Contribution ID: **219** Type: **Paper presentation**

Adversarial Attacks for Scripts

Thursday 18 September 2025 15:30 (15 minutes)

As the number of cyberattacks increases year by year, malware detection remains a pressing challenge, as traditional methods are no longer sufficient due to the dynamic nature of the field. Machine learning comes as an improvement over traditional approaches, offering better detection capabilities, but it still comes with two main disadvantages: a lack of interpretability and vulnerability to adversarial attacks. In this study, we examined the effect of such attacks on a malware detector based on a CharCNN model. Using Grad-CAM, we identified the most influential character regions in both clean and malicious script samples. These relevant regions were then inserted into samples of the opposite class to generate adversarial examples. Our experiments demonstrate a significant drop in detection performance: the accuracy of the CharCNN model decreased from 99.24% to 85.31% on JavaScript files and from 98.48% to 78.66% on Python files following the attacks.

Author: CHIPER, Maria (University of Bucharest)

Co-authors: Mrs STĂNESCU, Daria (National University of Science and Technology Politehnica Bucharest); Mr BECHERU, Traian (National University of Science and Technology Politehnica Bucharest); PECA, Ludmila (Universitatea Tehnică a Moldovei)

Presenter: CHIPER, Maria (University of Bucharest)

Session Classification: Doctoral Symposium

Track Classification: Doctoral Symposium