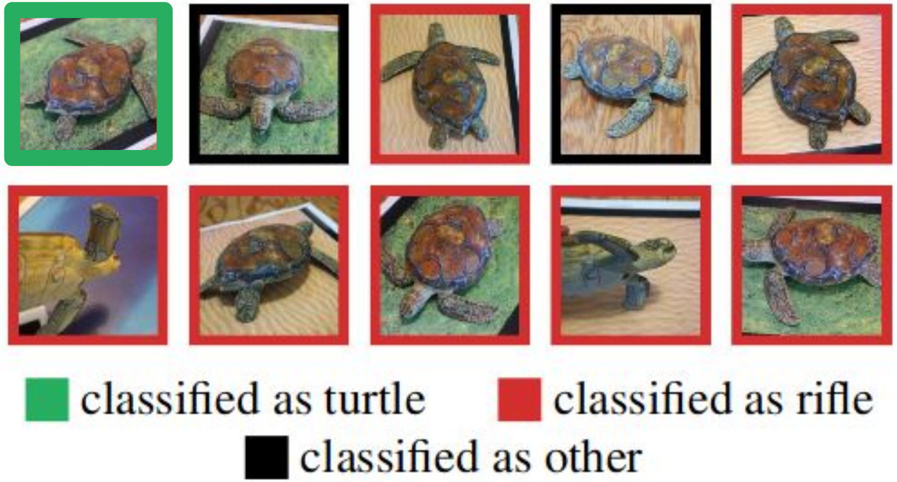
Generating Adversarial Examples with SAT Solvers

Owen Smith

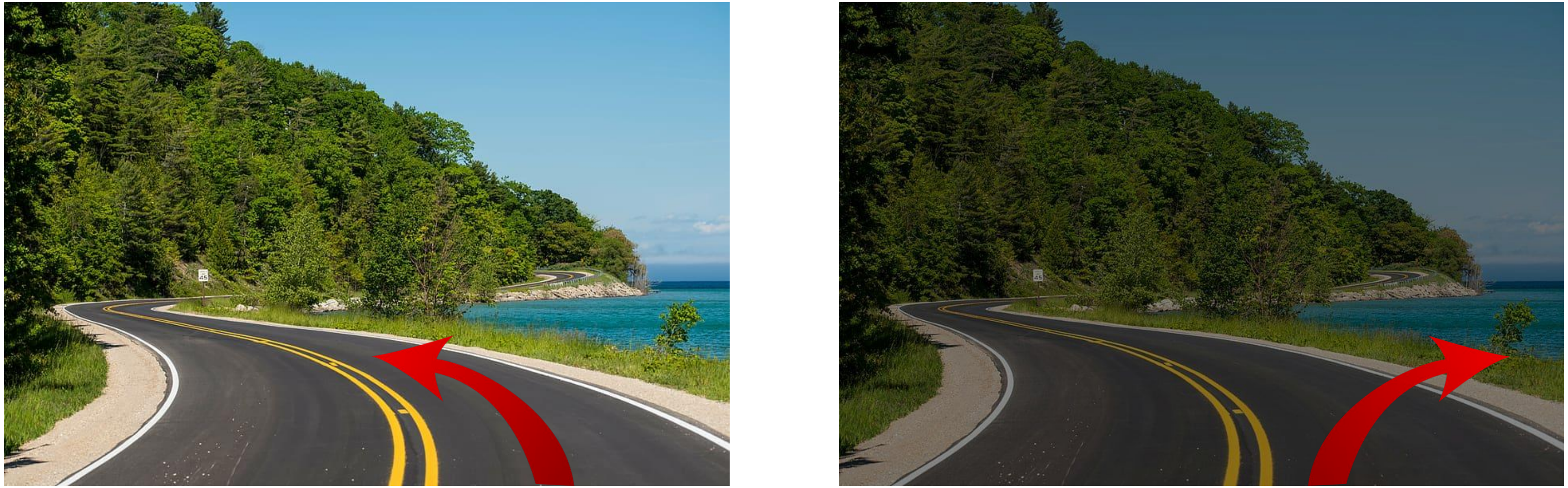
COMP 597 Fall 2021

Classifiers are....not perfect



[Athalye et al 2018]

Motivation & Applications

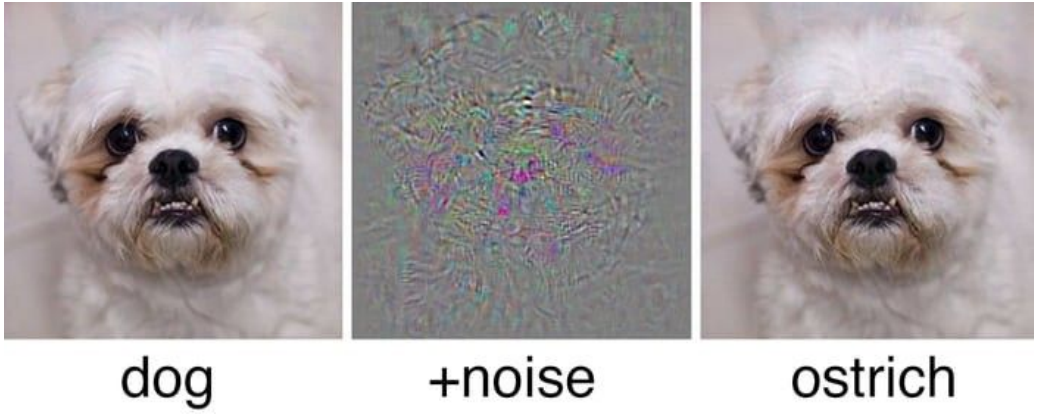


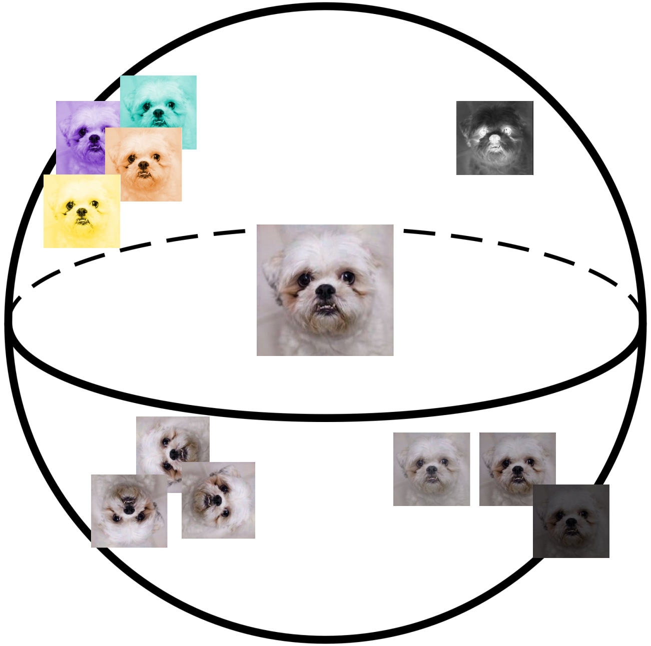
[DeepXplore 2017]

What are adversarial examples?

- Adversarial examples are specialised inputs created with the purpose of confusing a neural network, resulting in a misclassification of the input.

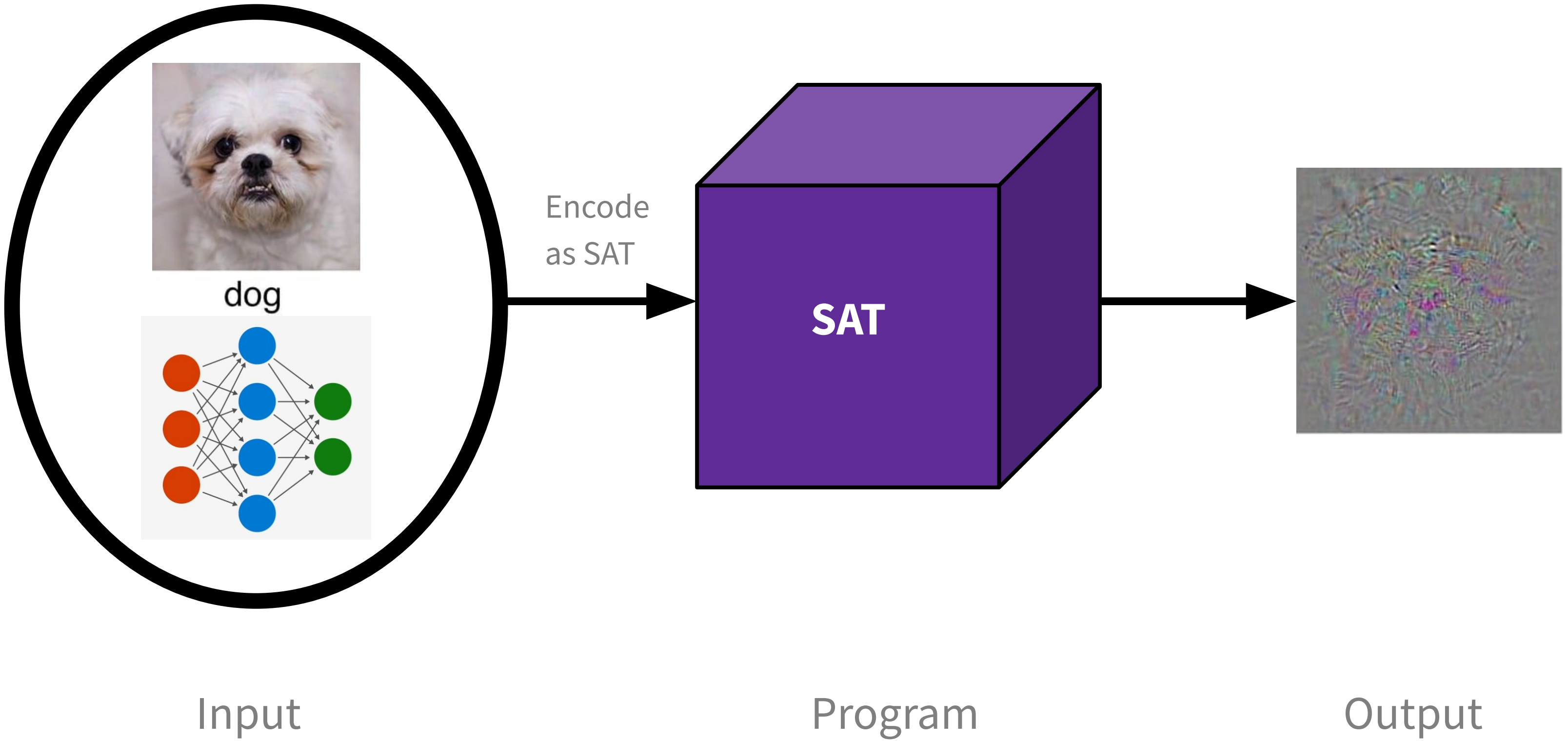
- We can include these adversarial examples back into the train set to help improve model performance.





Visualizing The Problem

Workflow finding adversarial examples



dog

Encode
as SAT

SAT

Input

Program

Output

Why use a SAT Solver?

- A neural network can easily be translated into SAT
    - We only encode the forward pass
    - Process is declarative, don’t need to know anything about the loss function
- We can leverage the solver’s internal algorithm to analyze the interaction between neurons and solve the constraints at a much quicker rate than a naive search approach.
- Extendable across different models. All we need are the weights and biases.

Starting Simple

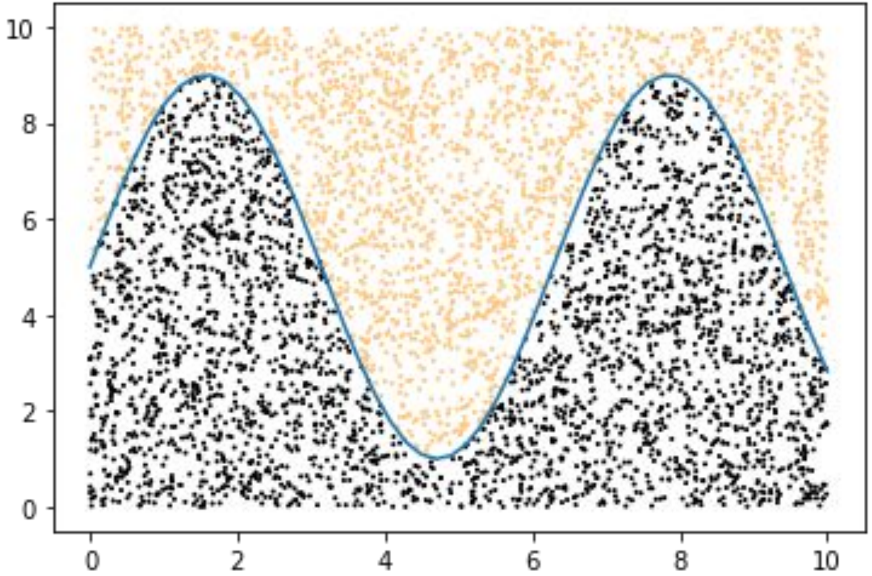
Binary Classifier w/ 2 Features

Training Data

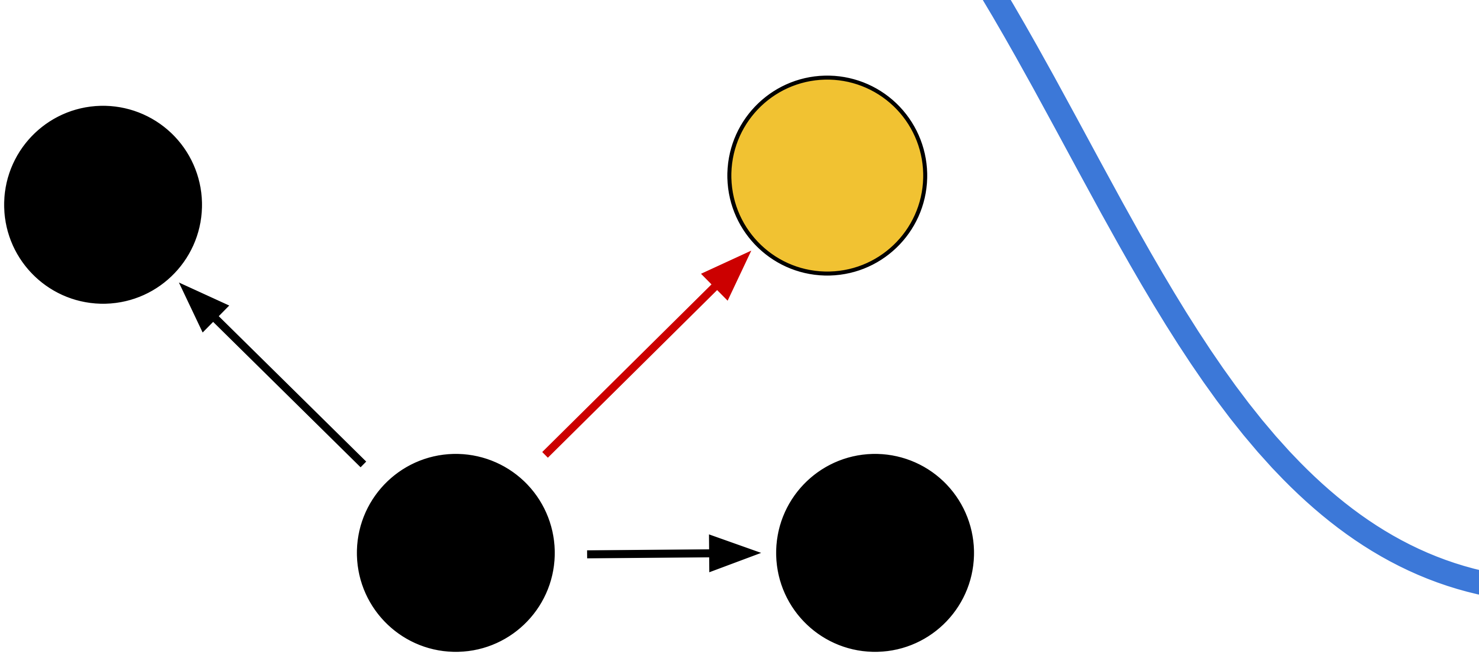
■ Class 0

🟨 Class 1

🟦 Ground Truth



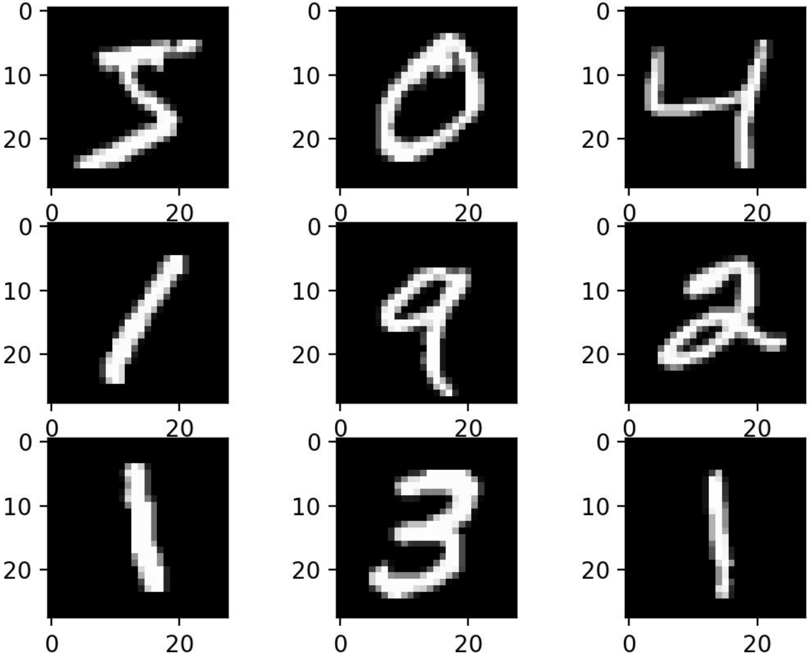
Finding an adversarial example



Demo pt. 1

Finding a mask for images

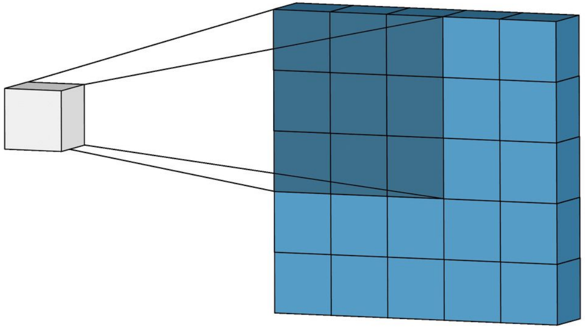
MNIST



Demo pt. 2

Future Work

- Extend to coloured images
- Explore more complex layers (convolutional)
- Explore Quantized Neural Networks to improve scalability
- Encode purely as SAT instead of SMT lowering the problem onto SAT



References

https://github.com/owenps/Adversarial_Generator

Athalye, Anish. “**Synthesizing Robust Adversarial Examples**.” Arxiv.org, 7 June 2018, https://arxiv.org/pdf/1707.07397.pdf.

Pei, Kexin. “**DeepXplore: Automated Whitebox Testing of Deep Learning Systems.**” Arxiv.org, 24 Sept. 2017, https://arxiv.org/pdf/1705.06640.pdf.