UNIVERSITY OF NEBRASKA AT OMAHA
COLLEGE OF INFORMATION SCIENCE & TECHNOLOGY

MANTIC LAB

Towards Robust Classification in Adversarial Learning using Bayesian Games

Anna Buhman, Masters Student
Computer Science Department, University of Nebraska at Omaha
Advisor: Dr. Raj Dasgupta

Abstract

A well-trained neural network is very accurate when classifying data into different categories. However, a malicious adversary can fool a neural network through tiny changes to the data, called perturbations, that would not even be detectable to a human. This makes neural networks vulnerable to influence by an attacker. Generative Adversarial Networks (GANs) have been developed as one possible solution to this problem [1]. A GAN consists of two neural networks, a generator and a discriminator. The discriminator tries to learn how to classify data into categories. The generator stands in for the attacker and tries to discover the best way to cause the discriminator to make wrong classifications through perturbing the input. Our work improves on this method through the application of Bayesian games to model multiple generators and discriminators rather than one of each. Through training against multiple types of input perturbation, the discriminators will improve their classification of adversarial samples.

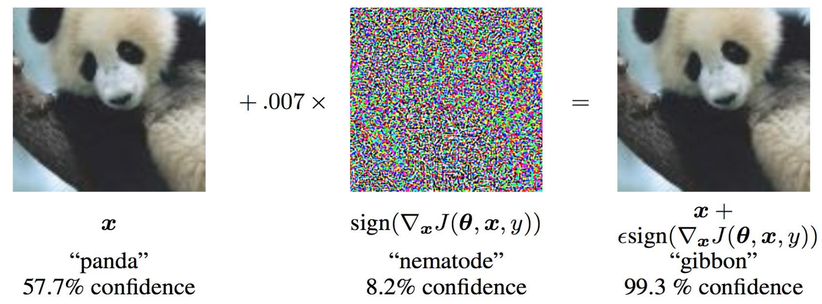

Figure 1. An example of adversarial perturbation of an image. An image which was correctly classified as a panda can be perturbed by distorting it slightly with noise. After the perturbation, it still appears to human eyes to be a panda, but the neural network incorrectly classifies the image as a gibbon. [2]

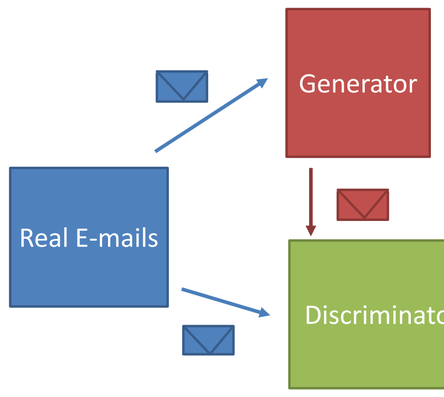

Figure 2. A generative adversarial network for spam email. The generator receives both real emails and those that are perturbed by the generator. The discriminator must correctly classify which emails are spam and which are non-spam

Objectives

- Defend against an adversary who is trying to fool a neural network
- Make existing techniques more flexible and robust
- Effective perturbation of text email for use by the generator to fool the discriminator
- Accurate classification of text email, both perturbed and unperturbed, into spam vs non-spam

Existing Approaches/Limitations

- Generative Adversarial Networks improve the discriminator's ability to defend against an attacker, with the limitation that they can only defend against an attacker similar to the generator they trained against
- Most Generative Adversarial Networks work with images, while ours works with text. A limitation of perturbing text is that while an image can be slightly modified and still look like an image, slightly modified text may become nonsensical [3]
- Bayesian games have been applied to adversarial learning [4]
- Most research on text-based GANs focuses on generating meaningful text, not on defending against adversarial text.

Proposed Approach

- There will be multiple generators and discriminators as opposed to one of each
- The discriminators will become more flexible in their defense by defending against different generators who will use different perturbation methods to fool the discriminator
- Bayesian games will be used to model this interaction and find an optimal strategy for improving the discriminator's classification.
- Use Word2Vec algorithm to convert words to numerical vectors so they can be perturbed without using any nonsense words [5]

Research Plan

- Implementation
  - Existing libraries for Word2Vec and neural networks
  - HCC's Crane supercomputer
- Testing
  - Classify spam email vs non spam
  - Discriminators will be matched against multiple generators
- Data to collect
  - Accuracy of classification
  - Compare accuracy with other text GAN methods

References

[1] I. Goodfellow *et al.*, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 2672–2680.
[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," presented at the ICLR 2015, San Diego, USA, 2014.
[3] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient.," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 2852–2858.
[4] Michael Großhans, Christoph Sawade, Michael Br¨uckner, and Tobias Scheffer. Bayesian Games for Adversarial Regression Problems. In Proceedings of the 30th International Conference on International Conference on Machine Learning – Volume 28, ICML'13, pages III–55–III–63, Atlanta, GA, USA, 2013. JMLR.org.
[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 3111–3119.