UNIVERSITÀ
DI SIENA
1240

DIPARTIMENTO
INGEGNERIA DELL'INFORMAZIONE E
SCIENZE MATEMATICHE

SCHEDA ATTIVITÀ

INCARICO DI LAVORO AUTONOMO

Titolo del progetto	*Analysis of the vulnerability of AI-based classifiers against adversarial attacks*
Soggetto proponente	Mauro Barni
Obiettivi e finalità	1. Analysis of the data distribution of some popular image datasets used in AI applications (MNIST, CIFAR10, Food101) and other datasets identified during the activity 2. Identification and implementation of suitable algorithms to analyse the distribution of the datasets in point 1, by computing the Minimum Volume Enclosing Ellipsoids (MVEE) of the images at different resolutions 3. Exploitation of the MVEE analysis to explain the emergence of adversarial examples based on the concentration of measure phenomenon 4. Empirical validation of the analysis in point 4. on a pool of DNN-based classifiers trained on the datasets in point 1.
Responsabili delle attività di progetto	Mauro Barni
Durata dell'incarico	12 mesi
Requisiti/competenze richieste	PhD on themes related to the objective of the contract. Research experience in AI. Knowledge of Python programming language.
Descrizione dell'attività complessiva di	The goal of the research is to analyze the data distribution of widely used image datasets in AI -

progetto	like MNIST, CIFAR-10, Food101 and possibly others - to understand structural and statistical properties that may influence model robustness. The focus will be on assessing how data geometry and high-dimensional structure contribute to the emergence of adversarial examples. In a second phase the findings of the analysis will be evaluated by the light of existing theoretical work on the concentration of measure phenomenon, which suggests why, in high dimensions, small perturbations can significantly alter model predictions. Understanding how these theoretical insights manifest in real-world datasets can help identify intrinsic vulnerabilities in current AI models and guide the design of more robust learning systems. Eventually, the validity of the results of the theoretical analysis will be assessed experimentally, on a pool of classifiers trained on the datasets used for the analysis.

Il Proponente

Il Responsabile del Progetto