Consiglio Nazionale delle Ricerche

CNR-Istituto di **Elettronica** e di **Ingegneria** dell'**Informazione** e delle **Telecomunicazioni**

Thursday seminars

IEIIT Youth

Dr. Sara Narteni
(PhD student)

Dr. Alberto Carlevaro
(PhD student)

Countermeasures against adversarial machine learning based on eXplainable and Reliable Artificial Intelligence

Machine learning (ML) algorithms are nowadays widely adopted in different contexts to perform autonomous decisions and predictions. Due to the high volume of data shared in the recent years, ML algorithms are more accurate and reliable since training and testing phases are more precise. An important concept to analyze when defining ML algorithms concerns adversarial machine learning attacks. These attacks aim to create manipulated datasets to mislead ML algorithm decisions.

In this talk, we will present our research on new approaches able to detect and mitigate malicious adversarial machine learning attacks against a ML system. In particular, we investigate the Carlini-Wagner (CW), the fast gradient sign method (FGSM) and the Jacobian based saliency map (JSMA) attacks. The aim of the work is to exploit detection algorithms as countermeasures to these attacks. Initially, we performed some tests by using canonical ML algorithms with a hyperparameters optimization to improve metrics. Then, we adopt original reliable AI algorithms, either based on eXplainable AI (Logic Learning Machine) or Support Vector Data Description (SVDD). The obtained results show how the classical algorithms may fail to identify an adversarial attack, while the reliable AI methodologies are more prone to correctly detect a possible adversarial machine learning attack. The evaluation of the proposed methodology was carried out in terms of good balance between FPR and FNR on real world application datasets: Domain Name System (DNS) tunneling, Vehicle Platooning and Remaining Useful Life (RUL). In addition, a statistical analysis was performed to improve the robustness of the trained models, including evaluating their performance in terms of runtime and memory consumption.

Registration form

Teams Webinar • 13 April 2023 – 17:30

CNR