Stochastic Gradient Switching

For defense against

White Box Adversarial Evasion Attacks

White Paper

By Eric Muccino

Abstract

Existing artificial neural network frameworks are vulnerable to a variety of adversarial attacks. Attackers employ white box adversarial evasion attacks by exploiting model gradients with gradient ascent methods to engineer data samples that the model will misclassify to the adversary's specification. Stochastic Gradient Switching is a novel defense approach, where each layer in a neural network is designed to be an ensemble of unique layers, all fully connected to the previous layer ensemble. During inference, one layer is randomly selected from each ensemble to be used for forward propagation, effectively selecting one of many unique sub-networks upon each inference call. Stochastic gradient switching removes an attacker's ability to deterministically track model gradients, subduing evasion attack efforts that require gradient ascent optimization.

Contents

1. Introduction

As deep learning technologies power increasingly more services, associated security risks become more critical to address. Adversarial machine learning is a branch of machine learning that exploits the mathematics underlying deep learning systems in order to evade, explore, and/or poison machine learning models [1,2]. Evasion attacks are the most common adversarial attack method due to their ease of implementation and potential for being highly disruptive. During an evasion attack, the adversary tries to evade a fully trained model by carefully engineering samples to be misclassified by the model. This attack does not assume any influence over the training data. Evasion attacks have been demonstrated in the context of autonomous vehicles where the adversary manipulates traffic signs to confuse the learning model [3]. Research suggests that deep neural networks are susceptible to adversarial based evasion attacks due to their high degree of nonlinearity as well as insufficient model averaging and regularization [4]. In this paper we propose Stochastic Gradient Switching, a technique for defending against adversarial evasion attacks.

2. Problem Statement

Adversarial attacks can be classified as being either black-box or white-box attacks. During a white-box attack, an attacker has full access to the model, including the architecture, weights, and training algorithm. During a black-box attack, the attacker only has the ability to use the model as an oracle, observing model outputs by querying the model with inputs. If given the opportunity, white-box attacks are easier to perform since the attacker has complete information about the model. For this reason, defense techniques against white-box attacks also work to defend against black-box attacks. During a white-box evasion attack, an adversary will select an input instance that he or she wants to be misclassified. The attacker will perform a gradient ascent optimization algorithm to tune input features of a sample through the use of a loss function that maximizes output of a desired class while minimizing the total change to the input features [4,5]. In the context of computer vision, this process results in the creation of images that are visibly indistinguishable from original samples but are capable of fooling the model through careful exploitation of narrow gaps in decision boundaries. Our proposed framework removes an adversary's ability to perform evasion attacks.

3. Prior Art

Recent research has established various defenses against adversarial attacks. However, each defense technique has considerable drawbacks and/or limitations. Some of the most prolific defense strategies include; adversarial training, gradient hiding, and feature squeezing.

Adversarial training is the process of generating adversarial samples for a model then including those samples in the training data set to develop a second, more robust model [4,6,7,8]. While this can help to an extent, the problem is that for highly non-linear tasks, such as computer vision, new adversarial samples will inherently exist for the second model that did not exist for the first [9,10,11]. These gaps are still left vulnerable for an attacker to exploit.

Gradient hiding is the method of using a non-gradient based machine learning framework such as random forest or k-nearest neighbors [11]. The problem with this is that it renders deep neural networks unusable. Many problems such as computer vision and natural language processing are best handled by one of the many architecture designs that deep neural networks have to offer. Additionally, non-gradient based machine learning models can be approximated by a surrogate deep neural network model. Research shows that adversarial examples found on the surrogate model are often adversarial on the original non-gradient based model as well [10].

Feature squeezing is a technique that involves reducing the complexity of an input feature space [12,13]. This is done either by binning values into a lower resolution or by converting the input feature space into a lower dimensional latent space. The largest drawback of feature squeezing is that it often comes at the expense of degraded model performance and it doesn't remove all adversarial examples.

4. Approach

Architecture

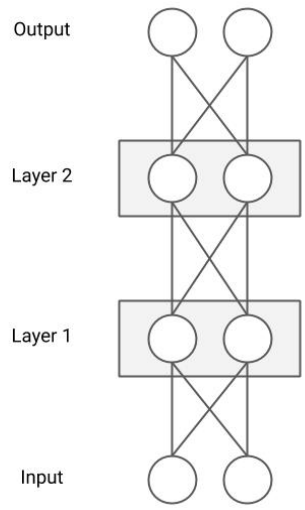

Figure 1

A conventional neural network consists of a series of layers, each layer containing perceptrons that make computations based upon learned weights associated with previous layers (figure 1). The weights are trained via a backpropagation gradient descent algorithm that minimizes a loss function over a training data set. When an adversary has complete knowledge of the weights and connections of the network, they can exploit the network gradients to craft adversarial samples that produce desired outputs.

To address the challenges and shortcomings identified in the previous sections, Stochastic Gradient Switching removes the determinism of a neural network's gradients in order to eliminate an attackers ability to exploit

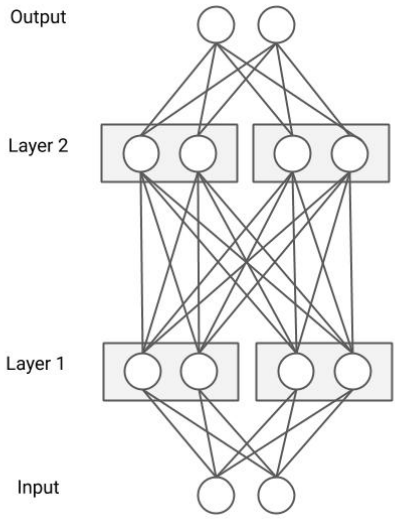

Figure 2

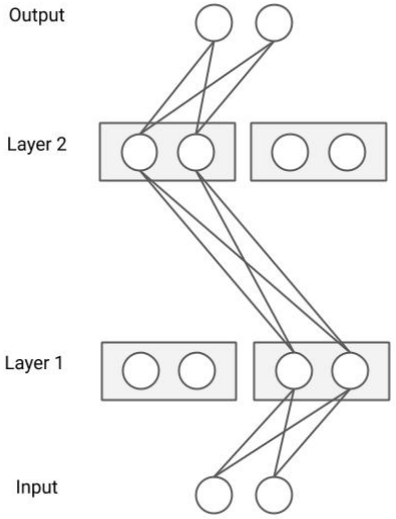

Figure 3

them. Our approach is to replace each hidden layer of a conventional neural network with an ensemble of layers, each layer being fully connected to the previous ensemble of layers. An example is shown in figure 2, where each hidden layer is replaced by an ensemble of 2 layers that run in parallel. Upon inference, a single layer is randomly selected from each layer ensemble as shown in figure 3. Forward propagation is then computed as in a conventional neural network and an output is returned. In the illustrated example, there are 2 hidden layer ensembles, each containing 2 individual layers. This effectively creates 4 sub-networks, each with different gradients and an equal chance to be selected upon inference. In general, a network with n layer ensembles and m layers in each ensemble will have m^n sub-networks with unique gradients. This basic architecture can be extended to more complex neural network schemas such as convolutional neural networks or recurrent neural networks simply by expanding each hidden layer across an additional dimension to create ensemble layers.

Model Training

For smaller models, Neural Networks with Stochastic Gradient Switching can be trained in parallel as a collection of sub-networks with shared weights. For larger models and limited computational resources, a drop-out style training algorithm can be implemented to train one sub-network at a time. For Stochastic Gradient Switching to be an effective defense strategy against white-box evasion attacks, each sub-network must maintain a unique set of gradients from the other sub-networks. This is achieved by applying a regularization term to the loss function that measures the vector distances between each layer within an ensemble. Minimizing this loss term will encourage the network to find solutions that keep

sub-networks nearly orthogonal to each other. The weight of the vector distance regularization term within the loss function must be tuned as a hyperparameter to balance sub-network orthogonality with high training accuracy. Orthogonal layer spaces result in sub-networks that transform inputs through unique feature space bases. Having unique feature spaces across sub-networks decreases overlap in adversarial example feature state space. For this reason, an adversarial example trained over one sub-network is unlikely to be an adversarial example in a different sub-network.

5. Research

Mindboard's research goals are to investigate and quantify the effectiveness of Stochastic Gradient Switching in defending against adversarial evasion attacks. This investigation will extend to a range of model architectures including Feedforward Neural Networks, Convolutional Neural Networks, and Recurrent Neural Networks. Mindboard seeks to explore the companionship of Stochastic Gradient Switching with other adversarial defense techniques and consider usage for defense against other adversarial attack types such as exploratory and poisoning attacks. We will pursue the development of Stochastic Gradient Switching libraries to be compatible with machine learning libraries such as TensorFlow, Keras, and/or PyTorch.

6. Summary

Stochastic Gradient Switching is a novel machine learning technique with the potential to have a large impact on how we protect ourselves against adversarial attacks. With machine learning services beginning to occupy critical spaces such as military operations, law enforcement, and medical practices, secure machine learning pipelines are extremely important. Mindboard seeks to contribute to the adversarial machine learning field through the research and development of robust neural networks with Stochastic Gradient Switching.

7. References

[1] Battista Biggio, Giorgio Fumera, and Fabio Roli. 2014. Security Evaluation of Pattern Classifiers under Attack. IEEE Trans. Knowl. Data Eng. 26, 4 (2014), 984–996. https://doi.org/10.1109/TKDE.2013.57

[2] Battista Biggio, Giorgio Fumera, and Fabio Roli. 2014. Security evaluation of pattern classifiers under attack. IEEE transactions on knowledge and data engineering 26, 4 (2014), 984–'–996.

[3] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2016. Adversarial Machine Learning at Scale. CoRR abs/1611.01236 (2016). arXiv:1611.01236 http://arxiv.org/abs/1611.01236

[4] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. CoRR abs/1312.6199 (2013). http://arxiv.org/abs/1312.6199

[5] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016. 582–597. https://doi.org/10.1109/SP.2016.41

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. CoRR abs/1412.6572 (2014). http://arxiv.org/abs/1412.6572

[7] Chunchuan Lyu, Kaizhu Huang, and Hai-Ning Liang. 2015. A Unified Gradient Regularization Family for Adversarial Examples. In 2015 IEEE International Conference on Data Mining, ICDM 2015, Atlantic City, NJ, USA, November 14-17, 2015. 301–309. https://doi.org/10.1109/ICDM.2015.84

[8] Uri Shaham, Yutaro Yamada, and Sahand Negahban. 2015. Understanding Adversarial Training: Increasing Local Stability of Neural Nets through Robust Optimization. CoRR abs/1511.05432 (2015). http://arxiv.org/abs/1511.05432

[9] Nina Narodytska and Shiva Prasad Kasiviswanathan. 2017. Simple Black-Box Adversarial Attacks on Deep Neural Networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, Honolulu, HI, USA, July 21-26, 2017. 1310–1318. https://doi.org/10.1109/CVPRW.2017.172

[10] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2017. Practical Black-Box Attacks against Machine Learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017. 506–519. https://doi.org/10.1145/3052973.3053009

[11] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick D. McDaniel. 2017. Ensemble Adversarial Training: Attacks and Defenses. CoRR abs/1705.07204 (2017). arXiv:1705.07204 http://arxiv.org/abs/1705.07204

[12] Weilin Xu, David Evans, and Yanjun Qi. 2017. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. CoRR abs/1704.01155 (2017). arXiv:1704.01155 http://arxiv.org/abs/1704.01155

[13] Weilin Xu, David Evans, and Yanjun Qi. 2017. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. CoRR abs/1704.01155 (2017). arXiv:1704.01155 http://arxiv.org/abs/1704.01155