Understanding Generalization in Neural Networks for Robustness against Adversarial Vulnerabilities

Subhajit Chaudhury[*]
The University of Tokyo
subhajit@hal.t.u-tokyo.ac.jp

Abstract

Neural networks have contributed to tremendous progress in the domains of computer vision, speech processing, and other real-world applications. However, recent studies have shown that these state-of-the-art models can be easily compromised by adding small imperceptible perturbations. My thesis summary frames the problem of adversarial robustness as an equivalent problem of learning suitable features that leads to good generalization in neural networks. This is motivated from learning in humans which is not trivially fooled by such perturbations due to robust feature learning which shows good out-of-sample generalization.

Neural networks have contributed to tremendous progress in the domains of computer vision, speech processing, and other real-world applications. However, recent studies have shown that these state-of-the-art models can be easily compromised by adding small imperceptible perturbations. Such malicious artifacts can be added by an attacker both in the training data (poisoning attacks) and test data for a trained model (adversarial attacks), to cause significant loss of performance during inference. My thesis summary frames the problem of adversarial robustness as an equivalent problem of learning suitable features that leads to good generalization in neural networks. This is motivated by learning in humans which is not trivially fooled by such perturbations due to robust feature learning which shows good out-of-sample generalization. My thesis is planned to be divided into three parts. Firstly, I studied methods to find explainable single-pixel perturbations to the training data (poisons) that compromise the generalization abilities of neural networks. In the second phase, I plan to study performance to adversarial vulnerabilities by learning robust non-linear feature transformations. In the third part, I plan to study the effect of different optimization techniques towards generalization in neural networks in the presence of adversarial perturbations.

1) Interpretable Single-pixel Poisons: Data poisoning involves injecting small perturbations to training samples by an attacker to subvert the model performance on clean test data. Biggio et al. (Biggio, Nelson, and Laskov 2012) first introduced it in the context of Support Vector Machines (SVM) for binary classification problems. Recently, there have been some works in the field of neural networks (Steinhardt, Koh, and Liang 2017) as well. Koh et al. (Koh and Liang 2017) used influence functions to synthesize adversarial training examples that can flip the predicted labels of a set of testing images. Shafahi et al. (Shafahi et al. 2018) used a forward-backward-splitting iterative procedure (Goldstein, Studer, and Baraniuk 2014) to create targeted data poisoning attacks that performed better than previous methods. GenAttack (Alzantot et al. 2018) proposed a gradient-free adversarial attacks for test time perturbation optimization. Considering previous single-pixel poisoning attacks, (Jacobsen et al. 2018) studied the effect of single-pixel perturbations on MNIST training images on test performance. They showed that adding one pixel to training images that encodes the class label, and then testing on the clean test set can yield a high generalization gap. Tanay et al. (Tanay, Andrews, and Griffin 2018) showed that neural network models can be made almost arbitrarily sensitive to a single-pixel while maintaining identical test performance between models.

Table 1: Showing testing error (in %) for various methods. "Clean" is non-poisoned data. Training errors are close to 0%.

Dataset	MNIST	Fashion-MNIST
Clean	0.8 ± 0.0	8.3 ± 0.3
Baseline (random)	56.8 ± 0.6	52.0 ± 1.7
Jacobsen et al. (Jacobsen et al. 2018)	43.0 ± 13.9	49.6 ± 2.6
Proposed	$\mathbf{88.6 \pm 1.2}$	$\mathbf{74.9 \pm 0.8}$

However, there is a limited study on explaining why certain poisons are more effective in fooling the model that others. In this work, we propose an explainable gradient-free data poisoning approach that learns single-pixel perturbations on the training images that forces the neural network to focus on non-salient spatial locations for the classification task. In this work, we assume the attacker has access to the clean training images and labels, which can be used to learn a clean classification model ($\mathcal{M}_c$). GradCAM (Selvaraju et al. 2017) distribution corresponding to the true label is obtained from $\mathcal{M}_c$. The region with high value suggests that the neural network focuses on those regions to make its classification decision. Our goal is to inject poisons on the non-salient image loca-

[*] This doctoral thesis is conducted under the supervision of Prof. Toshihiko Yamasaki, The University of Tokyo.

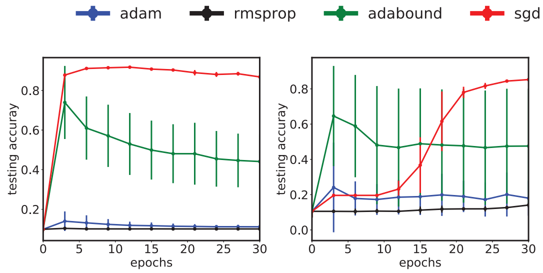
Table 2: Showing testing error (in %) for various non-NN learning methods on MNIST dataset

Method	Lin-SVM	Random Forest	RBF-SVM
Clean	8.34	5.50	1.65
Baseline	59.77	12.55	2.30
Jacobsen (Jacobsen et al. 2018)	75.16	15.79	15.79
Proposed	**90.3**	**14.64**	**2.06**

tions to divert the neural networks GradCAM distribution to the non-discriminative image features. To this end, we use the complementary Region of Interest (ROI) to sample class-wise single-pixel locations and perturbation intensities. We use a gradient-free optimization technique (CMA-ES) to search for the best performing perturbations based on a fitness score that encourages a higher generalization gap between clean and poisoned images. Table 1 shows experimental evaluation on MNIST and Fashion-MNIST datasets. Our proposed attack strategy outperforms previous single pixel-based poisoning methods (Jacobsen et al. 2018) and the baseline of random perturbations.

2) Can Robust Non-linear Transformations Prevent Adversarial Vulnerabilities?: We performed experiments to evaluate how the above single-pixel poison attacks would perform on other traditional learning methods like LinearSVM, Random Forests, and RBF-SVM as shown in Table 2. We found that Linear SVM was successfully attacked by our method, while Random Forests and RBF-SVM is resilient to such attacks. Especially RBF-SVM can defend against such attacks because it uses non-linear feature transformation based on the l_2 distance. The difference between the original and poisoned input is minimal in l_2 distance perspective due to single-pixel perturbations. Thus the decision boundary learned in the transformed feature space can also generalize to the clean test images. In the future, I plan to examine the effect on adversarial robustness due to non-linear feature transformations more generally, because l_2 distance between attacked and clean samples are typically very less.

3) Does Adaptivity in Optimization Overfit Easily?: Wilson et al. (Wilson et al. 2017) showed that adaptive methods are affected by spurious features that do not contribute to out-of-sample generalization by crafting a smart artificial linear regression example. By examining the effect of common optimization strategies on our single-pixel poisons, we wish to study if a certain optimization algorithm is more liable to memorizing small perturbations while ignoring other salient statistical patterns in the training data. To this end, we trained CNN models on single-pixel perturbed data using ADAM (Kingma and Ba 2014), SGD, RMSProp (Tieleman and Hinton 2017), and Adabound (Luo et al. 2019) optimization as shown in Figure 1. ADAM and RMSProp show low testing accuracy for all cases while vanilla SGD is surprisingly resilient to such perturbations showing better out-of-sample performance consistently for all the datasets. Adabound uses strategies from both SGD and Adam, thus showing intermediate performance. In the future, I plan to study why SGD based methods are more resilient to such poisoning attacks and if this property can be used in adversarial robustness in general for both evasive and poisoning attacks.



(a) MNIST (b) SVHN

Figure 1: Testing accuracy under single pixel perturbation shows SGD consistently performs better than adaptive optimization techniques.

References

Alzantot, M.; Sharma, Y.; Chakraborty, S.; and Srivastava, M. 2018. Genattack: Practical black-box attacks with gradient-free optimization. *arXiv preprint arXiv:1805.11090*.

Biggio, B.; Nelson, B.; and Laskov, P. 2012. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*.

Goldstein, T.; Studer, C.; and Baraniuk, R. 2014. A field guide to forward-backward splitting with a fasta implementation. *arXiv preprint arXiv:1411.3406*.

Jacobsen, J.-H.; Behrmann, J.; Zemel, R.; and Bethge, M. 2018. Excessive invariance causes adversarial vulnerability. *arXiv preprint arXiv:1811.00401*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koh, P. W., and Liang, P. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1885–1894. JMLR. org.

Luo, L.; Xiong, Y.; Liu, Y.; and Sun, X. 2019. Adaptive gradient methods with dynamic bound of learning rate. *arXiv preprint arXiv:1902.09843*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Shafahi, A.; Huang, W. R.; Najibi, M.; Suciu, O.; Studer, C.; Dumitras, T.; and Goldstein, T. 2018. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, 6103–6113.

Steinhardt, J.; Koh, P. W. W.; and Liang, P. S. 2017. Certified defenses for data poisoning attacks. In *Advances in neural information processing systems*, 3517–3529.

Tanay, T.; Andrews, J. T.; and Griffin, L. D. 2018. Built-in vulnerabilities to imperceptible adversarial perturbations. *arXiv preprint arXiv:1806.07409*.

Tieleman, T., and Hinton, G. 2017. Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *Technical Report.*

Wilson, A. C.; Roelofs, R.; Stern, M.; Srebro, N.; and Recht, B. 2017. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, 4148–4158.