Formal Verification of Deep Neural Networks

(Invited Tutorial)

Nina Narodytska
VMware Research, Palo Alto, California
Email: nnarodytska@vmware.com

***stract*—Deep neural networks are among the most successful cial intelligence technologies making impact in a variety of practical applications. However, many concerns were raised about the 'magical' power of these networks. It is disturbing that we are really lacking of understanding of the decision making process behind this technology. Therefore, a natural question is whether we can trust decisions that neural networks make. One way to address this issue is to define properties that we want a neural network to satisfy. Verifying whether a neural network fulfills these properties sheds light on the properties of the function that it represents. In this tutorial, we overview several approaches to verifying neural networks properties. The first set of methods encode neural networks into Integer Linear Programs or Satisfiability Modulo Theory formulas. They come up with domain-specific algorithms to solve verification problems. The second approach is to treat the neural network as a non-linear function and to use global optimization techniques for verification. The third line of work uses abstract interpretation to certify neural networks. Finally, we consider a special class of neural networks – Binarized Neural Networks – that can be represented and analyzed using Boolean Satisfiability. We discuss how we can take advantage of the structure of neural networks in the search procedure.**

I. Introduction

ep neural networks have become ubiquitous in machine ing with applications ranging from computer vision to speech recognition and natural language processing. Neural networks demonstrate excellent performance on many practical problems, often beating specialized algorithms for these problems, which led to their rapid adoption in industrial applications. With such a wide adoption, important questions arise regarding our understanding of the decision making process of these neural networks: Is there a way to analyze deep neural networks? How robust are these networks to perturbations of inputs? Recently, a new line of research on understanding neural networks has emerged that looks into a wide range of such questions, from interpretability of neural networks to verifying their properties [1], [2], [3], [4], [5], [6], [7], [8].

One emerging technique to analyze a neural network is ed on formal verification. The idea is to encode the network and the property we aim to verify as a formal statement, using ILP, SMT or SAT, for example. If the encoding provides an exact representation of the network then we can study any property related to this network, e.g. how sensitive the network is to perturbations of the input.

In this tutorial, we look at main trends in verification of deep learning networks.

- We recap basic neural networks concepts and discuss a set of interesting properties of neural network, including properties that relate inputs and outputs of the network, e.g. robustness and invertibility, and properties that relate two networks, like network equivalence.
- We discuss common encodings of deep neural networks as Boolean, SMT or ILP formulas. We will consider how various NN properties that can be represented in these formalisms.
- We survey the main methods developed in neural networks verification. We start with a group of methods that use SMT or ILP solvers to encode verification problems. These methods range from methods that use only one technology to solve the problem to methods that combine SMT and ILP techniques during the search process. Then we will look into methods that treat neural networks as non-linear functions and use global optimization techniques to perform verification. Finally, we consider the line of work that uses abstract interpretation to certify neural networks.
- We consider a special class of neural networks – Binarized Neural Networks. These networks have a number of important features that are useful in resource constrained environments, like embedded devices. We discuss how binarized neural networks can be represented as Boolean formulas. We show that structural properties of binarized neural networks can be exploited to reason about this class of networks.

References

[1] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," *CoRR*, vol. abs/1704.05796, 2017.
[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014.
[3] L. Pulina and A. Tacchella, "An abstraction-refinement approach to verification of artificial neural networks," in *CAV*, 2010, pp. 243–257.
[4] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu, "Safety Verification of Deep Neural Networks," in *CAV'17*, ser. Lecture Notes in Computer Science. Springer, 2017, pp. 3–29.
[5] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks," in *CAV'17*, 2017, pp. 97–117.
[6] C. Cheng, G. Nührenberg, and H. Ruess, "Verification of binarized neural networks," *CoRR*, vol. abs/1710.03107, 2017.
[7] N. Narodytska, S. P. Kasiviswanathan, L. Ryzhyk, M. Sagiv, and T. Walsh, "Verifying properties of binarized deep neural networks," *CoRR*, vol. abs/1709.06662, 2017.
[8] F. Leofante, N. Narodytska, L. Pulina, and A. Tacchella, "Automated verification of neural networks: Advances, challenges and perspectives," *CoRR*, vol. abs/1805.09938, 2018.