# Safety and Robustness for Deep Learning with Provable Guarantees

Marta Kwiatkowska[*]
marta.kwiatkowska@cs.ox.ac.uk
Department of Computer Science, University of Oxford
Oxford

## ABSTRACT

Computing systems are becoming ever more complex, with decisions increasingly often based on deep learning components. A wide variety of applications are being developed, many of them safety-critical, such as self-driving cars and medical diagnosis. Since deep learning is unstable with respect to adversarial perturbations, there is a need for rigorous software development methodologies that encompass machine learning components. This lecture will describe progress with developing automated verification and testing techniques for deep neural networks to ensure safety and robustness of their decisions with respect to bounded input perturbations. The techniques exploit Lipschitz continuity of the networks and aim to approximate, for a given set of inputs, the reachable set of network outputs in terms of lower and upper bounds, in anytime manner, with provable guarantees. We develop novel algorithms based on feature-guided search, games, global optimisation and Bayesian methods, and evaluate them on state-of-the-art networks. The lecture will conclude with an overview of the challenges in this field.

## CCS CONCEPTS

• **Theory of computation → Logic and verification**; • **Computing methodologies → Neural networks**.

## KEYWORDS

Neural networks, robustness, formal verification, Bayesian neural networks

**ACM Reference Format:**
Marta Kwiatkowska. 2020. Safety and Robustness for Deep Learning with Provable Guarantees. In *35th IEEE/ACM International Conference on Automated Software Engineering (ASE '20), September 21–25, 2020, Virtual Event, Australia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3324884.3418901

## 1 INTRODUCTION

Much of the recent success of Artificial Intelligence (AI) derives from deep learning [10]. Deep neural networks (DNNs) have been developed for a variety of tasks, including computer vision, face recognition, autonomous driving, malware detection, speech recognition, text analysis and medical diagnosis. Unfortunately, neural networks are susceptible to *adversarial examples* [2, 23]. An adversarial example is an input which, though initially classified correctly, is misclassified after a minor, perhaps imperceptible, perturbation. As an illustrative example, Figure 1 from [29] shows an image of a traffic light correctly classified by a convolutional neural network, which is then misclassified after changing only a few pixels. This is an example of a sensitivity-based adversarial example [24], typically achieved by small norm-bounded perturbations of the original input (here Euclidean distance 0.88), which a human observer would still classify correctly. Another important class of adversarial examples are invariance-based, which preserve the network's prediction while making semantic changes, but are not necessarily close to the original input with respect to the distance function. Adversarial examples have now been demonstrated for virtually all applications of deep learning, and the ease with which they can be exhibited highlights the need for appropriate safety mechanisms to use during deployment, as well as software development frameworks to ensure the safety and robustness of neural networks, as argued in [12].

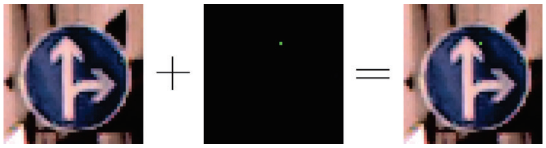

**Figure 1: An adversarial example for a neural network trained on the GTSRB dataset. After a slight perturbation, the image classification changes from "go right or straight" to "go left or straight".**

This paper describes recent progress with developing automated verification and testing techniques for deep neural networks, with the overall goal of improving the safety and robustness of their decisions. The overview focuses on the work related to the ASE 2020 keynote and is not intended to be exhaustive.

## 2 ROBUSTNESS ASSURANCE FOR NEURAL NETWORKS

This overview is concerned with robustness (or resilience) of neural networks to norm-bounded adversarial perturbations. Since DNN

[*]This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 834115) and the EPSRC Programme Grant on Mobile Autonomy (EP/M019918/1).

*ASE '20, September 21–25, 2020, Virtual Event, Australia*
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6768-4/20/09…$15.00
https://doi.org/10.1145/3324884.3418901

components are used in automated decision making, robustness translates to safety of DNN decisions in adversarial settings, where adversary's power is constrained to perturbations of up to some magnitude. We consider classification problems and work with local (also called pointwise) robustness, defined with respect to an input point and its neighbourhood as the invariance of the classification over the neighbourhood. Robustness can be measured in terms of the distance computed in the input vector space from a given input to the nearest adversarial example (called the *maximum safe radius*). Global robustness can then be estimated for the model in terms of the expectation of local robustness over the test dataset weighted by the input distribution.

## 2.1 Diagnostic search for adversarial examples

Searching for adversarial examples in the local neighbourhood of an input point can be utilised for diagnostic testing purposes. Adversarial examples can also used in an abstraction refinement framework to improve scalability of the DNN verification task [8]. Adversarial examples are typically detected by transforming the search into an optimisation problem, originally introduced for the $L_2$ distance in [23] and extended to other norms, including $L_0$ and $L_\infty$, in [6]. In contrast to these white-box approaches, which assume access to the network parameters, [25] presents a black-box method which uses the SIFT [15] algorithm to extract features of the input image, and then, working on a mixture of Gaussian representation of the image, employs Monte Carlo Tree Search to find adversarial examples. An effective $L_0$ adversarial search method is given in [26] for 3D learning by occluding points in a Lidar image, for both pointset and volumetric representations, and evaluated on the KITTI dataset. The method can be extended to perturbations other than occlusion.

## 2.2 Automated verification with provable guarantees

While searching for adversarial examples can pinpoint DNN's instabilities, it is unable to provide guarantees that no adversarial example exists if not found. The goal of automated verification approaches is to provide formal, *provable guarantees* on the robustness of DNNs. These include exact methods such as [14], which encode a ReLU network as a set of constraints and reduce the verification to the satisfiability problem, as well techniques based on abstract interpretation [19] and robust optimisation [18]. An alternative approach based on satisfiability solving [12] develops a verification framework that employs discretisation of the neighbourhood of a given input and a layer-by-layer refinement to enable exhaustive exploration. [20] presents a verification approach based on computing the reachable set of outputs using global optimisation, relying on the knowledge of an estimate of the Lipschitz constant of the network. In [29], an approximate game-based approach is developed, which enables anytime computation of upper and lower bounds of the maximum safe radius for a given input and perturbation magnitude, providing a theoretical guarantee that it can reach the exact value of the maximum safe radius. A lower bound guarantees that all perturbations up to that magnitude will not result in a class change, whereas an upper bound, found by searching for an adversarial example, provides evidence that there exist adversarial examples of that magnitude. The method works by 'gridding' the input space based on the Lipschitz constant and relies on the fact that it suffices to check only the 'corners' of the grid. Lower bound computation employs a variant of A$^\star$ search. The game-based approach has also been adapted [28] to video inputs and recurrent networks to provide robustness guarantees against perturbations of optical flow, for example for naturally plausible distortions such as camera occlusion. It is a versatile method, which can be adapted to other settings.

## 2.3 Quantifying robustness of neural networks

The classical approach to measuring robustness is by means of generalisation bounds, that is, computing theoretical upper bounds on the test error [13], usually with respect to the input data distribution. DNNs are known to generalise well. Though the accuracy of the generalisation bounds can be questioned, tight PAC-Bayes generalisation bounds has been obtained in [7]. In [16], PAC-Bayes generalization bounds are provided for models trained with data augmentation and it is shown that, compared to data augmentation, feature averaging reduces generalization error when used with convex losses and tightens PAC-Bayes bounds. In [31] it is shown that by employing causal inference one can improve generalisation bounds of reinforcement learning for both linear and non-linear problems. The feasibility of robust learning from the perspective of computational learning theory is studied in [11], where it is shown that no non-trivial concept class can be robustly learned in the distribution-free setting against an adversary who can perturb just a single input bit.

Robustness with respect to norm-bounded adversarial perturbations can be measured in terms of the maximum safe radius. The game-based approach of [29] enables such an approximation, also for the $L_0$ norm. By computing the expectation of the maximum safe radius over a test dataset, through iteratively generating lower and upper bounds on the network's robustness for the (non-differentiable) $L_0$ norm using a tensor implementation, global robustness of a network can be measured [21].

Since verification for neural networks is NP-hard, test generation and evaluation methods that ensure high levels of coverage have also been developed [22].

## 2.4 Probabilistic guarantees for deep learning

The approaches listed above pertain to deterministic neural networks. Since machine learning models encode probabilistic dependence on training data, they lend themselves to frameworks for computing *probabilistic guarantees* on their robustness. For Gaussian process (GP) models, adversarial probabilistic robustness guarantees have been studied in [5] for regression and classification [3]. Similar approaches have also been developed for Bayesian neural networks (BNNs), defined as neural networks with distributions over their weights, which can capture the uncertainty within the learning model [17]. A BNN model thus returns an uncertainty estimate [9] along with the output, which is important for safety-critical applications. In [4], *probabilistic robustness* is considered for BNNs, using a probabilistic generalisation of the usual statement of (deterministic) robustness to adversarial examples [12], namely the computation of the probability (induced by the distribution over the

BNN weights) of the classification being invariant over the neighbourhood around a given input point. Since the computation of the posterior probability for a BNN is intractable, the method employs statistical model checking [30], based on the observation that each sample taken from the (possibly approximate) posterior weight distribution of the BNN induces a deterministic neural network. The latter can thus be analysed using existing verification techniques for deterministic networks (e.g. [12, 14, 20]). This methodology has been applied to quantify the uncertainty of a BNN autonomous driving controller for tasks such as collision avoidance, and evaluated on a range of scenarios in the Carla simulator. [1] considers reinforcement learning and leverages probabilistic model checking of Markov decision processes to produce probabilistic guarantees on safe behaviour over a finite time horizon. In [27], probabilistic safety for BNNs is studied, in the sense of computing the probability of the BNN posterior distribution that all elements of a compact set of input points are mapped to the same region in the output space.

## 3 CONCLUDING REMARKS

We have provided a brief overview of recent advances towards methodologies for safety and robustness assurance for deep learning, which draw on formal verification, optimisation, probabilistic verification and testing. Clearly, although the presented methods have been useful to quantify local and global adversarial robustness of deep neural networks and Bayesian neural networks, many challenges remain. Future work will aim for tighter integration of symbolic verification, program synthesis and probabilistic verification, in order to improve scalability, support more complex properties, and provide explanations in addition to guarantees.

## REFERENCES

[1] Edoardo Bacci and David Parker. 2020. Probabilistic Guarantees for Safe Deep Reinforcement Learning. In *Proc. 18th International Conference on Formal Modelling and Analysis of Timed Systems (FORMATS'20) (LNCS)*. Springer.

[2] Battista Biggio and Fabio Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition* 84 (2018), 317–331.

[3] Arno Blaas, Luca Laurenti, Andrea Patane, Luca Cardelli, Marta Kwiatkowska, and Stephen J. Roberts. 2020. Robustness Quantification for Classification with Gaussian Processes. In *AISTATS 2020*. 3372–3382. See arxiv:1905.11876.

[4] Luca Cardelli, Marta Kwiatkowska, Luca Laurenti, Nicola Paoletti, Andrea Patane, and Matthew Wicker. 2019. Statistical Guarantees for the Robustness of Bayesian Neural Networks. In *IJCAI 2019*. See arXiv:1809.06452.

[5] Luca Cardelli, Marta Kwiatkowska, Luca Laurenti, and Andrea Patane. 2019. Robustness guarantees for Bayesian inference with Gaussian processes. In *AAAI 2019*. 7759–7768.

[6] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE, 39–57.

[7] Gintare Karolina Dziugaite and Daniel M. Roy. 2017. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. In *UAI 2017*.

[8] Yizhak Yisrael Elboher, Justin Gottschlich, and Guy Katz. 2020. An Abstraction-Based Framework for Neural Network Verification. In *CAV 2020*, Shuvendu K. Lahiri and Chao Wang (Eds.). Springer International Publishing, 43–65.

[9] Yarin Gal. 2016. *Uncertainty in deep learning*. Ph.D. Dissertation. University of Cambridge.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

[11] Pascale Gourdeau, Varun Kanade, Marta Kwiatkowska, and James Worrell. 2019. On the Hardness of Robust Classification. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 7446–7455.

[12] Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. 2017. Safety Verification of Deep Neural Networks. In *CAV 2017*. Springer, 3–29.

[13] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. 2020. Fantastic Generalization Measures and Where to Find Them. In *ICLR 2020*.

[14] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. 2017. Reluplex: An efficient SMT solver for verifying deep neural networks. In *CAV*. Springer, 97–117.

[15] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[16] Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. 2020. On the Benefits of Invariance in Neural Networks. *CoRR* abs/2005.00178 (2020).

[17] David JC MacKay. 1992. A practical Bayesian framework for backpropagation networks. *Neural computation* 4, 3 (1992), 448–472.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv e-prints* (June 2017). arXiv:1706.06083 [stat.ML]

[19] Matthew Mirman, Timon Gehr, and Martin Vechev. 2018. Differentiable Abstract Interpretation for Provably Robust Neural Networks. In *ICML 2018*. 3578–3586.

[20] Wenjie Ruan, Xiaowei Huang, and Marta Kwiatkowska. 2018. Reachability analysis of deep neural networks with provable guarantees. In *IJCAI 2018*. AAAI Press, 2651–2659.

[21] Wenjie Ruan, Min Wu, Youcheng Sun, Xiaowei Huang, Daniel Kroening, and Marta Kwiatkowska. 2019. Global Robustness Evaluation of Deep Neural Networks with Provable Guarantees for the Hamming Distance. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 5944–5952.

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

[23] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *ICLR*.

[24] Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jörn-Henrik Jacobsen. 2020. Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations. In *ICML 2020*. See arxiv.org/abs/2002.04599.

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

[26] Matthew Wicker and Marta Kwiatkowska. 2019. Robustness of 3D Deep Learning in an Adversarial Setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[27] Matthew Wicker, Luca Laurenti, Andrea Patane, and Marta Kwiatkowska. 2020. Probabilistic Safety for Bayesian Neural Networks. In *UAI 2020*. See arxiv.org/abs/2004.10281.

[28] M. Wu and M. Kwiatkowska. 2020. Robustness Guarantees for Deep Neural Networks on Videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 308–317.

[29] Min Wu, Matthew Wicker, Wenjie Ruan, Xiaowei Huang, and Marta Kwiatkowska. 2020. A Game-Based Approximate Verification of Deep Neural Networks with Provable Guarantees. *Theoretical Computer Science* 807 (2020), 298–329.

[30] Håkan LS Younes, Marta Kwiatkowska, Gethin Norman, and David Parker. 2006. Numerical vs. statistical probabilistic model checking. *International Journal on Software Tools for Technology Transfer* 8, 3 (2006), 216–228.

[31] Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. 2020. Invariant Causal Prediction for Block MDPs. In *ICML 2020*. See arxiv.org/abs/2003.06016.