

## 論文の内容の要旨

論文題目 Understanding Generalization in Neural Networks for  
Robustness against Adversarial Attacks  
(敵対的攻撃に対するニューラルネットワークの汎化性能の理解)

氏 名 チャウダリー シュボジト Subhajt Chaudhury

This thesis outlines various methods for adversarial attacks in training images, explain various properties of DNNs using such attacks and proposes various techniques for adversarial defense. In Chapter 1, we outline the summary of this thesis and explain related works for adversarial attacks and defenses related to this work.

In Chapter 2, we propose a novel evolutionary strategy-based algorithm, called EvoShift, for optimizing pixel attacks that are added to the training images. To obtain our training time attack, we solve a joint min-max optimization with the outer maximization designed to find the pixel noise and inner minimization designed to train the neural network on the noisy images. We impose a constraint on the optimization such that the cross-entropy (CE) loss on the noisy images is low, and the loss on the clean images is high. Such a formulation results in CNN trained on the noisy images to have a very high error on clean test images, thus exposing serious vulnerabilities in CNNs that are detrimental to robust learning. Interestingly, we find that optimization choice plays a vital role in generalization robustness. We show empirical evidence that SGD is resilient to our training time attacks, unlike adaptive optimization techniques (Adam). Although adaptive optimization methods are a popular choice for practitioners and researchers, we show that they can easily be overfit in the presence of our training time attacks. We believe that this is an important finding for the machine learning research community. We also apply regularization methods to counteract our proposed adversarial training time attack and find that most well-known regularization methods are ineffective against our attack. We find that random-crop data augmentation is moderately effective for a few pixel attacks. As a defense against our attack, we propose a novel robust loss function for CNN classification training that is resilient against our training attacks using an

information maximization framework. This result suggests that the traditional cross-entropy minimization framework for CNN training might cause non-robust feature learning, which might be mitigated by our proposed information-theoretic loss function. We introduce the concept of vulnerability in Generative Adversarial Networks (GANs) under proposed EvoShift attacks, causing poor image generation quality due to overfitting in the GAN discriminator.

In Chapter 3, we train DNN models on noisy training data using various optimizers and measure the performance of such models on clean test data, thus benchmarking how liable the optimizers are to overfit the training noise. We first construct a linearly separable two-class toy dataset upon which we superimpose a crafted noisy signal. Following that, we analytically show that adaptive gradient methods completely fail to learn any patterns from the data and do not generalize to the clean test set. On the other hand, SGD and its variants show 100% test accuracy on the test-set showing greater robustness against such spurious noise. Secondly, for higher-dimensional image datasets, we use a gradient-free noise optimization, based on our ICASSP 2020 paper for finding optimal pixel perturbations that maximize the generalization gap between training and testing images. Convolutional Neural Networks (CNN) models are trained on such worst-case noisy images using various optimizers. These trained models are then evaluated on clean data to measure the generalization of optimization methods. Empirical studies on MNIST, CIFAR10, and SVHN dataset confirm our hypothesis that vanilla SGD and its variants are significantly more robust against such perturbations compared to adaptive gradient methods. Our analysis of the 2D loss surface reveals that SGD tends to find solutions around flatter loss regions, which might explain our empirical observations. Based on our benchmarking results, we recommend using SGD optimizers with learning rate tuning instead of adaptive gradient methods, especially when there exists some training noise or distribution shift between the training and validation/testing data.

In Chapter 4, we expose the vulnerability of deep learning models for analyzing medical images under worst-case few pixel perturbations on training images. We emulate practical scenarios where few dots (almost imperceptible to human eyes) have appeared on the medical image because of some device noise. Therefore, we show that if the model is trained using those single/few pixels perturbed images, the network learns absolutely nuisance features instead of useful semantic features and provides unexpectedly low-test accuracy. We utilize the evolutionary strategy based few pixels perturbation algorithm from our ICASSP 2020 paper to corrupt the training images that maximally ignores the task-relevant features

(like shape and appearance) due to over-reliance on spurious distractor artifacts. We benchmark popular deep learning models on medical image datasets under such training time noise and show that even with single-pixel perturbations, deep models are susceptible to overfitting behavior similar to random classifiers. Informed by this analysis, we study input Covariate Shift Normalization (CSN), to reduce the effect of such spurious predictive features. Additionally, we analyze vanilla SGD and adaptive optimizers under such pixel perturbations in medical images and show that SGD is surprisingly more robust than adaptive methods (like ADAM) which is a default choice of optimization for most practitioners.

In Chapter 5, we show that adversarial features occupy a separate region in the frequency spectrum that can be disentangled from the regions occupied by semantically meaningful features in natural images. We use this concept to propose an adversarial defense against popular adversarial attacks. We empirically show that learning in the frequency domain can be used as a defense against adversarial images by a feed-forward operation of the frequency domain transformation of the input adversarial image through the frequency CNN. This method of defense outperforms previous input transformation based adversarial defense methods. Finally, we show that our method is robust against spatial transformation attacks such as rotations and translations, to which naturally trained CNNs show poor performance.

Finally, in Chapter 6, we provide conclusion from our experiments. This thesis provides empirical evidence regarding various properties of neural network training that we believe are of significance to the deep learning community. We hope this thesis will instigate future research on robustness of neural networks with respect to generalization and optimization techniques.