TU Clausthal

Abschlussvortrag Masterarbeit Jialin Li

„Input Region Exploration Order Strategies for the Certification of Learned Artificial Neural Networks“

Over the past few years, Deep Neural Networks (DNNs) are increasingly deployed in safety-critical domains, such as medical diagnosis, aircraft collision avoidance, autonomous driving, Etc. Unfortunately, it has been found that DNNs can sometimes be unexpectedly fragile and behave erratically. For example, it classifies two very similar objects as different classes. Therefore, it becomes crucial to ensure that DNNs behave reliably and as expected. The interval-based certification approach has been proved to be an efficient method to certify the safety of DNNs.

The interval-based certification method leverages interval analysis. It propagates the input region through the DNNs to compute the corresponding estimation output. Then, it checks whether the estimation output satisfies the desired output. Because of the dependency problem, the estimation outputs is an overestimation of the actual outputs. Therefore, the certification method heavily splits the input region to obtain a more accurate estimation output. The result is that millions of sub-input regions are generated. Our study focuses on ordering the split sub-input regions to accelerate the certification progress.

We utilize three different order strategies to schedule the certification of sub-input regions. The first strategy is based on FIFO, and it prioritizes the larger sub-input regions. The second strategy is based on LIFO, prioritizing the sub-input regions with smaller volumes. The third is a mixed strategy; the impact factors determine the certification order of sub-input regions. This impact factor is defined by both the sub-input region's volume and the distance at which a sub-input region can be explicitly identified as safe/adversarial.

We evaluated the input region exploration order strategies in two experiments. The experimental results indicate that our approach is effective for DNN certification progress.

Betreuer der Arbeit:	Prof. Dr. Rüdiger Ehlers, Prof. Dr. Steffen Herbold
Datum:	Donnerstag, 24. März 2022, 14:00 Uhr
Ort:	Online-Meeting über BBB
	Link: https://webconf.tu-clausthal.de/b/sim-uc9-rvy