AMMINISTRAZIONE CENTRALE
AREA COMUNICAZIONE E MARKETING
SETTORE **UFFICIO STAMPA**
Via VIII febbraio, 2 – 35122 Padova
stampa@unipd.it
http://www.unipd.it/comunicati
tel. 049-8273066/3041

Padova, 11 giugno 2020

UNIVERSITÀ DI PADOVA E VERONA UNICI ITALIANI TRA I VINCITORI DEL FACEBOOK AWARDS 2020

Il premio "Probability and Programming research awards" di Facebook inc. è conferito alle migliori idee di ricerca in ambito Intelligenza Artificiale (AI) e Programmazione.

L'idea premiata, dal titolo "Adversarial Machine Learning by Morphological Abstract Interpretation", riguarda il problema dell'adversarial machine learning. L'adversarial machine learning è un argomento oggigiorno molto caldo che riguarda le vulnerabilità dei sistemi di Intellilgenza Artificiale, in scenari dove (per ragioni naturali o per specifici attacchi malevoli) questi stessi algoritmi non riescono a classificare correttamente le informazioni.

Esempi classici sono gli errori dei sistemi di AI nel riconoscere oggetti, a volte causando incidenti anche catastrofici, ad esempio nei veicoli a guida completamente autonoma.

Francesco Ranzato

«Scopo della nostra ricerca – **spiega il prof Francesco Ranzato docente di Informatica del dipartimento di Matematica dell'Università di Padova** - è evitare queste situazioni mediante tecniche formali, ovvero matematicamente dimostrate, che riescano a garantire la corretta classificazione entro specifici intervalli di tolleranza. Le attuali tecniche di difesa sono spesso ad hoc ed includono training sul modello con dati contraddittori, validazione degli input, e test della robustezza dei modelli. L'idea è di applicare alcune tecniche avanzate di analisi statica e verifica del software, che abbiamo già sviluppato col gruppo di ricerca di Verona guidato dal prof Roberto Giacobazzi, allo specifico problema dell'adversarial machine learning. In questo caso sfruttiamo una sorprendente quanto affascinante simmetria esistente tra i modelli utilizzati per

l’analisi del software e quelli sviluppati per il filtraggio morfologico delle immagini. Questo significa codificare le perturbazioni morfologiche dei dati che possono causare errata classificazione come astrazioni ed utilizzare tecniche di interpretazione astratta per certificare la robustezza dei sistemi di classificazione rispetto a queste stesse perturbazioni.»