Generating Adversarial Attacks for Sparse Neural Networks

Jack H. Chen  
*Portland State University*

Walt Woods  
*Portland State University*

Citation Details
https://pdxscholar.library.pdx.edu/mcecs_mentoring/31

This Poster is brought to you for free and open access. It has been accepted for inclusion in Undergraduate Research & Mentoring Program by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.
Generating Adversarial Attacks for Sparse Neural Networks

Jack Chen, Walt Woods
Advisor: Christof Teuscher

Image Recognition:
- Relies on Neural Networks
- Classifies images with state-of-the-art accuracy

Adversarial Attacks:
- Neural Networks are highly susceptible to imperceptible changes to its input that changes prediction called adversarial attacks.
- Has huge security implications, therefore it is beneficial to make attacks more perceivable.
- Attack perceivability measured with Mean Square Error (MSE).
- Attack effectiveness to change classification is measured with loss.

Problem

To test our hypothesis, we need to generate adversarial attacks that are both effective and computed in reasonable amount of time. We began by creating white box and black box attack generators.

Attack Generation

Figure 1: Adversarial Attack changing image classification from Sports Car to Fur Coat

Figure 2: Representation of a white box attack. Internal gradients of a neural network is known to a generator.

Figure 3: Representation of a black box attack. Only the inputs and outputs of network are known to the generator.

Figure 4: Number of iterations before successful classification change on white box attack generator.

Figure 5: White box attacks on CIFAR-10 dataset with class accuracies before (leftmost) and after (rightmost) an attack.

Figure 6: CIFAR-10 images of before and after black box attack and resulting MSE of the attack

Results and Next Steps

Black Box Generation:
- Generally the internals of a neural network are unknown.
- We used a genetic algorithm where attacks were evaluated as a logarithmic function of loss, max loss being 0.5, plus MSE.

Next Steps:
- We must find sparse architectures that models the relevant parts of images for neural network input and test these architectures with our black box generator.
- We will also examine our generator with state-of-the-art classifiers and examine the attack’s MSE.

Hypothesis

Mean Square Error (MSE):
- In an attack, MSE is the mean of the magnitude of changed pixels
- Inversely related to accuracy. As loss of a class increases, accuracy of that class decreases.
- Generating adversarial attacks must minimize both loss and MSE.

Loss:
- Generally the internals of a neural network are unknown.
- We used a genetic algorithm where attacks were evaluated as a logarithmic function of loss, max loss being 0.5, plus MSE.

High MSE

Low MSE

Figure 2: Example of relatively high (top) and low MSE (bottom)

Sparsity:
- Sparsity can provide benefits, such as less power usage, with little difference in accuracy [Woods & Teuscher, 2017].
- We predict that there exists a sparse architecture that takes only the relevant parts of an image as inputs to a neural network. Taking only relevant parts of an image reduces the space an adversarial attack can affect and therefore we predict the MSE of that attack will increase.