Property-Directed Verification and Robustness Certification of Recurrent Neural Networks

Xuan Xie[1]

[1]*Max Planck Institute for Software Systems*

Abstract

Recurrent neural networks (RNNs) are a state-of-the-art tool to represent and learn sequence-based models. They are increasingly used in safety-critical applications and act, for example, as controllers in cyber-physical systems. Thus, there is a growing need for giving formal correctness guarantee to the system. However, research in this domain is only at the beginning. While formal-methods based techniques, such as model checking, have been successfully used in traditional software and reached a certain level of industrial acceptance, a transfer to machine-learning algorithms has yet to take place.

To be more specific, we are faced with the problem of *neural network verification*. Given a property φ and a neural network, neural network verification tries to prove φ holds for the network. There is a considerable amount of work focusing on verifying feed-forward neural networks and convolutional neural networks, however, verification of recurrent neural networks, which is central to natural language processing domain, remains largely untouched.

In this presentation, I will present a **property-directed approach to verifying recurrent neural networks (PDV)** [1]. To this end, we learn a deterministic finite automaton as a surrogate model from a given RNN using *active automata learning*. This model then is analyzed using *model checking* as a verification technique. The term *property-directed* reflects the idea that our procedure is guided and controlled by the given property rather than performing the two steps separately.

For experiment, I will show the comparisons of PDV against statistical model checking and automaton abstraction model checking on three applications, which are synthetic specifications, adversarial robustness certification and identifying contact sequences in contact tracing. The experimental results demonstrate the effectiveness and efficiency of our technique.

This is joint work with Igor Khmelnitsky, Daniel Neider, Rajarshi Roy, Benoît Barbot, Benedikt Bollig, Alain Finkel, Serge Haddad, Martin Leucker and Lina Ye. It has been accepted by ATVA 2021 [1].

References

[1] I. Khmelnitsky, D. Neider, R. Roy, X. Xie, B. Barbot, B. Bollig, A. Finkel, S. Haddad, M. Leucker, L. Ye, Property-directed verification and robustness certification of recurrent neural networks, in: International Symposium on Automated Technology for Verification and Analysis, 2021.

LWDA'21: Lernen, Wissen, Daten, Analysen September 01–03, 2021, Munich, Germany

xuanxie@mpi-sws.org (X. Xie)