TU Clausthal

Abschlussvortrag Research Track Ehab Ghannoum

„Poison Source Code Detection with CodeGuardian“

Neural models for code, such as Copilot, have demonstrated outstanding capabilities in assisting developers with various software development tasks. However, these models are susceptible to adversarial examples, a phenomenon known as poison attacks. Poison attack aims to inject backdoors into deep learning models by poisoning the training data with poison samples. To defend against poison samples, we present CodeGuardian (CG), a hybrid deep-learning model designed to detect potential poison source code samples. We conducted a comparison between CG and the state-of-the-art approach, ONION, to detect poison samples generated by DAMP, MHM, ALERT, and a modified version of the TextFooler. The results obtained revealed that CG significantly outperforms ONION. Additionally, we assessed our model's performance against unknown attacks by systematically excluding one attack at a time and testing the model's performance. Our CG model demonstrated high accuracy in identifying DAMP, MHM, and ALERT.

Betreuer der Arbeit:	Prof. Dr. Mohammad Ghafari, Prof. Dr. Benjamin Säfken (Institut für Mathematik)
Datum:	Montag, 12. Februar 2024, 14:00 Uhr
Ort:	Online-Meeting über BBB
	Link: https://webconf.tu-clausthal.de/rooms/ikt-1hf-xbm-wai/join