Safety Verification for Deep Neural Networks with Provable Guarantees

Marta Kwiatkowska
Department of Computer Science, University of Oxford, UK
http://http://www.cs.ox.ac.uk/marta.kwiatkowska/
marta.kwiatkowska@cs.ox.ac.uk

Abstract

Computing systems are becoming ever more complex, increasingly often incorporating deep learning components. Since deep learning is unstable with respect to adversarial perturbations, there is a need for rigorous software development methodologies that encompass machine learning. This paper describes progress with developing automated verification techniques for deep neural networks to ensure safety and robustness of their decisions with respect to input perturbations. This includes novel algorithms based on feature-guided search, games, global optimisation and Bayesian methods.

2012 ACM Subject Classification Theory of computation → Logic and verification; Computing methodologies → Neural networks

Keywords and phrases Neural networks, robustness, formal verification, Bayesian neural networks

Digital Object Identifier 10.4230/LIPIcs.CONCUR.2019.38

Category Invited paper

Funding *Marta Kwiatkowska*: 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).

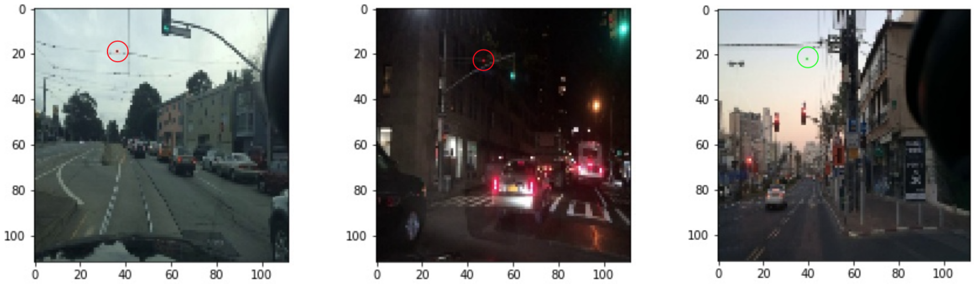
1 Introduction

Computing devices have become ubiquitous and ever present in our lives: smartphones help us stay in touch with family and friends, GPS-enabled apps offer directions literally at our fingertips, and voice-controlled assistants are now able to execute simple commands. Artificial Intelligence is making great strides, promising many more exciting applications with an increased level of autonomy, from wearable medical devices to robotic care assistants and self-driving cars.

Deep learning, in particular, is revolutionising AI. Deep neural networks (DNNs) have been developed for a variety of tasks, including computer vision, face recognition, malware detection, speech recognition and text analysis. While the accuracy of neural networks has greatly improved, they are susceptible to *adversarial examples* [17, 1]. An adversarial example is an input which, though initially classified correctly, is misclassified after a minor, perhaps imperceptible, perturbation. Figure 1 from [19] 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 illustrative example, though somewhat artificial, since in practice the controller would rely on additional sensor input when making a decision, highlights the need for appropriate mechanisms and frameworks to prevent the occurrence of similar issues during deployment.

Clearly, the excitement surrounding the potential of AI and autonomous computing technologies is well placed. Autonomous devices make decisions on their own and on users' behalf, powered by software that today often incorporates machine learning components. Since autonomous device technologies are increasingly often incorporated within safety-

30th International Conference on Concurrency Theory (CONCUR 2019).
Editors: Wan Fokkink and Rob van Glabbeek; Article No. 38; pp. 38:1–38:5
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



(a) (b) (c)

Figure 1 from [19]. Adversarial examples generated on Nexar challenge data (dashboard camera images). (a) Green light classified as red with confidence 56% after one pixel change. (b) Green light classified as red with confidence 76% after one pixel change. (c) Red light classified as green with 90% confidence after one pixel change.

critical applications, they must trustworthy. However, software faults can have disastrous consequences, potentially resulting in fatalities. Given the complexity of the scenarios and uncertainty in the environment, it is important to ensure that software incorporating machine learning components is robust and safe.

2 Overview of progress in automated verification for neural networks

Robustness (or resilience) of neural networks to adversarial perturbations is an active topic of investigation. Without claiming to be exhaustive, this paper provides a brief overview of existing research directions aimed at improving safety and robustness of neural networks. Local (also called pointwise) robustness is defined with respect to an input point and its neighbourhood as the invariance of the classification over the neighbourhood. Global robustness is usually estimated as the expectation of local robustness over the test dataset weighted by the input distribution.

2.1 Heuristic search for adversarial examples

A number of approaches have been proposed to search for adversarial examples to exhibit their lack of robustness, typically by transforming the search into an optimisation problem, albeit without providing guarantees that adversarial examples do not exist if not found. In [17], search for adversarial examples is performed by minimising the L_2 distance between the images while maintaining the misclassification. Its improvement, Fast Gradient Sign Method (FGSM), uses a cost function to direct the search along the gradient. In [5], the optimisation problem proposed in [17] is adapted to attacks based on other norms, such as L_0 and L_∞. Instead of optimisation, JSMA [13] uses a loss function to create a "saliency map" of the image, which indicates the importance of each pixel in the classification decision. [19] introduces a game-based approach for finding adversarial examples by extracting the features of the input image using the SIFT [9] method. Then, working on a mixture of Gaussians representation of the image, the two players respectively select a feature and a pixel in the feature to search for an adversarial attack. This method is able to find the adversarial example in Figure 1 in a matter of seconds.

2.2 Automated verification approaches

In contrast to heuristic search for adversarial examples, verification approaches aim to provide *formal guarantees* on the robustness of DNNs. An early verification approach [14] encodes the entire network as a set of constraints and reduces the verification to the satisfiability problem. [8] improves on [14] by by extending the approach to work with piecewise linear ReLU functions, scaling up to networks with 300 ReLU nodes. [7] develops a verification framework that employs discretisation and a layer-by-layer refinement to exhaustively explore a finite region of the vector spaces associated with the input layer or the hidden layers, and scales to work with larger networks. [15] presents a verification approach based on computing the reachable set of outputs using global optimisation. In [12], techniques based on abstract interpretation are formulated, whereas [11] employ robust optimisation.

Several approaches analyse the robustness of neural networks by considering the maximal size of the perturbation that will not cause a misclassification. For a given input point, the *maximal safe radius* is defined as the largest radius centred on that point within which no adversarial examples exist. Solution methods include encoding as a set of constraints and reduction to satisfiability or optimisation [18]. In [20], the game-based approach of [19] is extended to anytime computation of upper and lower bounds on the maximum safe radius problem, providing a theoretical guarantee that it can reach the exact value. The method works by 'gridding' the input space based on the Lipschitz constant and checking only the 'corners' of the grid. Lower bound computation employs $A^\star$ search.

Since verification for state-of-the-art neural networks is an NP problem, testing methods that ensure high levels of coverage have also been developed [16].

2.3 Towards probabilistic verification for deep neural networks

All works listed above assume a trained network with fixed weights and therefore yield deterministic robustness guarantees. Since neural networks have a natural probabilistic interpretation, they lend themselves to frameworks for computing *probabilistic guarantees* on their robustness. Bayesian neural networks (BNNs) are neural networks with distributions over their weights, which can capture the uncertainty within the learning model [10]. The neural network can thus return an uncertainty estimate (typically computed pointwise, see [6]) along with the output, which is important for safety-critical applications.

In [3], *probabilistic robustness* is considered for BNNs, using a probabilistic generalisation of the usual statement of (deterministic) robustness to adversarial examples [7], 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 [21], 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 mentioned above (e.g. [7, 8, 15]).

A related safety and robustness verification approach, which offers formal guarantees, has also been developed for Gaussian process (GP) models, for regression [4] and classification [2]. In contrast to DNNs, where trade offs between robustness and accuracy have been observed [11, 3], robustness of GPs increases with training. More research is needed to explore these phenomena.

3 Conclusion

The pace of development in Artificial Intelligence has increased sharply, stimulated by the advances and wide acceptance of the machine learning technology. Unfortunately, recent forays of technology companies into real-world applications have exposed the brittleness of deep learning. There is a danger that tacit acceptance of deep learning will lead to flawed AIs deployed in critical situations, at a considerable cost. Machine learning plays a fundamental role in enabling artificial agents, but developments so far have focused on 'narrow' AI tasks, such as computer vision and speech recognition, which lack the ability to reason about interventions, counterfactuals and 'what if' scenarios. To achieve 'strong' AI, greater emphasis is necessary on rigorous modelling and verification technologies that support such reasoning, as well as development of novel synthesis techniques that guarantee the correctness of machine learning components by construction. Importantly, automated methods that provide probabilistic guarantees which properly take account of the learning process have a role to play and need to be investigated.

References

1 Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.
2 Arno Blaas, Luca Laurenti, Andrea Patane, Luca Cardelli, Marta Kwiatkowska, and Stephen J. Roberts. Robustness quantification for classification with Gaussian processes. *CoRR abs/1905.11876*, 2019.
3 Luca Cardelli, Marta Kwiatkowska, Luca Laurenti, Nicola Paoletti, Andrea Patane, and Matthew Wicker. Statistical guarantees for the robustness of Bayesian neural networks. In *IJCAI 2019*, 2018. See arXiv:1809.06452.
4 Luca Cardelli, Marta Kwiatkowska, Luca Laurenti, and Andrea Patane. Robustness guarantees for Bayesian inference with Gaussian processes. In *AAAI 2019*, 2018. See arXiv:1809.06452.
5 Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 39–57. IEEE, 2017.
6 Yarin Gal. *Uncertainty in deep learning*. PhD thesis, University of Cambridge, 2016.
7 Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. Safety verification of deep neural networks. In *CAV*, pages 3–29. Springer, 2017.
8 Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *CAV*, pages 97–117. Springer, 2017.
9 David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
10 David JC MacKay. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.
11 A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv e-prints*, June 2017. `arXiv:1706.06083`.
12 Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *ICML 2018*, pages 3578–3586, 2018.
13 Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.
14 Luca Pulina and Armando Tacchella. An abstraction-refinement approach to verification of artificial neural networks. In *CAV*, pages 243–257. Springer, 2010.
15 Wenjie Ruan, Xiaowei Huang, and Marta Kwiatkowska. Reachability analysis of deep neural networks with provable guarantees. In *IJCAI*, pages 2651–2659. AAAI Press, 2018.
16 Youcheng Sun, Min Wu, Wenjie Ruan, Xiaowei Huang, Marta Kwiatkowska, and Daniel Kroening. Concolic testing for deep neural networks. In *ASE 2018*, pages 109–119, 2018.

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

18 Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. *CoRR abs/1711.07356*, 2017.

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

20 Min Wu, Matthew Wicker, Wenjie Ruan, Xiaowei Huang, and Marta Kwiatkowska. A game-based approximate verification of deep neural networks with provable guarantees. *Theoretical Computer Science*, 2018. To appear. See arxiv abs/1807.03571.

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