Minhao Cheng

CONTACT INFORMATION

Department of Computer Science
UCLA

Tel: (1) 530-601-8331
E-mail: mhcheng@cs.ucla.edu

EDUCATION

University of California Los Angeles, Los Angeles, USA
Ph.D, Computer Science, September 2018 - March 2021

University of California Davis, Davis, USA
Ph.D student, Computer Science, September 2015 - August 2018

University of Electronic Science and Technology of China, Chengdu, China
B.S., Computer Science and Technology, July, 2015

RESEARCH INTERESTS

Machine learning Security, Adversarial robustness, Deep Learning, Optimization, AutoML.

WORK EXPERIENCE

Microsoft Research Intern, Redmond, USA
Robust Training Method for Better Generalization **June.2020 - Sep.2020**

IBM Research Intern, Yorktown Heights, USA
Scalable Training Method for Adversarial Robustness **June.2019 - Sep.2019**

Rakuten Slice Intern, San Mateo, USA
Hierarchical Classification Using Neural Networks **June.2017 - Sep.2017**

PUBLICATION

Google Scholar Profile: Number of Citations=580+; h-index = 10, i10-index = 10. Details available at `https://scholar.google.com/citations?user=_LkC1yoAAAAJ&hl=en`

Ruochen Wang, **Minhao Cheng**, Xiangning Chen, Xiaocheng Tang, Cho-Jui Hsieh. Rethinking Architecture Selection in Differentiable NAS. To Appear In International Conference on Learning Representations (ICLR), 2021. (Outstanding Paper Award)

Xiangning Chen*, Ruochen Wang*, **Minhao Cheng***, Xiaocheng Tang, Cho-Jui Hsieh. DrNAS: Dirichlet Neural Architecture Search. To Appear In International Conference on Learning Representations (ICLR), 2021. (* Equal Contribution)

Minhao Cheng, Pin-Yu Chen, Sijia Liu, Shiyu Chang, Cho-Jui Hsieh, Payel Das. Self-Progressing Robust Training. To Appear In AAAI Conference on Artificial Intelligence (AAAI), 2021.

Xiaoqing Zheng, Jiehang Zeng, Yi Zhou, Cho-Jui Hsieh, **Minhao Cheng**, Xuanjing Huang. Evaluating and enhancing the robustness of neural network-based dependency parsing models with adversarial examples. In ACL (long), 2020.

Minhao Cheng*, Simranjit Singh*, Patrick H. Chen, Pin-Yu Chen, Sijia Liu, Cho-Jui Hsieh.

Sign-OPT: A Query-Efficient Hard-label Adversarial Attack. In International Conference on Learning Representations (ICLR), 2020. (* Equal Contribution)

Minhao Cheng, Jinfeng Yi, Huan Zhang, Pin-Yu Chen, Cho-Jui Hsieh. Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples. In AAAI Conference on Artificial Intelligence (AAAI), 2020.

Yu-Lun Hsieh, **Minhao Cheng**, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, Cho-Jui Hsieh. On the Robustness of Self-Attentive Models. In Proceedings of Association for Computational Linguistics (ACL), 2019.

Minhao Cheng, Wei Wei, Cho-Jui Hsieh: Evaluating and Enhancing the Robustness of Dialogue Systems: A Case Study on a Negotiation Agent. In Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019.

Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh: Query-Efficient Hard-label Black-box Attack: An Optimization-based Approach. In International Conference on Learning Representations (ICLR), 2019.

Huang Fang, **Minhao Cheng**, Cho-Jui Hsieh, Michael Friedlander: Fast Training for Large-Scale One-versus-All Linear Classifiers using Tree-Structured Initialization. In SIAM International Conference on Data Mining (SDM), 2019.

Yao Li, **Minhao Cheng**, Kevin Fujii, Fushing Hsieh, Cho-Jui Hsieh. Learning from Group Comparisons: Exploiting Higher Order Interactions. In Advances in Neural Information Processing Systems (NIPS), 2018.

Xuanqing Liu, **Minhao Cheng**, Huan Zhang, Cho-Jui Hsieh. Towards Robust Neural Networks via Random Self-ensemble. European Conference on Computer Vision (ECCV), 2018.

Minhao Cheng, Ian Davidson, Cho-Jui Hsieh. Extreme Learning to Rank via Low Rank Assumption. International Conference on Machine Learning (ICML), 2018.

Minhao Cheng, Cho-Jui Hsieh. Distributed Primal-Dual Optimization for Non-uniformly Distributed Data. International Joint Conference on Artificial Intelligence (IJCAI), 2018.

Huang Fang, **Minhao Cheng**, Cho-Jui Hsieh. A Hyperplane-based Algorithm for Semi-supervised Dimension Reduction. IEEE International Conference on Data Mining (ICDM), 2017.

Dazhuang Su, Xinzheng Niu, **Minhao Cheng**. Intelligent Mobile Framework Based on Swarm Computation. CIT/IUCC/DASC/PICom 2015: 1000-1006.

PREPRINT

Minhao Cheng, Zhe Gan, Yu Cheng, Shuohang Wang, Cho-Jui Hsieh, Jingjing Liu. Adversarial Masking: Towards Understanding Robustness Trade-off for Generalization.

Minhao Cheng, Qi Lei, Pin-Yu Chen, Inderjit Dhillon, Cho-Jui Hsieh. CAT: Customized Adversarial Training for Improved Robustness.

Huan Zhang, **Minhao Cheng**, Cho-Jui Hsieh. Enhancing Certifiable Robustness via a Deep Model Ensemble.

Xiaoyun Wang, **Minhao Cheng**, Joe Eaton, Cho-Jui Hsieh, Felix Wu. Attack graph convolu-

tional networks by adding fake nodes.

Liu Liu, **Minhao Cheng**, Cho-Jui Hsieh, Dacheng Tao. Stochastic Zeroth-order Optimization via Variance Reduction method.

Patents

Minhao Cheng, Pin-Yu Chen, Sijia Liu, Shiyu Chang, Payel Das. Method and System of Training Robust Machine Learning Models.

Minhao Cheng, Xiaocheng Tang, Chu-Cheng Hsieh. Hierarchical Classification Using Neural Networks.

Research Experience

Differentiable Neural Architecture Search

- **Rethinking DARTS**: While much has been discussed about the supernets optimization, the architecture selection process has received little attention. We provide empirical and theoretical analysis to show that the magnitude of architecture parameters does not necessarily indicate how much the operation contributes to the supernets performance. We then propose an alternative perturbation-based architecture selection that directly measures each operations influence.
- **DrNAS:** One big trend on neural architecture search is based on the continuous relaxation of the architecture representation, allowing the efficient search of the architecture using gradient descent. However, it suffers from poor generalization. We, therefore, propose a novel differentiable architecture search method by formulating it into a distribution learning problem. We achieve state-of-the-art results on all three popular datasets.

Adversarial Defense

- **CAT:** Adversarial training is one of the most successful adversarial defense method. However, it suffers from poor generalization on both clean and perturbed data. Instead, we propose a new algorithm that adaptively customizes the perturbation level and the corresponding label for each training sample in adversarial training. We show that the proposed algorithm achieves better clean and robust accuracy than previous adversarial training methods.

- **SPROUT**: As a common method to improve adversarial robustness, adversarial training has been widely adopted. However, it suffers from bad prediction accuracy and scalability. In this project, we use a vicinal function to help the model learn better adversarial robustness and generalization.

- **AdvMask:** Although it has shown effectiveness in improving model's robustness, adversarial training suffers from bad model prediction accuracy. We study why this trade-off happens and find that batch normalization has a major impact. Then we propose a new method to utilize this finding to improve generalization and the tradeoff between accuracy and robustness.

Hard-Label Black-box Attack

- **OPT-attack:** It has been shown that DNNs models are vulnerable to a very small human-imperceptible perturbation. However, it is still a challenge when we could only query hard-label instead of probability output. We develop a query efficient algorithm which could apply to industrial-strength image classifiers.
- **Sign-OPT:** To further improve query efficiency of OPT-attack, we propose a single query ora-

cle for retrieving signs of directional derivative and develop a novel approach for the hard-label black-box attack. It only requires 5× to 10× fewer queries compared with the previous state-of-art method.

Evaluating Robustness of NLP models

- **Seq2sick:** Recent research on DNNs has indicated ever-increasing concern on the robustness to adversarial examples. We design algorithms to generate adversarial examples for the sequence to sequence model which is widely used in machine translation, text summarization.
- **AdvAgent:** Although we have shown that the sequence-to-sequence model is not robust against adversarial attacks, it is still an open question about the robustness of the dialog system implemented by deep neural networks. We then develop two algorithms that could attack the dialog system successfully with or without knowing the deep neural network model structure.

Ranking and Recommendation

- **Factorial RankSVM:** We develop new low-rank approximation method using for large-scale recommendation system. Previously, it uses the ranksvm and its variations to train the data. Now we use the low-rank method to get better speed and accuracy dealing with the pair-wised data.

Distributed Machine Learning

- **Stochastic BlockLS:** New line search method in stochastic algorithms in large-scale distributed machine learning scenario to overcome the large Primal-Dual gap in the training periods so that we can achieve a faster convergence and a better accuracy in the distributed machine learning case.

Professional Services

- Paper reviewer & Programming Committee: NIPS '16, ICML '17, NIPS '17, ICML '18, NIPS '18, ICML '19, AAAI '19, NIPS '19, AAAI '20, IJCAI '20, ACL '20, ICML '20, NeurIPS '20, ICLR '21, AAAI '21.

Teaching Experience

University of California, Los Angeles, Los Angeles, USA

- Teaching Assistant in CS 180 Introduction to Algorithms. **Mar.2020 - Jun.2020**
- Teaching Assistant in CS 33 Introduction to Computer Organization. **Sep.2019 - Dec.2019**
- Teaching Assistant in CS 260 Machine Learning Algorithms. **Jan.2019 - Mar.2019**

University of California, Davis, Davis, USA

- Teaching Assistant in ECS 122B Algorithms. **Mar.2017 - Jun.2017**
- Teaching Assistant in ECS 122A Algorithms. **Apr.2016 - June.2016/ Sep.2016 - Dec.2016**
- Teaching Assistant in STA 250 Optimization. **Jan.2016 - Mar.2016**