

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.

Production Note:

Signature: 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

Certificate	ii
Abstract	iii
Acknowledgments	iv
List of Publications	v
List of Figures	xi
List of Tables	xiii
1 Introduction	1
1.1 Background	1
1.1.1 Adversarial Attacks on Images	2
1.1.2 Adversarial Attacks on Texts	3
1.2 Research Objectives	4
1.3 Summary of Research Findings	6
1.4 Thesis Organization	6
2 Literature Survey	8
2.1 Adversarial Attacks on Images	8
2.1.1 Gradient Based Attack	8
2.1.2 Score Based Attack	10
2.1.3 Decision Based Attack	11
2.1.4 Transformation Based Attack	13

2.2	Adversarial Attacks on Texts	15
2.2.1	Character-level Attack	15
2.2.2	Sentence-level Attack	17
2.2.3	Word-level Attack	19
2.2.4	Multi-level Attack	22
2.3	Summary	23
3	Targeted Attention Attack on Deep Learning Models in Road Sign Recognition	25
3.1	Introduction	26
3.2	Method	28
3.2.1	RP ₂ and attention mechanism	28
3.2.2	Challenges from real-world conditions	30
3.2.3	Target Attention Attack Work-flow	31
3.3	Experiments	34
3.3.1	Datasets and Preprocess	34
3.3.2	Experimental Settings	35
3.3.3	Experimental Results	37
3.4	Summary	44
4	Semantic-Preserving Adversarial Text Attacks	46
4.1	Introduction	47
4.2	Method	51
4.2.1	Black-box Text Attack	51
4.2.2	Semantic Similarity	52
4.2.3	Bigram and Unigram Candidate Selection	53

4.2.4	Semantic Preservation Optimization	56
4.2.5	SPO with Semantic Filter (SPOF)	59
4.2.6	Targeted Attack Strategy	59
4.3	Experiments	60
4.3.1	Datasets	60
4.3.2	Victim Models	61
4.3.3	Evaluation Metrics	63
4.3.4	Baselines	64
4.3.5	Experimental Settings	66
4.3.6	Experimental Results and Analysis	67
4.3.7	Transferability	69
4.3.8	Adversarial Retraining	70
4.3.9	Targeted Attack Evaluations	71
4.4	Summary	72

5 BESA: BERT-based Simulated Annealing for Adversarial Text Attacks **73**

5.1	Introduction	73
5.2	The BESA Method	76
5.2.1	Black-box Untargeted Attack	76
5.2.2	BERT-based Candidate Selection	78
5.2.3	Simulated Annealing (SA) Optimization	79
5.3	Experiments	82
5.3.1	Datasets and Victim Models	82
5.3.2	Baselines	83

5.3.3	Evaluation Metrics and Experiment Settings	83
5.3.4	Experimental Results	85
5.3.5	Adversarial Training	87
5.3.6	Transferability	87
5.3.7	Qualitative Examples	88
5.3.8	Parameter Tuning	89
5.4	Summary	94
6	Conclusion and Future Work	96
6.1	Conclusion of this Thesis	96
6.2	Future Directions	97
	Bibliography	99

List of Figures

1.1	Successful adversarial examples from [92] to mislead AlexNet [46]. The perturbations are almost imperceptible by human vision system, but the AlexNet predicts the adversarial examples as “ostrich, struthio, camelus” from top to bottom.	2
3.1	(a) In our experiment, the “Stop” sign is perturbed by grayscale noises, then a well-trained CNN classifier gives a wrong prediction. (b) Three real-world road signs with tree shadows. Noises in (b) are similar to those in (a).	26
3.2	The TAA framework mainly contains two steps: soft attention map learning and perturbation optimization. The soft attention map is learned via a 92-layer RAN. The perturbation is optimized using the <i>target</i> attention map and a set of training images. The red arrow means if our target is “SpeedLimit45”, then TAA takes the soft attention map from “SpeedLimit45” instead of “Stop”.	31
3.3	Robustness test under a wide range of training epochs.	42
3.4	The RP_2 mask is crafted by several manually intervened post-processing steps.	42

4.1	The workflow of our BU-SPO with an text example “Study: CEOs rewarded for outsourcing. NEW YORK (CNN/Money) - The CEOs of the top 50 US companies that sent service jobs overseas pulled down far more pay than their counterparts at other large companies last year, a study said Tuesday.” This example is originally labeled as “Business” (66.68%) by LSTM model, but it is misclassified as “Sci/Tech” by replacing a bigram “New York” with “Empire State”. For brevity, we only select several words from the long text to display. The two green color boxes denote two successful attacks, but we accept the “Empire State” substitution, because it preserves more semantics (0.9877) than “AFP” (0.9724).	48
4.2	Transfer attack on Yahoo! Answers. Lower accuracy indicates higher transfer ability (the lower the better).	69
4.3	Adversarial retraining results. The higher the accuracy, the more robust of the model after retraining.	70
5.1	The number of grammar errors increased after attacks.	85
5.2	The semantic similarity before and after attacks.	85
5.3	Adversarial retraining results on MR dataset. The higher the accuracy, the more robust of the model after retraining.	87
5.4	Transfer attack on SST-2 dataset. Lower accuracy indicates higher transferability (the lower the better).	88

List of Tables

1.1	Three successful adversarial text examples generated by the character-level attack, sentence-level attack, and word-level attack strategies.	5
2.1	Summary of the properties for different attacking methods. The properties are T argeted attack, U ntargeted attack, W hite-box attack, B lack-box attack and G ray-box attack.	14
2.2	Summary of the properties for different text attacking methods. The properties are T argeted attack, U ntargeted attack, W hite-box attack, and B lack-box attack.	24
3.1	“Stop-SpeedLimit45” attack results.	38
3.2	“PedestrianCrossing-SpeedLimit65” attack results.	40
3.3	Transfer attack on GTSRB dataset.	41
3.4	Transfer attack on various of DNNs models.	43
3.5	Generalization test on more attack scenarios.	44
3.6	Physical world attack. Our TAA achieves high <i>ASR</i> with much less P_{loss}	45

4.1	The Comparison between unigram attacks and bigram attacks. one superiority of bigram substitution is that it can distinguish commonly used bigram phrases and avoid generating meaningless sentences.	47
4.2	Dataset information summarization. “# Avg. Words” is the average number of words for all samples.	61
4.3	Test Accuracy of Four DNNs Models before Attacks.	62
4.4	The Attack Success Rate (ASR) of various attack algorithms. For each row, the highest ASR is highlighted in bold, the second highest ASR is highlighted in underline, and the third highest ASR is denoted with italic.	63
4.5	The Average Word Replacement (AWR) number of various attack methods. For each row, the smallest AWR is highlighted in bold, the second smallest AWR is denoted in underline, and the third smallest AWR is represented with italic.	63
4.6	The average Universal Sentence Encoder (USE) score of various attack methods. For each row, the highest USE score is highlighted in bold, the second highest USE score is denoted in underline, and the third highest USE score is represented with italic.	64
4.7	Adversarial examples of IMDB (attack Word CNN). Green texts are original words, while red ones are substitutions.	65
4.8	Adversarial examples by attacking Word LSTM model on AG’s News dataset.	68
4.9	Adversarial examples by attacking Bi-LSTM model on Yahoo! Answers dataset.	69
4.10	Targeted attack results on AG’s News dataset.	71

5.1	Statistic information of the five datasets. “# Words” denotes the average number of words (i.e., average text length).	83
5.2	The attack success rate (ASR) and average word substitution rate (WSR) of various attack algorithms on five text datasets. The best results are highlighted in bold . The “ACC” column shows the original test accuracy without attacks, and DisBERT is short for DistilBERT.	84
5.3	Time (in seconds) needed in attacking the BERT model.	86
5.4	Adversarial examples crafted by BESA. The green and red color denotes the original and substitution words, respectively.	88
5.5	SNLI adversarial examples by attacking BERT.	89
5.6	QNLI adversarial examples by attacking BERT.	89
5.7	IMDB adversarial examples by attacking DistilBERT.	90
5.8	MR adversarial examples by attacking LSTM.	91
5.9	SST-2 adversarial examples by attacking LSTM.	91
5.10	The attack results with different USE threshold. “# Grammar” denotes the number of grammar errors increased after attacking. . . .	92
5.11	Attack results with different internal simulation steps.	92
5.12	Attack results with different lowest temperature.	93
5.13	Attack results with different highest temperature.	93
5.14	Attack results by simultaneously tuning the highest temperature and lowest temperature.	94
5.15	Attack results with different attack radius (σ).	94
5.16	Attack results with different balance parameter (δ).	94