ID: **P3.6-224** Type: **e-Poster**

the vulnerabilities of machine learning systems in adversarial settings

Thursday, 1 July 2021 09:00 (15 minutes)

Machine learning has advanced radically over the past 10 years, and machine learning algorithms now achieve human-level performance or better on a number of tasks. Machine learning techniques have been extensively deployed for a variety of applications in different areas of life. The success of machine learning algorithms has led to an explosion in demand.
Machine learning models are also subject to attacks at both training and testing phases. Attackers can break current machine learning systems, such as by poisoning the data used by the learning algorithm or crafting adversarial examples to directly force models to make erroneous predictions
The main threat during testing is evasion attack, in which the attacker subtly operates by making small perturbations to the test set and modifies input data so that a human observer would perceive the original content but the model generates different outputs. Such inputs, known as adversarial examples, has been used to attack voice interfaces, face-recognition systems, image and video and text-classifiers.
This presentation will explain adversarial attacks examples in current machine learning models and its future trends as well as answering what can be done to defend models against adversarial manipulation.

Promotional text

Attacking Real-World Machine Learning Systems,
Understand Machine Learning security,
Adversial ML,
Data poisoning

Primary author: Mr SERRHINI, Mohamed (Mohamed First University, Oujda, Morocco)

Presenter: Mr SERRHINI, Mohamed (Mohamed First University, Oujda, Morocco)

Session Classification: T3.6 e-poster session

Track Classification: Theme 3. Verification Technologies and Technique Application: T3.6 - Artificial Intelligence and Machine Learning