The Hong Kong University of Science and Technology

Department of Mathematics

PhD THESIS EXAMINATION

On the Generation, Understanding and Defense of Adversarial Examples in Deep Learning

By

Mr. Zhichao HUANG

ABSTRACT

Although deep learning has many practical applications, it is known that deep neural networks are vulnerable to adversarial examples, which are small perturbations of inputs that can fool neural networks into making wrong predictions. In this thesis, we propose new methods and theories to evaluate, understand and improve the adversarial robustness of deep neural networks. Firstly, we investigate the black-box adversarial attacks, where the attacker has no information about the target model except for its output. We propose two new methods, ZOHA and TREMBA, to accelerate the black-box attack. In ZOHA, second order information is incorporated into the zeroth-order optimization. In TREMBA, we utilize the transferability of adversarial examples, developing a new search space and greatly reducing the number of queries for black-box attacks. These algorithms demonstrate that black-box attack can be practical threats to the practical models. Secondly, we study the existence of adversarial examples and its relationship to benign overfitting. We provide a theoretical explanation why adversarial examples exist in standard training of neural networks: adversarial examples are by-products of overfitting the noise in the overparameterized models. Moreover, our theory explain the trade-off between the robustness and clean performance. Lastly, we improve the poor generalization of adversarial training by a novel test-time fine-tuning strategy. We propose to improve the generalization and robust accuracy of adversarially-trained networks via self-supervised test-time fine-tuning. To this end, we introduce a meta adversarial training method to find a good starting point for test-time fine-tuning. We provide theoretical justification for the necessity of test-time adaptation, and the empirical experiments also demonstrate the effectiveness of the proposed strategy.

Date : 30 September 2022, Friday

Time : 10:00 a.m.

Venue : Room 4475 (Lifts 25-26)

***Thesis Examination Committee*:**

Chairman	**:**	**Prof. Toyotaka ISHIBASHI, LIFS/HKUST**
Thesis Supervisor	**:**	**Prof. Tong ZHANG, MATH/HKUST**
Member	**:**	**Prof. Dong XIA, MATH/HKUST**
Member	**:**	**Prof. Can YANG, MATH/HKUST**
Member	**:**	**Prof. Qifeng CHEN, CSE/HKUST**
External Examiner	**:**	**Prof. Michael LYU, Department of Computer Science and Engineering/ The Chinese University of Hong Kong**

(Open to all faculty and students)

The student's thesis is now being displayed on the reception counter in the General Administration Office (Room 3461).