Verification of Recurrent Neural Networks for Cognitive Tasks via Reachability Analysis

Hongce Zhang[1] and **Maxwell Shinn**[2] and **Aarti Gupta**[2]
and **Arie Gurfinkel**[3] and **Nham Le**[3] and **Nina Narodytska**[4]

Abstract. Recurrent Neural Networks (RNNs) are one of the most successful neural network architectures that deals with temporal sequences, e.g. speech and text recognition problems. Recently, RNNs have been shown to be useful in cognitive neuroscience as a model of decision-making. RNNs can be trained to solve the same behavioral tasks performed by humans and other animals in decision-making experiments, allowing for a direct comparison between networks and experimental subjects. Analysis of RNNs is expected to be a simpler problem than the analysis of neural activity. However, in practice, reasoning about an RNN's behaviour is a challenging problem. In this work, we take the formal verification approach to the analysis of RNNs. We make two main contributions. First, we analyse the cognitive domain and formally define a set of useful properties to analyse for a popular experimental task. Second, we employ and adapt well-known verification techniques to our focus domain, i.e. a polytope propagation, an invariant detection and a counter-example guided abstraction refinement, to perform verification efficiently. Our experimental results show that our techniques scales better for the exponential number of polytopes compared to the state-of-the-art neural network verification tool.

[1] Princeton, USA, {hongcez, aartig}@princeton.edu
[2] Yale University, USA, maxwell.shinn@yale.edu
[3] University of Waterloo, Canada, arie.gurfinkel@uwaterloo.ca, nhamle-van@gmail.com
[4] VMware Research, USA, nnarodytska@vmware.com