Verification of Neural Networks: Challenges and Perspectives in the AIDOaRt Project

Romina Eramo[1], Tiziana Fanni[2], Dario Guidotti[3,*], Laura Pandolfo[3], Luca Pulina[3] and Katiuscia Zedda[2]

[1]University of Teramo, Via R. Balzarini 1, Teramo, 64100, Italy
[2]Abinsula, Viale Umberto 64, Sassari, Italy
[3]University of Sassari, Piazza Università 21, Sassari, 07100, Italy

Abstract

Neural networks are increasingly being used for dealing with complex real-world applications. Despite their success, there are still important open issues such as their limited application in safety and security-critical contexts, wherein assurance about networks' behavior must be provided. The development of reliable neural networks for safety-critical contexts is one of the topics investigated in the AIDOaRt Project, a 3 years long H2020-ECSEL European project focusing on Artificial Intelligence augmented automation supporting modeling, coding, testing, monitoring, and continuous development of Cyber-Physical Systems. In this paper, we present an interesting safety-critical use case – related to the automotive domain – from the AIDOaRt project. In addition, we outline the challenges we are facing in bridging the gap between the scalability of state-of-the-art verification methodologies and the complexity of the neural networks best suited for the task of interest.

Keywords

Trustworthy AI, Neural Networks, Formal Verification, Automotive

1. Introduction

Neural networks (NNs) are one of the most investigated and widely used techniques in Machine Learning (ML) and have found successful application in many different domains across computer science [1, 2]. However, despite their success, they still find limited application in safety and security-critical contexts, wherein assurance about networks' behavior must be provided. As an example, a specific concern about the reliability of NNs is their vulnerability to adversarial attacks [3], which are small variations of the inputs which cause unforeseeable changes in the behavior of the neural network. This kind of vulnerability is both a safety (e.g., rain droplets on the cameras causing a misclassification of a vehicle) and a security issue (e.g., minor modification applied to a vehicle or road sign by a malevolent actor [4]) as we show in Figure 1. Automated

IPS-RiCeRcA-SPIRIT 2022: 10th Italian Workshop on Planning and Scheduling, RiCeRcA Italian Workshop, and SPIRIT Workshop on Strategies, Prediction, Interaction, and Reasoning in Italy.
*Corresponding author.
reramo@unite.it (R. Eramo); tiziana.fanni@abinsula.com (T. Fanni); dguidotti@uniss.it (D. Guidotti); lpandolfo@uniss.it (L. Pandolfo); lpulina@uniss.it (L. Pulina); katiuscia.zedda@abinsula.com (K. Zedda)
0000-0002-3572-5875 (R. Eramo); 0000-0002-4301-6497 (T. Fanni); 0000-0001-8284-5266 (D. Guidotti); 0000-0002-5785-5638 (L. Pandolfo); 0000-0003-0258-3222 (L. Pulina)

Figure 1: Example of a real-life adversarial attack from [4]: small graffiti on a stop signal can easily mislead a neural network to misclassify it as a speed limit.

formal verification — see, e.g., [5] for a survey — provides an effective answer to the problem of establishing the correctness of the behavior of a NN and opens the path to their adoption in safety and security-critical applications [6].

Indeed, the development of reliable NNs for safety-critical contexts is one of the topics investigated by the team of the University of Sassari in the AIDOaRt[1] Project [7, 8], a Key Digital Technologies Joint Undertaking (KTD JU) project started on April 2021, involving 32 organizations grouped in national clusters from 7 different countries. The overall idea of the project is to efficiently support the system engineering life-cycle, from requirements to testing and deployment, including software design, coding, and verification. In particular, AIDOaRt focuses on supporting the continuous development of Cyber-Physical Systems (CPSs) via Artificial Intelligence (AI)-augmented automation. The AIDOaRt framework will enable the observation and analysis of collected data from both runtime and design-time in order to provide dedicated AI-augmented solutions, that will then be validated in concrete industrial cases involving complex CPSs in domains such as railway, automotive, restaurants, etc.

The use of AI in general, and ML in particular, is a key aspect of the AIDOaRt project. Different AI-augmented capabilities will be provided, and to this intent, the AIDOaRt AI-augmented Toolkit will be designed and developed in the context of the project in order to support additional capabilities related to different CPS development tasks. The dissemination of such techniques in a regulated industry can rapidly enable systems to decide and act in a more and more automated manner, sometimes even without direct human control. As a consequence, this demands for a responsible approach to ensure a safe and beneficial use of AI technologies. This approach has to consider both the implications of (co)decision making by machines and related ethical issues. Within AIDOaRt, one of the main challenges is ensuring that systems are designed responsibly. The integrated core framework and AI-based solutions will be applied and validated in practice in the context of the different AIDOaRt project use cases. At the end of the project, an industrial

[1]AI-augmented automation supporting modelling, coding, testing, monitoring and continuous development in Cyber-Physical Systems

uptake of AIDOaRt results via developing real complex and large-scale CPSs is expected.

In this paper, we briefly present an interesting safety-critical use case from the automotive domain and the challenges we are facing in bridging the gap between the scalability of state-of-the-art verification methodologies and the complexity of the NNs best suited for the task of interest. Section 2 briefly presents the relevant background and state of the art. In Section 3 we present the Abinsula Use Case and, finally, in Section 4 we present the challenges we are facing and our ideas on how to overcome them.

2. Background

2.1. Neural Networks

Neural Networks are machine learning models composed of interconnected computing units called neurons. In a feed-forward network, the neurons are arranged in disjointed layers and each layer is connected only with the following ones forming a direct acyclic graph (DAG). Every layer of a NN performs specific computations on the inputs it receives from the previous layers or, for the first layer of the network, from the outside. The number of neurons in a layer or, more in general, in a network is directly correlated with the complexity of the computations carried out by it and it is often used to formally define such complexity. As an example, two of the most commonly used kind of layers are *linear* and *activation* layers: linear layers apply an affine transformation $f(x) = Ax + b$ to their input tensor x, whereas activation layers apply element-wise a specific (usually non-linear) function $f(x) = \sigma(x)$ to it. Another kind of frequently used layer is the convolutional one, which has been especially successful when used in architectures applied in computer vision tasks. The idea behind the convolutional layer is to analyze the input images using a set of perception filters (also known as feature maps) which can be learned from the data.

2.2. Verification

Formal verification aims to guarantee that NNs satisfy stated input-output relations and in the last decade several verification methodologies have been proposed for different specifications and architectures [9, 10, 11, 12, 5]. In general, verification methodologies can be divided into two categories: *complete* and *incomplete*. Complete verification algorithms [13, 14, 15, 16, 17] leverage technologies like Satisfiability Modulo Theories (SMT) and Mixed Integer Linear Programming (MILP) solvers or methods like branch and bound (BnB) to provide a definitive answer on the compliance of the NN to the property of interest, at the price of a greatly increased computational complexity. Conversely, incomplete verification algorithms [18, 19, 20, 21, 22] leverage methods like abstract interpretation and bound propagation to provide an answer subject to a degree of uncertainty: in particular, when an incomplete method is unable to certify that a network satisfies the property of interest it is in doubt if the property is violated or if the precision of the verification methodology was not good enough. Whereas when an incomplete algorithm produces a positive result regarding the satisfaction of the property, it is certain that the network behavior is compliant with it. This degree of uncertainty is in exchange for a greatly reduced computational complexity. It should be noted that, in recent times, many verification

tools began to combine complete and incomplete algorithms to enhance the scalability of the firsts and the precision of the seconds.

3. Safety critical systems in the automotive domain using disruption technology

The technologies developed within the AIDOaRt project will be applied and evaluated on different 15 industrial use cases (UCs), including the one related to the automotive domain proposed by Abinsula Srl, a company based in Sassari (Italy) that provides innovative Information and Communication Technology (ICT) solutions worldwide [2].

Modern cars are connected systems that acquire inputs from the environment, thus they can be considered as CPSs. With the increment of sources of information and data, safety represents even more a critical objective and new challenges in the development process are arising. This is especially true where several stakeholders, such as hardware specialists, software developers and system designers have to work together with safety engineers to ensure a reliable and safe system. In this context, the emergence of standards, such as ISO 26262 and ISO 16505, has helped the automotive industry to focus on practice to address safety in a systematic and consistent way.

This opened the path to new research for guaranteeing safety in the automotive context. The use of AI and ML is on the rise in order to enhance the automated verification of systems applied in real safety-critical applications. However, if it is true that AI is now recognized as innovative technology, it is far from being applied in real safety-critical applications due to the lack of methodologies, for example for the predictability of the system in domains such as the automotive one.

The use case brought by Abinsula aims at enhancing the human interaction and driving experience, proposing an electronic rear-view mirror that gets data from a set of cameras and provides the rear image on a screen. Image processing and sensor fusion, more in general the AI application to sensors data processing, allow for increasing the informative content of the rear environment image providing alerts or suggestions for a safer and effective drive. The use case involves the usage of four cameras that capture 1920x1080 images. Data related to the camera include an up to 60 FPS stream and the camera ID. The exploitation of NNs in image processing is meant to replace what the human brain does in terms of recognition and processing while driving. Therefore, a fundamental step to implement a virtual rear mirror able to perform as the human brain, while guaranteeing safety, is the formal verification of the adopted NNs to ensure the predictability of the system.

The main expected improvements are related to the introduction of AI techniques both in the modeling and testing phase of the system development life-cycle. In particular, in the modeling and implementation phase, the use of AI techniques will be employed for the verification of Deep Neural Network (DNN) models/implementations.

[2] https://abinsula.com/

4. Challenges and Perspectives

Taking into account the scope of the Abinsula UC, we analyzed the state of the art regarding NNs applied in computer vision and we identified a first set of architectures that could be near to optimal for the applications of interest. In particular, we identified the YOLO [23] network architecture, which is one of the most popular learning models used for object detection. This architecture is able to process videos at 45 frames per second (fps), whereas a more optimized version manages to reach 150 fps. Moreover, it greatly outperforms the contemporary models leveraging classification and learns better generalizable representations of objects.

Regarding the verification tools, we considered the winner of the 2nd International Verification of Neural Networks Competition [24] (VNN-COMP'21) ALPHA-BETA-CROWN [25], which is a NN verifier based on an efficient bound propagation algorithm and a branch and bound methodology. It also leverages dedicated hardware (i.e., GPUs) in order to scale to relatively large convolutional networks and supports a wide range of architectures. Naturally, we also considered the runner-ups of the same competition, which present comparable performances to ALPHA-BETA-CROWN. However, through a first comparison between the benchmarks used during the VNN-COMP'21 (11 convolutional layers) and the YOLO architecture (109 convolutional layers in its last version), it immediately appears clear that the scalability of the current state-of-the-art verification tools is far from being enough to support our ideal architecture.

Therefore we identified different strategies to bridge the gap between our ideal architecture and the effective scalability of the state-of-the-art verification tools. In particular, we intend to investigate pruning and/or quantization as means to produce smaller network models which should enable verification without a significant loss in performances [26]. Furthermore, we believe that it could be possible to focus on subsets of the network architecture which are of particular interest for the task at hand and, at the same time, small enough to be feasible to verify [27]. Finally, we are evaluating how to enhance existing verification tools and methodologies to support network architectures similar to YOLO.

Acknowledgments

This research work has received funding through the AIDOaRt project from the ECSEL Joint Undertaking (JU) under grant agreement No 101007350. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Sweden, Austria, Czech Republic, Finland, France, Italy, and Spain. The research of Luca Pulina has been partially funded by the University of Sassari, research fund FAR2020PULINAL.

References

[1] E. Giunchiglia, A. Nemchenko, M. van der Schaar, RNN-SURV: A deep recurrent model for survival analysis, in: V. Kurková, Y. Manolopoulos, B. Hammer, L. S. Iliadis, I. Maglogiannis (Eds.), Artificial Neural Networks and Machine Learning - ICANN 2018 - 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III, volume 11141 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 23–32.

[2] Y. LeCun, Y. Bengio, G. E. Hinton, Deep learning, Nat. 521 (2015) 436–444.
[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: Y. Bengio, Y. LeCun (Eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014.
[4] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, Robust physical-world attacks on deep learning visual classification, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 1625–1634.
[5] F. Leofante, N. Narodytska, L. Pulina, A. Tacchella, Automated verification of neural networks: Advances, challenges and perspectives, CoRR abs/1805.09938 (2018).
[6] S. Demarchi, D. Guidotti, A. Pitto, A. Tacchella, Formal verification of neural networks: A case study about adaptive cruise control, in: I. A. Hameed, A. Hasan, S. A. Alaliyat (Eds.), Proceedings of the 36th ECMS International Conference on Modelling and Simulation, ECMS 2022, Ålesund, Norway, May 30 - June 3, 2022, European Council for Modeling and Simulation, 2022, pp. 310–316.
[7] H. Bruneliere, V. Muttillo, R. Eramo, L. Berardinelli, A. Gomez, A. Bagnato, A. Sadovykh, A. Cicchetti, Aidoart: Ai-augmented automation for devops, a model-based framework for continuous development in cyber–physical systems, Microprocessors and Microsystems 94 (2022) 104672.
[8] R. Eramo, V. Muttillo, L. Berardinelli, H. Bruneliere, A. Gomez, A. Bagnato, A. Sadovykh, A. Cicchetti, Aidoart: Ai-augmented automation for devops, a model-based framework for continuous development in cyber-physical systems, in: 2021 24th Euromicro Conference on Digital System Design (DSD), IEEE, 2021, pp. 303–310.
[9] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, X. Yi, A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability, Comput. Sci. Rev. 37 (2020) 100270.
[10] D. Guidotti, Enhancing neural networks through formal verification, in: M. Alviano, G. Greco, M. Maratea, F. Scarcello (Eds.), Discussion and Doctoral Consortium papers of AI*IA 2019 - 18th International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19-22, 2019, volume 2495 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 107–112.
[11] D. Guidotti, Verification and repair of neural networks, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, AAAI Press, 2021, pp. 15714–15715.
[12] D. Guidotti, Safety analysis of deep neural networks, in: Z. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, ijcai.org, 2021, pp. 4887–4888.
[13] G. Katz, D. A. Huang, D. Ibeling, K. Julian, C. Lazarus, R. Lim, P. Shah, S. Thakoor, H. Wu, A. Zeljic, D. L. Dill, M. J. Kochenderfer, C. W. Barrett, The marabou framework for verification and analysis of deep neural networks, in: Computer Aided Verification -

31st International Conference, CAV 2019, New York City, NY, USA, July 15-18, 2019, Proceedings, Part I, volume 11561 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 443–452.

[14] P. Henriksen, A. R. Lomuscio, Efficient neural network verification via adaptive refinement and adversarial search, in: G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, J. Lang (Eds.), ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020), volume 325 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2020, pp. 2513–2520.

[15] P. Henriksen, A. Lomuscio, DEEPSPLIT: an efficient splitting method for neural network verification via indirect effect analysis, in: Z. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, ijcai.org, 2021, pp. 2549–2555.

[16] R. Bunel, J. Lu, I. Turkaslan, P. H. S. Torr, P. Kohli, M. P. Kumar, Branch and bound for piecewise linear neural network verification, J. Mach. Learn. Res. 21 (2020) 42:1–42:39.

[17] A. D. Palma, R. Bunel, A. Desmaison, K. Dvijotham, P. Kohli, P. H. S. Torr, M. P. Kumar, Improved branch and bound for neural network verification via lagrangian decomposition, CoRR abs/2104.06718 (2021).

[18] D. Guidotti, L. Pulina, A. Tacchella, pynever: A framework for learning and verification of neural networks, in: Z. Hou, V. Ganesh (Eds.), Automated Technology for Verification and Analysis - 19th International Symposium, ATVA 2021, Gold Coast, QLD, Australia, October 18-22, 2021, Proceedings, volume 12971 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 357–363.

[19] L. Pulina, A. Tacchella, Never: a tool for artificial neural networks verification, Ann. Math. Artif. Intell. 62 (2011) 403–425.

[20] G. Singh, T. Gehr, M. Püschel, M. T. Vechev, An abstract domain for certifying neural networks, Proc. ACM Program. Lang. 3 (2019) 41:1–41:30.

[21] H. Tran, X. Yang, D. M. Lopez, P. Musau, L. V. Nguyen, W. Xiang, S. Bak, T. T. Johnson, NNV: the neural network verification tool for deep neural networks and learning-enabled cyber-physical systems, in: S. K. Lahiri, C. Wang (Eds.), Computer Aided Verification - 32nd International Conference, CAV 2020, Los Angeles, CA, USA, July 21-24, 2020, Proceedings, Part I, volume 12224 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 3–17.

[22] H. Zhang, T. Weng, P. Chen, C. Hsieh, L. Daniel, Efficient neural network robustness certification with general activation functions, in: S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, 2018, pp. 4944–4953.

[23] J. Redmon, S. K. Divvala, R. B. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 779–788.

[24] S. Bak, C. Liu, T. T. Johnson, The second international verification of neural networks competition (VNN-COMP 2021): Summary and results, CoRR abs/2109.00498 (2021).

[25] S. Wang, H. Zhang, K. Xu, X. Lin, S. Jana, C. Hsieh, J. Z. Kolter, Beta-crown: Efficient bound propagation with per-neuron split constraints for neural network robustness verification, in: M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, 2021, pp. 29909–29921.

[26] D. Guidotti, F. Leofante, L. Pulina, A. Tacchella, Verification of neural networks: Enhancing scalability through pruning, in: G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, J. Lang (Eds.), ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020), volume 325 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2020, pp. 2505–2512.

[27] D. Guidotti, F. Leofante, L. Pulina, A. Tacchella, Verification and repair of neural networks: A progress report on convolutional models, in: M. Alviano, G. Greco, F. Scarcello (Eds.), AI*IA 2019 - Advances in Artificial Intelligence - XVIIIth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19-22, 2019, Proceedings, volume 11946 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 405–417.