Adversarial Robustness of Deep Learning:
Theory, Algorithms, and Applications

Wenjie Ruan
University of Exeter
Exeter, UK
wruan@exeter.ac.uk

Xinping Yi
University of Liverpool
Liverpool, UK
xinping.yi@liverpool.ac.uk

Xiaowei Huang
University of Liverpool
Liverpool, UK
xiaowei.huang@liverpool.ac.uk

ABSTRACT
This tutorial aims to introduce the fundamentals of adversarial robustness of deep learning, presenting a well-structured review of up-to-date techniques to assess the vulnerability of various types of deep learning models to adversarial examples. This tutorial will particularly highlight state-of-the-art techniques in adversarial attacks and robustness verification of deep neural networks (DNNs). We will also introduce some effective countermeasures to improve robustness of deep learning models, with a particular focus on adversarial training. We aim to provide a comprehensive overall picture about this emerging direction and enable the community to be aware of the urgency and importance of designing robust deep learning models in safety-critical data analytical applications, ultimately enabling the end-users to trust deep learning classifiers. We will also summarize potential research directions concerning the adversarial robustness of deep learning, and its potential benefits to enable accountable and trustworthy deep learning-based data analytical systems and applications.

CSC CONCEPTS
- Computing methodologies → Artificial intelligence.

ACM Reference Format:
Wenjie Ruan, Xinping Yi, and Xiaowei Huang. 2021. Adversarial Robustness of Deep Learning: Theory, Algorithms, and Applications. In Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM ’21), November 1–5, 2021, Virtual Event, QLD, Australia. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3459637.3482029

1 RATIONALE
In recent years, we witness significant progress has been made in deep learning, which can achieve human- or superhuman-level performance on various data analytical tasks, such as image data recognition [25], natural language processing [4], and medical data analysis [5]. Given the prospect of a broad deployment of DNNs in a wide range of applications, concerns regarding the safety and trustworthiness of deep learning have been recently raised [7, 39]. There is significant research that aims to address these concerns, with many publications appearing since the year of 2014 [11]. As we seek to deploy deep learning systems not only on virtual domains, but also in real systems, it becomes critical that a deep learning model can obtain satisfactory performance, but which are truly robust and reliable. Although many notions of robustness and reliability exist in different communities, one particular topic in machine learning community that has attract enormous attention in recent years is the adversarial robustness of deep learning: a deep learning model is fragile or extremely non-robust to an input that is adversarially perturbed, and such perturbations usually are invisible or insensible by humans [7]. This is although a very specific notion of robustness in general, but one that concerns the safety and trustworthiness of modern deep learning systems.

This tutorial seeks to provide a broad, hands-on introduction to the topic concerning adversarial robustness: the widespread vulnerability of state-of-the-art deep learning models to adversarial misclassification (i.e., adversarial examples). The goal is to combine both formal mathematical treatments and practical tools and applications that highlight some of the key methods and challenges concerning the adversarial robustness. This tutorial will specifically concentrate on three major research progress on this direction - adversarial attacks, defences and verification. As detailed in Section 5, some tutorials regarding this emerging direction already appeared in flagship conferences in machine learning, AI and computer vision communities, including IJCAI 2021 1, ECML/PKDD 2021 2, ICDM 2021 3, CVPR 2020, KDD 2019 4, etc. Our tutorial is fundamentally different to those existing ones, which is i) more comprehensive: we not only cover adversarial attacks but particularly concentrate on verification-based approaches which is able to establish formal robustness guarantees; ii) more application-oriented: in the second part of our tutorial, we emphasize one particular defence technique that can significantly improve robustness of DNNs - adversarial training, which will shed a light on the development of robust deep learning models for real-world data analytical applications. The specific differences to each current similar tutorials are detailed in Section 5.

This tutorial can alarm the community to be aware of the safety vulnerabilities of deep learning on real-world data analytical solutions despite its appealing performance. We also envision, through this tutorial, AI and data mining researchers and engineers get a sense on how to evaluate the robustness of deep learning models (e.g., via adversarial attacks/perturbations and verification-based approaches) and how to design/train robust deep learning models (e.g., via defence, especially adversarial training).

2 CONTENT DETAILS
- Introduction to adversarial robustness: this part will introduce the concept of adversarial robustness by showing some examples from computer vision [45], natural language processing [13], medical systems [35], and autonomous systems [34]. Specifically, we will demonstrate the vulnerabilities of various types of deep learning models to different adversarial examples. We will also highlight the dissimilarities of research focuses on adversarial robustness from different communities, i.e., attack, defence and verification.

1 Towards Robust Deep Learning Models: Verification, Falsification, and Rectification in IJCAI 2021 (https://tutorial-ijcai.trustai.uk/)
2 https://tutorial-ecml.trustai.uk/
3 Recent Progress in Zeroth Order Optimization and Its Applications to Adversarial Robustness in Data Mining and Machine Learning, in KDD 2019, CVPR 2020
• **Adversarial attacks**: this part will detail some famous adversarial attack methods with an aim to provide some insights of why adversarial examples exit and how to generate adversarial perturbation effectively and efficiently. Specifically, we will present six typical adversarial attacks, including L-BFGS [30], FGSM [8], C&W [3], ZeroAttack [31], spatial-transformed attacks [38], universal attacks [44]. In the end of this part, we will also briefly introduce some adversarial attacks on other domains, including attacks on sentiment analysis systems [13], attacks on 3D point cloud models [9], attacks on audio recognition systems [1].

• **Adversarial defense**: this part will present an overview of state-of-the-art robust optimisation techniques for adversarial training [18], with emphasis on distributional robustness and the interplay between robustness and generalisation. In particular, adversarial training with Fast Gradient Method (FGM) [8], Projected Gradient Method (PGM) [16] will be introduced briefly, followed by the advanced methods promoting distributional robustness [27] from the view points of robustness versus accuracy [42], supervised versus semi-supervised learning [20], and the exploitation of local and global data information [22, 41]. In addition, the interplay between robustness and generalisation will be discussed with respect to generalisable robustness and robust generalisation. A variety of regularisation techniques such as spectral normalisation [19], Lipschitz regularisation [32], and weight correlation regularisation [14] to promote generalisable robustness will be discussed, together with some recent advances to improve robust generalisation [26, 40, 43].

• **Verification and validation**: this part will review the state-of-the-art on the verification techniques for checking whether a deep learning model is dependable. First, we will discuss verification techniques for checking whether a convolutional neural network is robust against an input, including constraint solving based techniques (MILP, Reluplex [15]), approximation techniques (MaxSens [37], AI² [6], DeepSymbol [17]), and global optimisation based techniques (DLV [12], DeepGO [23, 24], DeepGame [33, 36]). Second, we will discuss some software testing based methods which generate a large set of test cases according to coverage metrics, including e.g., DeepXplore [21], and DeepConcolic [2, 28, 29], and their extension to recurrent neural networks [10]. The dependability of a learning model can then be estimated through the test cases. Third, we will discuss how to extend these above techniques to work with a reliability notion which considers all possible inputs in an operational scenario [46]. This requires the consideration of robustness and generalisation in a holistic way [14, 47]. Finally, we will summarize and outlook current state of this research field and future perspectives.

3 **TARGET AUDIENCE AND PREREQUISITES**

This tutorial motivates and explains a topic of emerging importance for AI, and it is particularly devoted to anyone who is concerning the safety and robustness of deep learning models. The target audience would be data mining and AI researchers and engineers who wish to learn how the techniques of adversarial attacks and verification as well as adversarial training can be effectively used in evaluating and improving the robustness of deep learning models. No knowledge of the tutorial topics is assumed. A basic knowledge of deep learning and statistical pattern classification is requested.

4 **BENEFITS**

Deep learning techniques now is not only pervasive in the community of computer vision and machine learning, but also widely applied on data analytical systems. For researchers and industrial practitioners who are developing safety-critical systems such as health data analytics, malware detection, and automatic disease diagnosis, the robustness and reliability of deep learning models are profoundly important. CIKM, as a flagship conference in data mining and knowledge management, has attracted huge amount of data scientists and engineers and many of them are using deep learning techniques. We envision, by presenting a tutorial concerning the robustness of deep learning at CIKM’21, the community can (i) be aware the vulnerability of deep learning models, (ii) understand why such vulnerability exits in deep learning and how to evaluate its adversarial robustness, and (iii) know how to train a robust deep learning model. We believe, those mentioned goals are appealing to the audience in CIKM’21. In the meantime, a tutorial concerning similar topic already appeared in SIGKDD’19 and ICDM 2020, two flagship conferences in data mining. As such, we believe it is necessary and urgent to propose a more comprehensive tutorial concentrating this topic in CIKM’21 as well.

5 **DIFFERENCE TO SIMILAR TUTORIALS**

- **Rigorous Verification and Explanation of ML Models, in AAAI 2020.** Link: https://alexyIgnatiev.github.io/aaai20-tutorial/
  - The overlapping will be on the verification part, where the above tutorial only considers from the logic/binary perspective, while we will cover comprehensively constraint-solving based methods, approximation methods, and global optimisation methods. The other two topics of this tutorial, i.e., adversarial attack (Part-I) and defense (Part-III), are not covered in the above tutorial.
- **Adversarial Machine Learning, in AAAI 2019, AAAI 2018.** Link: https://aaai19.adversarial.github.io/index.html#org
  - The above tutorial focuses on adversarial attacks of classifier evasion and poisoning as well as the corresponding defense techniques, while this tutorial places the emphasis on the fundamentals of adversarial examples (Part-I) and generalisable robust optimisation techniques for defense (Part-III), which are not covered by the above tutorial. In addition, the topic of verification of this tutorial (Part-II) is not covered at all in the above one.
- **Adversarial Machine Learning, in IJCAI 2018, ECCV 2018, ICCV 2017.** Link: https://www.pluribus-one.it/research/sec-ml/wild-patterns
  - The above tutorial concentrates on demonstrating vulnerability of various machine learning models and the design of learning-based pattern classifiers in adversarial environments. Our tutorial, however, is primarily about adversarial robustness of deep neural networks, especially the safety verification (Part-II) and adversarial defense (Part-III) on DNNs are not covered by the above tutorial. The only overlapping will be the part of adversarial attacks, but ours is more comprehensive and deep learning-focused.
- **Adversarial Robustness: Theory and Practice, in NIPS 2018** Link: https://adversarial-ml-tutorial.org/
  - The above tutorial was given two years ago, concentrating on the verification-based approaches to establishing formal guarantees for adversarial robustness. Then it presents adversarial training and regularization-based methods for improving robustness. Our tutorial will be more up-to-date. The major differences are: 1) in adversarial attacks, we present more recent and advanced adversarial examples, such as universal and spatial-transformed one; 2) in verification, we are more comprehensive, except for

---

\[ ^{1}\text{Recent Progress in Zeroth Order Optimization and Its Applications to Adversarial Robustness in Data Mining and Machine Learning, in CVPR2020, KDD 2019.} \]
the constraint-solving based methods, we also cover the approxi-
mation and global optimisation methods; and (3) in adversarial
defense, our focus is on generalisable adversarial training via
advanced spectral regularisation.

• Recent Progress in Zeroth Order Optimization and Its Applications
  to Adversarial Robustness in Data Mining and Machine Learning,
  in CVPR 2020, KDD 2019.
  Link: https://sites.google.com/view/adv-robustness-zoopt

  The above tutorial concentrates on Zero-order optimization meth-
ods with a particular focus on black-box adversarial attacks to
DNNs. Our tutorial is more comprehensive, which not only covers
a wider range of adversarial attacks, but also presents verification
approaches that can establish formal guarantees on adversarial
robustness, as well as review state-of-the-art adversarial training
methods that can defend adversarial attacks and improve DNN's
robustness.

6 RELEVANT EXPERIENCE ON THE TOPIC

6.1 Relevant Tutorial Experience

• W. Ruan, X. Yi, X. Huang, Tutorial "Adversarial Robustness of
  Deep Learning Theory, Algorithms, and Applications", The 20th
  IEEE International Conference on Data Mining (ICDM 2020),
  17-20 Nov 2020, Sorrento, Italy.
  Link: https://tutorial.trustdeeplearning.com/

• W. Ruan, E. Botoeva, X. Yi, X. Huang, Tutorial "Towards Robust
  Deep Learning Models: Verification, Falsification, and Rectifi-
  cation", The 30th International Joint Conference on Artificial
  Intelligence (IJCAI 2021), 21-26 Aug 2021, Canada.
  Link: http://tutorial-ijcai.trustai.uk/

• W. Ruan, X. Yi, X. Huang, Tutorial "Adversarial Robustness of
  Deep Learning Theory, Algorithms, and Applications", The 2021
  European Conference on Machine Learning and Principles and
  Practice of Knowledge Discovery in Databases (ECML/PKDD
  2021), 13-17 Sep 2021, Virtual.
  Link: http://tutorial-ecml.trustai.uk/

6.2 Citations of Relevant Works

The following are the selected papers published by the presenters
which are related to this tutorial6.

(1) Safety verification of deep neural networks, CAV 2017, Citation = 623

(2) Structural Test Coverage Criteria for Deep Neural Networks,
    in ACM Transactions on Embedded Computing Systems 2018,
    Citation = 196

(3) Concolic Testing for Deep Neural Networks, in ASE 2018, Citation = 183

(4) Feature-guided black-box safety testing of deep neural net-
    works, in TACAS 2018, Citation = 162

(5) Reachability Analysis of Deep Neural Networks with Provable
    Guarantees, in IJCAI 2018, Citation = 163

(6) A survey of safety and trustworthiness of deep neural networks:
    Verification, testing, adversarial attack and defence, and inter-
    pretability, in Computer Science Review, 2020. Citation = 111

(7) A game-based approximate verification of deep neural networks
    with provable guarantees, in Theoretical Computer Science,
    2020. Citation = 61

(8) Global Robustness Evaluation of Deep Neural Networks with
    Provable Guarantees for the Hamming Distance, in IJCAI 2019,
    Citation = 53

The overall citations of the above papers that are closely related to
this tutorial are over 1,200 since 2017. The details are as below:

• A seminal paper (1), on the safety verification of deep learning,
  has attracted 600+ Google Scholar citations, which is one of
  the first papers on the verification of deep learning.

• A few papers including (5) (7) (8), on the verification of neu-
  ral networks through global optimisation algorithms, have at-
  tracted 250+ citations.

• A few papers including (2) (3) (4) (5), on the adversarial at-
  tacks of the robustness of neural networks, have attracted 500+ citations.

• A recent survey paper (6), closely aligned with the topic of this
tutorial, has attracted 100+ Google scholar citations since its
publication in 2020.

7 BRIEF RESUMES OF PRESENTERS

7.0.1 Dr Wenjie Ruan. Dr Wenjie Ruan is a Senior Lecturer of Data
Science at University of Exeter, UK. Previously, he has worked at
Lancaster University as a lecturer, and University of Oxford as a post-
doctoral researcher. Dr Ruan got his PhD from University of Adelaide,
Australia. His series of research works on Device-free Human Localiza-
tion and Activity Recognition for Supporting the Independent Living of
the Elderly have received Doctoral Thesis Excellence from The Univer-
sity of Adelaide. He was also the recipient of the prestigious DECRP
fellowship from ARC (Australian Research Council). Dr Ruan has pub-
lished 30+ top-tier papers in top venues such as AAAI, IJCAI, ICDM,
UbiComp, CIKM, ASE, etc. His recent work on reachability analysis
on deep learning is one of the most citable papers in IJCAI'18 (150+
citations since 2018), and his work on testing-based falsification on deep
learning is also one of the most citable papers in ASE'18 (150+ citations
since 2018). Dr. Ruan has served as Senior PC, or PC member for over
10 conferences including IJCAI, AAAI, ICML, NeurIPS, CVPR, ICCV,
AAMAS, ECML-PKDD, etc. His homepage is: http://wenjierruan.com/.

7.0.2 Dr Xinping Yi. Dr Xinping Yi is a Lecturer (Assistant Professor)
of Electrical Engineering at the University of Liverpool, UK. He re-
ceived his Ph.D. degree from Telecom ParisTech, Paris, France. Prior
to Liverpool, he worked at Technische Universitat Berlin, Germany,
EUROCOM, France, UC Irvine, US, and Huawei Technologies, China.
Dr Yi’s recent research lies in deep learning theory with emphasis
on generalisation and adversarial robustness. Dr Yi has published
40+ papers in IEEE transactions such as IEEE Transactions on In-
fOrmation Theory (TIT), and machine learning conferences such as
ICML, NeurIPS. He has served as programme committee members
and reviewers at a number of international conferences and jour-
nals, such as ICML, ICLR, IJCAI, CVPR, ICCV, ISIT, ICC, TIT, Pro-
ceedings of IEEE, JSAC, TWC, Machine Learning. His homepage is:
https://sites.google.com/site/xinpingyi00/.

7.0.3 Dr Xiaowei Huang. Dr Xiaowei Huang is Reader of Computer
Science at the University of Liverpool, UK. Prior to Liverpool, he
worked at Oxford and UNSW Sydney. Dr Huang’s research concerns
the safety and trustworthiness of autonomous intelligent systems. He
is now leading a research group working on the verification, validation,
and interpretability of deep neural networks. Dr Huang is the prin-
ciple investigator of several Dstl projects concerning the safety and
assurance of artificial intelligence, and the Liverpool lead on a H2020
Acknowledgement. Wenjie Ruan is supported by Offshore Robotics for Certification of Assets (ORCA) Partnership Resource Fund (PRF) on Towards the Accountable and Explainable Learning-enabled Autonomous Robotic Systems (AELARS) [EP/R026173/1]. XH has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 956123, and is also supported by the UK EPSRC (through the Offshore Robotics for Certification of Assets [EP/R026173/1] and End-to-End Conceptual Guarding of Neural Architectures [EP/T026995/1]).

REFERENCES

[1] Hadi Abdollahi, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin BB Butler, and Joseph Wilson. 2019. Practical hidden voice attacks against speech and speaker recognition systems. arXiv preprint arXiv:1904.05754 (2019).

[2] Nicolas Berthier, Yasheng Sun, Wei Huang, Yanghao Zhang, Wenjie Ruan, and Xiaowei Huang. 2021. Tutorials on Testing Neural Networks. arXiv preprint arXiv:2108.07374 (2021).

[3] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In 2017 IEEE symposium on security and privacy (pp. 38–47). IEEE.

[4] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, Pierre-Antoine Lamblin, and Jean S Clark. 2011. Natural language processing (almost) from scratch. Journal of machine learning research 12, 1 (2011), 2493–2537.

[5] Andre Esteve, Alexandre Robicquet, Bharrat Ramsundar, Volodymyr Kuleshov, Mark Depristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. Nature medicine 25, 1 (2019), 24–29.

[6] Timon Ceho, Matthew Nirman, Dana Drachler-Cohen, Petar Tsankov, Saurabh Chaudhuri, and Martin Vechev. 2018. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In SP 2018. 3–18.

[7] Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. 2018. Making machine learning robust against adversarial inputs. Commun. ACM 61, 7 (2018), 56–66.

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. In n.d. Explaining and harnessing adversarial examples. ICLR 2015 (In n.d.).

[9] Fahad Hamid, Saeed Rehman, Anil V. Joshi, Shumaila Azam, and Abdullah Hamdi. 2020. Adtpc: Transferable adversarial perturbations on 34 point clouds. In European Conference on Computer Vision. Springer, 241–257.

[10] Wenjie Ruan, Yasheng Sun, Xingyu Zhao, Zhiyuan Jiang, James Sharp, Wenjie Ruan, Jie Meng, and Xiaowei Huang. 2021. Coverage-Guided Testing for Recurrent Neural Networks. IEEE Transactions on Reliability (2021), 1–16. https://doi.org/10.1109/TR.2021.3086664

[11] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Yuqing Zhao, Min Wu, and Xiuping Yi. 2020. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defense, and interpretability. Computer Science Review 37 (2020), 100270.

[12] Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. 2017. Safety verification of deep neural networks. In International Conference on Computer Aided Verification. Springer, 3–29.

[13] Di Jin, Zhijing Jin, Jie Tsayni Zhou, and Peter Stolovits. 2020. Is bert really robust? A strong baseline for natural language attack on text classification and entailment. In Proceedings of the AAAI conference on artificial intelligence, Vol. 34. 8018–8025.

[14] Gaojie Jin, Xinping Yi, Liang Zhang, Lijun Zhang, Sven Schewe, and Xiaowei Huang. 2021. Provable Guarantees for Hamming Distance. In International Joint Conference on Artificial Intelligence (IJCAI). 2651–2659.

[15] Wenjie Ruan, Yasheng Sun, and Marta Kwiatkowska. 2019. Global Robustness Evaluation of Deep Neural Networks with Provable Guarantees for Hamming Distance. In IEEE 9544–9592.

[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International journal of computer vision 115, 3 (2015), 211–252.

[17] Ludwig Schmidt, Shahriar Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. 2018. Adversarially robust generalization requires more data. In NeurIPS 2018. 5019–5031.

[18] Aman Sinha, Hongseok Namkoong, and John Duchi. 2018. Certifying Some Distributional Robustness with Pronevolent Adversarial Training. In International Conference on Learning Representations.

[19] Youcheng Sun, Xiaowei Huang, Daniel Kroening, James Sharp, Matthew Hill, and Rob Amore. 2019. Structural Test Coverage Criteria for Deep Neural Networks. ACM Trans. Embed. Comput. Syst. 18, 5s, Article 94 (Oct. 2019), 23 pages.

[20] Youcheng Sun, Min Wu, Wenjie Ruan, Xiaowei Huang, Marta Kwiatkowska, and Daniel Kroening. 2018. Concise Testing for Deep Neural Networks. In ASE 109–119.

[21] Christopher Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013).

[22] Chun-Tsu Tain, Fu-Wen Tsai, Siyu Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. 2019. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 742–749.

[23] Abdalnournaa and Kevin Tannous. 2018. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In Advances in Neural Information Processing Systems. 3835–3844.

[24] Matthew Wicker, Xiaowei Huang, and Marta Kwiatkowska. 2018. Feature-guided black-box safety testing of deep neural networks. In TACAS. 408–426.

[25] Han Wu and Wenjie Ruan. 2021. Adversarial Driving: Attacking End-to-End Autonomous Systems. arXiv preprint arXiv:2105.09151 (2021).

[26] Han Wu, Wenjie Ruan, Jiachao Wang, Dingshang Zheng, Bei Liu, Yuyan Geng, Xiangfei Chai, Jian Chen, Kunwei Li, Shaolin Li et al. 2021. Interpretable machine learning for covid-19: An empirical study on severity prediction task. IEEE Transactions on Biomedical Engineering.

[27] 2019. Theoretically principled trade-off between robustness and accuracy. In International Conference on Machine Learning. 3835–3844.

[28] Haichao Zhang and Jianyu Wang. 2019. Defense against adversarial attacks using feature scattering-based adversarial training. Advances in Neural Information Processing Systems 32 (2019), 1831–1841.

[29] Hengyong Zhang, Yaodong Yu, Jianfao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. In International Conference on Machine Learning. PMLR, 7472–7482.

[30] Shuifei Zhang, Zhihui Qian, Xiaohua Huang, Qinfeng Wang, Rui Zhang, and Xiaoping Yi. 2021. Towards Better Robust Generalization with Shift Consistency Regularization. In International Conference on Machine Learning. PMLR, 12524–12534.

[31] Yanghao Zhang, Wenjie Ruan, Fu Wang, and Xiaowei Huang. 2020. Generalizing Universal Adversarial Attacks Beyond Additive Perturbations. In 2020 IEEE International Conference on Data Mining (ICDM). IEEE, 1412–1417.

[32] Yanghao Zhang, Fu Wang, and Wenjie Ruan. 2021. Fooling Object Detectors: Adversarial Attacks by Half-Neighbor Masks. arXiv preprint arXiv:2107.00999 (2021).

[33] Xingyu Zhao, Alex Banks, James Sharp, Valentin Robu, David Flynn, Michael Fisher, and Xiaowei Huang. 2020. A Safety Framework for Critical Systems Utilizing Deep Verification. In Conference onix Verifications. Springer, 296–319.

[34] Xingyu Zhao, Wei Huang, Alex Banks, Victoria Cox, David Flynn, Sven Schewe, and Xiaowei Huang. 2021. Assessing the Reliability of Deep Learning Classifiers Through Robustness Evaluation and Operational Profiles. In Asafety.