Specification and Verification of Properties of Neural Networks

Serge Haddad

LSV, ENS Paris-Saclay, CNRS, Inria, France
haddad@lsv.fr

1 Context

With the development of machine learning and its daily applications, gaining confidence in the systems produced by such techniques has become a critical issue. A first problem consists in formalizing what is expected from the systems. such requirements may be either generic or specific to the task to be achieved. For instance, adversarial robustness is a generic property [1]. It measures how much information is needed by an attacker to "falsify" the answer of a classifying system. On the other hand, assume the system proposes actions to be performed in the presence of an intruder, a specific property would be that there is no action to be proposed when no intruder is detected.

In the internship, we will focus on neural networks since this is the most widely used and moreover it presents similar features to hybrid systems letting the possibility to adapt efficient techniques from this domain. Let us illustrate an example of specification formula:

$$\forall \mathbf{x}, \mathbf{y}\ \mathit{Pre}(\mathbf{x}) \wedge \mathit{InOut}(\mathbf{x}, \mathbf{y}) \Rightarrow \mathit{Post}(\mathbf{y})$$

where $\mathbf{x}$ (resp. $\mathbf{y}$) is the input (resp. output) vector of the system, Pre is a precondition on the inputs and Post is a postcondition on the outputs. Thus checking the negation of a formula consists in solving some existential first-order theory [3].

The design of verification for neural networks is a challenging issue since the number of neurons generally is between few hundreds and millions (see [2] for a comparative study for piecewise linear neural networks). The techniques are either sound and complete [6] or can proceed via astraction [5] thus rising the issue of incompleteness and how to tackle with it. There are now software framework dedicated to the verification of deep neural networks [4].

2 Goals

Thus the goals of this internship are twofold:

- Specifying a language or a logic that can express the main properties expected to be satisfied by neural networks;
- Identifying specificities of formula related to these properties in order to design new exact and/or approximate algorithms for verifying these properties.

References

1. Bastani, O., Ioannou, Y., Lampropoulos, L., Vytiniotis, D., Nori, A.V., Criminisi, A.: Measuring neural net robustness with constraints. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. pp. 2613–2621 (2016), http://papers.nips.cc/paper/6339-measuring-neural-net-robustness-with-constraints
2. Bunel, R., Turkaslan, I., Torr, P.H.S., Kohli, P., Kumar, M.P.: Piecewise linear neural network verification: A comparative study. CoRR **abs/1711.00455** (2017), http://arxiv.org/abs/1711.00455
3. Katz, G., Barrett, C.W., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient SMT solver for verifying deep neural networks. In: Majumdar, R., Kuncak, V. (eds.) Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I. Lecture Notes in Computer Science, vol. 10426, pp. 97–117. Springer (2017). https://doi.org/10.1007/978-3-319-63387-9_5, https://doi.org/10.1007/978-3-319-63387-9_5
4. Katz, G., Huang, D.A., Ibeling, D., Julian, K., Lazarus, C., Lim, R., Shah, P., Thakoor, S., Wu, H., Zeljic, A., Dill, D.L., Kochenderfer, M.J., Barrett, C.W.: The marabou framework for verification and analysis of deep neural networks. In: Dillig, I., Tasiran, S. (eds.) Computer Aided Verification - 31st International Conference, CAV 2019, New York City, NY, USA, July 15-18, 2019, Proceedings, Part I. Lecture Notes in Computer Science, vol. 11561, pp. 443–452. Springer (2019). https://doi.org/10.1007/978-3-030-25540-4_26, https://doi.org/10.1007/978-3-030-25540-4_26
5. Mirman, M., Gehr, T., Vechev, M.T.: Differentiable abstract interpretation for provably robust neural networks. In: Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018. Proceedings of Machine Learning Research, vol. 80, pp. 3575–3583. PMLR (2018), http://proceedings.mlr.press/v80/mirman18b.html
6. Tjeng, V., Tedrake, R.: Verifying neural networks with mixed integer programming. CoRR **abs/1711.07356** (2017), http://arxiv.org/abs/1711.07356