Neural Network Training Effects On Adversarial Imagery In Wildlife Vision

Wentong Zhang[1], Travis Desell[2]

[1]Faculty of Arts and Sciences, Harvard University

[2]Department of Computer Science, University of North Dakota

Introduction

Neural networks have recently been shown to have a critical flaw exposed in adversarial examples, images that are nearly imperceptibly different from a training image that the network misclassifies with high confidence[1]. We investigate the difference in robustness to such adversarial examples between networks trained via back propagation and particle swarm optimization. **Initial testing on aerial wildlife imaging indicates that training using particle swarm optimization creates networks less robust to adversarial images.**

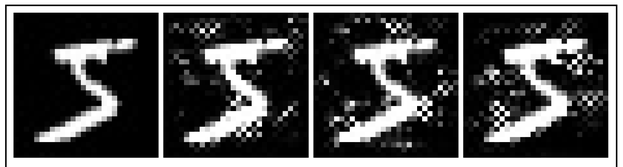

Figure 1. A digit from the MNIST database and 3 adversarial images generated from the digit.

Methodology

The data set used consists of a small set of 168 images, half of which contained bison. The neural network structure consists of **a gray-scale filter, two convolutional layers, and a fully connected layer.** Each convolutional layer contains three filters for the images, and each of the filters are connected to each filter in the next layer. The network **uses a softmax layer to classify the images.**

Two neural networks were trained using back propagation and particle swarm respectively. Each achieved perfect accuracy on the small test set of 40 images. Adversarial images were then **generated by back propagating directly back into the image** without changing the weights of the network. Finally, the average distortion between original and adversarial images was calculated.

Data

To calculate the **distortion** values, we compare images pixel by pixel with the following formula:

$$d = \sqrt{\frac{\sum (x_i' - x_i)^2}{n}}$$

Back Propagation

Average Distortion Values

Step Size	99% Conf.	99.9% Conf.	99.99% Conf.	1000 Iter.
0.0005	0.013938	0.0174745	0.0297978	0.0149107
0.001	0.0139449	0.017478	0.029799	0.0159777
0.01	0.0142452	0.0176715	0.0299041	0.0209069
0.1	0.0618472	0.0618797	0.0622544	0.0895939
1	0.404764	0.404764	0.404764	0.475895

Corresponding Pictures

Step Size	99% Conf.	99.9% Conf.	99.99% Conf.	1000 Iter.	99% Conf.	99.9% Conf.	99.99% Conf.	1000 Iter.
0.0005								
0.001								
0.01								
0.1								
1								

Particle Swarm

Average Distortion Values

Step Size	99% Conf.	99.9% Conf.	99.99% Conf.	1000 Iter.
0.0005	0.00420	0.00484099	0.00557814	0.0052877
0.001	0.0052173	0.00554686	0.00608713	0.0060961
0.01	0.0367885	0.036793	0.0368015	0.0372764
0.1	0.307981	0.308002	0.308001	0.287488
1	0.735891	0.735707	0.735673	0.503972

Corresponding Pictures

Step Size	99% Conf.	99.9% Conf.	99.99% Conf.	1000 Iter.	99% Conf.	99.9% Conf.	99.99% Conf.	1000 Iter.
0.0005								
0.001								
0.01								
0.1								
1								

Analysis

Quantitatively, the adversarial images generated on the network trained by back propagation had much lower average distortion values when using a step size of 0.001 or 0.0005.

On the other hand, when using a relatively high step size of 0.1, the adversarial images from the network trained by particle swarm looked much less realistic than those from the network trained by back propagation.

Overall, the **particle swarm network had much higher sensitivity to image modification** than that of the back propagation network, which indicates **reduced robustness to adversarial examples.** This accounts for both the smaller distortion values at small step sizes and large distortion values at large step sizes.

Future Work

While applicable to wildlife vision, the generation of adversarial images has more imminent applications to the field of computer vision and pattern recognition. We hope to **run similar experiments on the well-known standardized data sets**, including the MNIST and CIFAR-10. Usage of **other global optimization techniques**, such as other evolutionary algorithms, would also be interesting to consider in the future.

Acknowledgements

This research was funded by the **U.S. National Science Foundation** (NSF Award #1359224) with support from the **U.S. Department of Defense**. Imagery was provided by University of North Dakota Department of Biology, Susan Ellis-Felege, Robert Newman, Michael Corcoran, and Christopher Felege. Additional thanks to Jeremy Straub for his helpfulness and support in the research process.

References

[1]Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. Intriguing properties of neural networks. Proceedings of the 2014 International Conference on Learning Representations.