CYBERSECURITY NATIONAL LABORATORY

Pattern Recognition and Applications Lab

Artificial Intelligence and Intelligent Systems

Sicurezza dell'Intelligenza Artificiale in ambito sanitario

Giorgio Giacinto
giacinto@unica.it

VI Conferenza Nazionale, 22 aprile 2024

UNIVERSITÀ DEGLI STUDI DI CAGLIARI

1

Intelligenza Artificiale

UNICA

Grandi quantità di **dati** da un numero crescente di **dispositivi** e **sensori interconnessi** sono alla base del successo dell'**Intelligenza Artificiale**

I successi dell'Intelligenza Artificiale

- **Traduzione automatica**
- **Sistemi conversazionali**
- **Riconoscimento e sintesi facciale**
- **Riconoscimento di oggetti**
- **Diagnosi mediche per immagini**
- **Genomica**





I limiti dell'intelligenza artificiale

The Tao of DALL-E 2 › The weird images it creates reveal the limits of AI P. 5

Toyota's Tech for an Aging World › How to help the elderly act and feel young P. 30

The Trouble With Robots in the OR › Surgeons in training don't get to practice P. 33

FOR THE TECHNOLOGY INSIDER
AUGUST 2022

IEEE Spectrum

ARTIFICIAL INTELLIGENCE

DALL-E 2's Failures Reveal the Limits of AI ›

OpenAI's text-to-image generator struggles with text, science, and bias

OpenAI has not released the technology to the public, to commercial entities, or even to the AI community at large. "We share people's concerns about misuse, and it's something that we take really seriously," OpenAI researcher Mark Chen tells *IEEE Spectrum*. But the company did invite select people to experiment with DALL-E 2. The results that have emerged over the past few months say a lot about the limits of today's deep-learning technology, giving us a window into what AI understands about the human world—and what it totally doesn't get.

Bias: DALL-E 2 is considered a multimodal AI system because it was trained on images and text, and it exhibits a form of multimodal bias. For example, if a user asks it to generate images of a CEO, a builder, or a technology journalist, it will typically return images of men, based on the image-text pairs it saw in its training data. OpenAI

August 2022

Machine Learning: la chiave del successo della IA

INSIEME DI DATI DI ESEMPIO

Training set

Dati **categorizzati** a priori

Collezione di dati correlati che rappresenta il problema che la macchina dovrà risolvere

ASPETTI CRITICI

- controllo del processo di **selezione dei dati**
- possibile **polarizzazione** dei dati
- possibili **manomissioni** nel contenuto dei dati

Machine Learning: la chiave del successo della IA

INSIEME DI DATI DI ESEMPIO → RAPPRESENTAZIONE

Training set

$$\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_d \end{pmatrix}$$

Dati **categorizzati** a priori

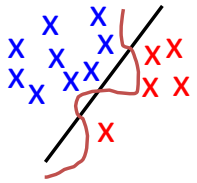
Estrazione di **misure**: distanze, frequenze…

Machine Learning: la chiave del successo della IA

INSIEME DI DATI DI ESEMPIO → RAPPRESENTAZIONE → ALGORITMO DI APPRENDIMENTO

Training set

$$\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_d \end{bmatrix}$$



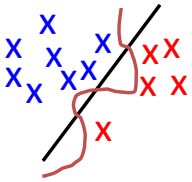
Dati **categorizzati** a priori

Estrazione di **misure**: distanze, frequenze…

individua **correlazioni** fra i dati ma non relazioni **causali**

Machine Learning: la chiave del successo della IA

ALGORITMO DI APPRENDIMENTO



L'algoritmo di apprendimento **ottimizza** una **funzione obiettivo** che definisce una **superficie di separazione** nello spazio usato per la **rappresentazione** dei dati

individua **correlazioni** fra i dati ma non relazioni **causali**

La capacità di classificare correttamente nuovi casi dipende dalle scelte di progetto

Machine Learning: Modelli Fondazionali

- **E' una categoria di sistemi**
  - addestrati con **grandi quantità di dati**
  - la maggior parte **non classificati**
    (addestramento semi-supervised)
  - basati su **deep learning** e **IA generativa**
  - possono essere **adattati** per l'utilizzo **per compiti diversi**

- **Applicazioni: traduzione linguistica, generazione contenuti audio-visivi da testo, generazione testo da richiesta testuale, ecc.**

Vulnerabilità nei sistemi di Machine Learning

Punti di forza
- Capacità di elaborare **grandi quantità di dati** da **fonti diverse**

Aree di debolezza
- Dipendenza dal processo di **selezione delle fonti e dei dati**
- **Scarsa trasparenza** e **Interpretabilità** dei modelli

Punti di forza
- Capacità di **generalizzazione** a casi nuovi rispetto a quelli usati in fase di addestramento

Aree di debolezza
- Possibilità di **manipolare** il sistema sfruttando le **correlazioni spurie** create dal modello

Intelligenza Artificiale e etica

- **Aspetti etici legati alla progettazione del sistema**
  - gestione dei **dati**, di solito molto dettagliati
  - obiettivi degli algoritmi di **ottimizzazione**

- **Aspetti etici legati all'utilizzo dalla IA in**
  - **relazioni** interpersonali
  - **relazione** fra la persona e la società
  - compiti **predittivi**

  come ad esempio in ambito educativo, sanitario, giuridico

Intelligenza Artificiale e etica

26 October 2023

The blind use of AI in healthcare can lead to invisible discrimination

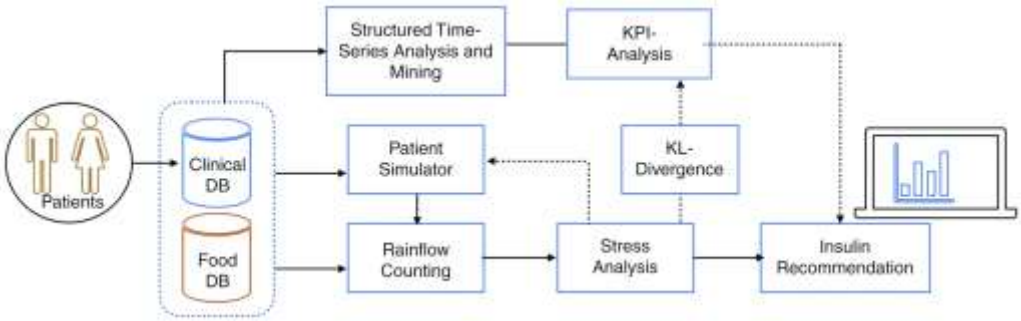
ARTIFICIAL INTELLIGENCE | HEALTH

DATA SCIENCE Artificial intelligence can help healthcare systems under pressure allocate limited resources, but also lead to more unequal access. This is demonstrated by a research collaboration between the University of Copenhagen, Rigshospitalet and DTU that investigated whether AI can spot the risk of depression equally across different population segments. The research presents options for combing algorithms for bias prior to their deployment.

Computer

Trustworthy AI-Based Personalized Insulin Recommender for Elderly People Who Have Type-2 Diabetes

Mar. 2024, pp. 35-45, vol. 57



*Direct **explanations** of the insulin recommender's performance are crucial for **trustworthiness**, ensuring that stakeholders understand actions and data.*

Adversarial Machine Learning

	Obiettivi dell'avversario		
	Errori nella risposta del sistema che non ne compromettono la funzionalità	Errori che compromettono il corretto funzionamento del sistema	Strategie di interrogazione del sistema che possono rivelare informazioni sull'addestramento del sistema o altri dati riservati
Capacità dell'avversario	**Integrità**	**Disponibilità**	**Riservatezza / Dati personali**
Interrogazioni *ad hoc* per ingannare il sistema	**Evasione** (adversarial examples)	***Sponge*** per aumentare consumo di energia o latenza nella risposta	**Estrazione del modello** / **furto** del modello o **inversione** (hill-climbing attacks)
Manipolazione dei dati usati in fase di addestramento	**Avvelenamento** per rendere possibili azioni intrusive in fase di utilizzo, ad es., *backdoors* o *trojans*	**Avvelenamento** finalizzato ad aumentare il numero di errori del sistema	**Avvelenamento** finalizzato a consentire, in fase di utilizzo del sistema, l'estrazione di dati usati in fase di addestramento

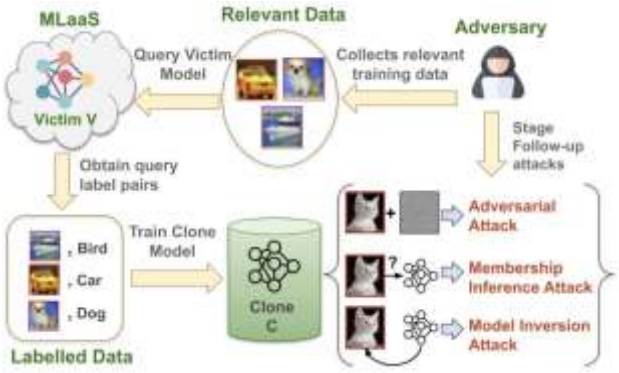
Adattamento da: Biggio B., Roli F., "Wild patterns: Ten years after the rise of adversarial machine learning", 2018

Furto del modello dei dati

Realizzare un modello con ottime prestazioni è **molto costoso**

- acquisizione dati, validazione e classificazione
- scelta del modello, ottimizzazione ecc.

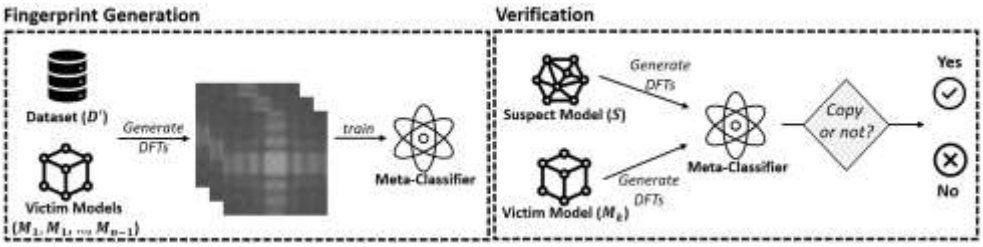
Rischio: furto del modello o furto dei dati usati per addestrare il modello attraverso meccanismi di inferenza.



S. Sanyal et al. "Towards Data-Free Model Stealing in a Hard Label Setting," *CVPR 2022*

Contromisure

- Meccanismi di ***watermarking*** o ***fingerprinting***
- Il modello deve essere perturbato in modo da poter rilevare usi illeciti
- Possibile effetto collaterale: degrado delle prestazioni



Esempio: S. Park et al., *Fingerprinting to Identify Proprietary Dataset Use in Deep Neural Networks* (ACSAC 2023)

Sfide

- **Creare un dataset e un sistema affidabile è una operazione molto costosa**
  - Rischi ridotti in caso di **applicazioni specifiche**
- **I dati rivelano sempre qualcosa dell'individuo**
  - Necessità di meccanismi per la **condivisione** di dati che garantiscano la **riservatezza** individuale
- **I modelli di machine learning *non sono super partes***
  - Necessità di **rendere esplicito il *modello*** della realtà usato per la formulazione dell'apprendimento
- **I modelli sono vulnerabili a diverse tipologie di attacchi**
  - Necessità di meccanismi di **gestione e riduzione del rischio**