Abstract Neural Networks

Matthew Sotoudeh and Aditya V. Thakur

University of California, Davis, USA
{masotoudeh,avthakur}@ucdavis.edu

Abstract. Deep Neural Networks (DNNs) are rapidly being applied to safety-critical domains such as drone and airplane control, motivating techniques for verifying the safety of their behavior. Unfortunately, DNN verification is NP-hard, with current algorithms slowing exponentially with the number of nodes in the DNN. This paper introduces the notion of Abstract Neural Networks (ANNs), which can be used to soundly overapproximate DNNs while using fewer nodes. An ANN is like a DNN except weight matrices are replaced by values in a given abstract domain. We present a framework parameterized by the abstract domain and activation functions used in the DNN that can be used to construct a corresponding ANN. We present necessary and sufficient conditions on the DNN activation functions for the constructed ANN to soundly over-approximate the given DNN. Prior work on DNN abstraction was restricted to the interval domain and ReLU activation function. Our framework can be instantiated with other abstract domains such as octagons and polyhedra, as well as other activation functions such as Leaky ReLU, Sigmoid, and Hyperbolic Tangent.
Code: https://github.com/95616ARG/abstract_neural_networks

Keywords: Deep Neural Networks · Abstraction · Soundness.

1 Introduction

Deep Neural Networks (DNNs), defined formally in Section 3, are loop-free computer programs organized into *layers*, each of which computes a linear combination of the layer's inputs, then applies some *non-linear activation function* to the resulting values. The activation function used varies between networks, with popular activation functions including ReLU, Hyperbolic Tangent, and Leaky ReLU [13]. DNNs have rapidly become important in a variety of applications, including image recognition and safety-critical control systems, motivating research into the problem of verifying properties about their behavior [18,9].

Although they lack loops, the use of non-linear activation functions introduces *exponential branching behavior* into the DNN semantics. It has been shown that DNN verification is NP-hard [18]. In particular, this exponential behavior scales with the number of *nodes* in a network. DNNs in practice have very large numbers of nodes, e.g., the aircraft collision-avoidance DNN ACAS Xu [17] has 300 and a modern image recognition network has tens of thousands [20]. The