TU Clausthal

Abschlussvortrag Masterarbeit Ahmad Hatahet

„Enhancing Runtime Monitoring Performance Through Binary Decision Diagrams and Principal Component Analysis“

In today’s rapidly developing technology era, most applications are utilizing artificial intelligence very effortlessly, especially neural networks (NN). The deployment of any model had proven to be not a trivial task, particularly when the data used in training the NN is not representative of all situations that might occur in the real world. In addition to this problem, facing adversarial inputs like novel classes or anomalies is another major one. For these reasons, adopting a monitoring system to check the input’s integrity is highly beneficial and helpful. In this thesis we would like to address these problems by enhancing on an abstraction-based monitor that utilizes binary decision diagram (BDD), which uses a set of Boolean expressions to construct a directed acyclic graph (DAG) and arrive to one of two terminal nodes, “True” or “False”, where the former represents that the NN is familiar with the input and the latter indicates that the input is infected and an alarm therefore must be raised. We utilize the last hidden layer’s (LHL) outputs of the NN and transform these outputs to binary (0 and 1) with the help of a threshold function, and we refer to them as a pattern. We experiment with various thresholds and multiple length of the LHL. However, the main challenge is to find a generalized monitor that prevents harmful and troublesome inputs. Therefore, we utilize the principal component analysis (PCA) to help with selecting the most influential neurons to monitor. Moreover, we utilize another method to generate new patterns that are -to some degree- alike the original ones, and then see if this addition can improve the generalization of the monitor. Our results shows that our monitor is resilient to unseen data yet significantly elastic to accepts similar inputs to the data used in training, and finally we test it intensively to know its capabilities and limitations against multiple types of anomalies and novel classes.

Betreuer der Arbeit:	Prof. Dr. Rüdiger Ehlers, PD Dr. Christoph Knieke
Datum:	Dienstag, 21. November 2023, 10:30 Uhr
Ort:	Institut für Software and Systems Engineering Besprechungsraum 120 Arnold-Sommerfeld-Straße 1 38678 Clausthal-Zellerfeld Webkonferenz: https://webconf.tu-clausthal.de/rooms/sim-uc9-rvy/join