Portfolio Solver for Verifying Binarized Neural Networks

Gergely Kovásznai[a], Gajdár Krisztián[b], Nina Narodytska[c]

[a]Eszterházy Károly University
kovasznai.gergely@uni-eszterhazy.hu

[b]Eszterházy Károly University
krisztian.gajdar@gmail.com

[c]VMware Research
n.narodytska@gmail.com

Abstract

Although deep learning is a very successful AI technology, many concerns have been raised about to what extent the decisions making process of deep neural networks can be trusted. Verifying of properties of neural networks such as adversarial robustness and network equivalence sheds light on the trustiness of such systems. We focus on an important family of deep neural networks, the Binarized Neural Networks (BNNs) that are useful in resource-constrained environments, like embedded devices. We introduce our solver VerBiNe that is able to encode BNN properties for SAT, SMT and MIP solvers and run them in parallel, in a portfolio setting. Our experimental results demonstrate that VerBiNe is capable to verify adversarial robustness of medium-sized BNNs in reasonable time and seems to scale for larger BNNs. We also report on experiments on network equivalence with promising results.

Keywords: deep learning, binarized neural network, adversarial robustness, network equivalence, SAT, SMT, MIP, Boolean cardinality constraint, pseudo-Boolean constraint.