UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Adversarial Machine Learning on AI Model Attacks

by

Xinghao Yang

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2022

Certificate of Authorship/Originality

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Xinghao Yang declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science/Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Signature: Production Note: Signature removed prior to publication.

Date: 28-Jan-2022

ABSTRACT

Adversarial Machine Learning on AI Model Attacks

by

Xinghao Yang

Deep Neural Networks (DNNs) have achieved great success in multiple domains, stretching from Computer Vision (CV) to Natural Language Processing (NLP). However, recent studies demonstrated that DNNs are extremely vulnerable towards adversarial examples, which are original input with small perturbations. These perturbations are usually imperceptible to humans but mislead well-trained DNNs to erroneous output with high confidence. This phenomenon poses great concern of DNNs' robust performance on security-critical applications, such as traffic sign recognition and sentiment analysis. In this research, we focus on adversarial attacks, which is an effective strategy to understand DNNs behavior and promote their robust performance. Firstly, we proposed a Targeted Attention Attack (TAA) strategy to investigate the robustness of the traffic sign recognition system. Our TAA strategy takes the advantage of a soft attention map to reduce the attack cost and generates more natural perturbations to fit the real-world situations. Secondly, we designed the Bigram and Unigram based Semantic Preservation Optimization (BU-SPO) method to examine the vulnerability of deep models in text classification. The BU-SPO attacks text documents not only at the unigram word level but also at the bigram level to avoid producing meaningless sentences, where the Semantic Preservation Optimization (SPO) is designed to reduce the modification cost and improve the semantic consistency. Thirdly, we presented a BERT-based Simulated Annealing (BESA) algorithm to craft fluent text adversarial examples. The BESA mechanism employs the BERT Masked Language Model to generate context-aware word substitutions and adopts the Simulated Annealing to approach the global optima solution with a reasonable objective function.

Acknowledgements

I would like to dedicate my thesis to all those who have offered me tremendous assistance during the three years in University of Technology Sydney.

First of all, my heartiest thanks flow to my principal supervisor, Associate Professor Wei Liu, for his helpful guidance, valuable suggestions and constant encouragement both in my study and in my life. His profound insight and accurateness about my thesis taught me so much that they are engraved on my heart. He provided me with beneficial help and offered me precious comments during the whole process of my writing, without which the thesis would not be what it is now.

Also, I would like to express my sincere gratitude to my co-supervisor, Professor Dacheng Tao, who have periodically guide me that greatly broadened my horizon and enriched my knowledge in my study. His inspirational talking have provided me with a firm basis for the composing of this thesis and will always be of great value to my future academic research.

Lastly, my thanks would go to my beloved family for their loving considerations and great confidence in me all through these years. I also owe my sincere gratitude to my friends who gave me their help and time in listening to me and helping me work out my problems during the difficult course of the thesis.

Xinghao Yang
Sydney, Australia, 2022.

List of Publications

Journal Papers

J-1. **X. Yang**, W. Liu, J. Bailey, D. Tao, and W. Liu, "Semantic-Preserving Adversarial Text Attacks," *IEEE Transactions on Knowledge and Data Engineering.* Under Review.

J-2. **X. Yang**, W. Liu, S. Zhang, W. Liu and D. Tao, "Targeted Attention Attack on Deep Learning Models in Road Sign Recognition," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4980-4990, 15 March, 2021, doi: 10.1109/JIOT.2020.3034899.

J-3. **X. Yang**, W. Liu and W. Liu, "Tensor Canonical Correlation Analysis Networks for Multi-view Remote Sensing Scene Recognition," *IEEE Transactions on Knowledge and Data Engineering*, doi: 10.1109/TKDE.2020.3016208.

J-4. A. Chivukula, **X. Yang**, W. Liu, T. Zhu and W. Zhou, "Game Theoretical Adversarial Deep Learning with Variational Adversaries," *IEEE Transactions on Knowledge and Data Engineering*, doi: 10.1109/TKDE.2020.2972320.

J-5. **X. Yang**, W. Liu, W. Liu and D. Tao, "A Survey on Canonical Correlation Analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2349-2368, 1 June, 2021, doi: 10.1109/TKDE.2019.2958342.

Conference Papers

C-1. **X. Yang**, W. Liu, D. Tao, W. Liu. BESA: BERT-based Simulated Annealing for Adversarial Text Attacks. *The 30th International Joint Conference on Artificial Intelligence (IJCAI 2021).* August 21-26, 2021. **Accepted**.

C-2. C. Sung, **X. Yang**, C. Liao, and W. Liu. IntRoute: An Integer Programming based Approach for Best Bus Route Discovery. *The 26th International*

Conference on Database Systems for Advanced Applications (DASFAA 2021). April 11-14, 2021.

C-3. **X. Yang**, W. Liu, J. Bailey, D. Tao, and W. Liu. Bigram and Unigram Based Text Attack via Adaptive Monotonic Heuristic Search. *The 35th AAAI Conference on Artificial Intelligence (AAAI 2021).* February 2-9, 2021.

C-4. **X. Yang** and W. Liu. Population Location and Movement Estimation through Cross-domain Data Analysis. *The 29th International Joint Conference on Artificial Intelligence (IJCAI 2020).* January 7-15, 2021.

C-5. A. Chivukula, **X. Yang**, and W. Liu: Adversarial Deep Learning with Stackelberg Games. *The 26th International Conference on Neural Information Processing (ICONIP 2019).* December 12-15, 2019, Sydney, Australia.

J-1, J-2, C-1, and C-3 are most related to this thesis.

Contents

List of Figures

List of Tables