indigo.ai

# Parole e immagini, l'intelligenza artificiale si avvicina all'uomo

*Insegnare ai modelli di linguaggio più complessi (come Gpt-3) a "capire" le immagini si può: con un nuovo approccio che accoppia testi e fotografie*

*A cura di Matteo Muffo, AI Researcher di Indigo.ai*

L'intelligenza artificiale continua a crescere e migliorarsi, ma la capacità delle macchine di ragionare e pensare in maniera autonoma è ancora lontana. Lo dimostrano gli enormi limiti di Gpt-3, l'ultimo algoritmo di OpenAi e Microsoft, che è un modello senza precedenti capace di immagazzinare *centinaia di miliardi* di parole: può interpretare e scrivere in maniera chiara qualunque cosa, ma crolla alla prova dei fatti quando si esce dal campo della parola scritta. Per metterlo in difficoltà basta chiedergli **di che colore siano le pecore**: il sistema risponderà nere con la stesse frequenza con cui dirà bianche. Il motivo è semplice: ha imparato a dire "pecora nera", perché questo colore ricorre sul web altrettante volte di "bianco" in relazione alla parola "pecora". Ma non ha capito il senso dell'espressione. Un errore banale che dimostra da un lato che le potenzialità sono enormi e quanto rapidamente evolva la tecnologia, ma dall'altro anche quanto sia importante **lo sforzo da fare in fase di addestramento** per sviluppare la capacità dell'intelligenza artificiale di ragionare in modo razionale.

La soluzione al problema l'hanno suggerita **i ricercatori dell'Università del North Carolina**, progettando **una nuova tecnica d'addestramento** per dare a Gpt-3 – ovvero una AI che si basa esclusivamente sulle parole – anche la capacità di "vedere" le cose, rafforzando così le sue possibilità di imparare. La sfida è quindi quella di combinare due diversi set di dati (testo e immagini) in un sistema unico per poter addestrare un nuovo modello da zero.

## Didascalie descrittive

L'approccio scelto è quello di compilare una **raccolta di immagini con didascalie molto più descrittive** di quelle cui siamo abituati. Per esempio, prendiamo la foto di un gatto in cucina che mangia: solitamente, con tutta probabilità, verrebbe intitolata solo "gatto"; invece, un set che combina parole e immagini la chiamerebbe "un gatto in cucina che mangia croccantini da una ciotola rossa". In questo modo, grazie alla combinazione puntuale di linguaggio testuale e visivo, possiamo **insegnare a un modello di intelligenza artificiale non solo come riconoscere gli oggetti, ma anche come si relazionano e agiscono l'uno sull'altro, attraverso l'uso di verbi e preposizioni**.

Un processo quasi banale sulla carta e per l'intelligenza umana, ma che però nella realtà richiederebbe un'eternità (se fatto dalle persone): basti pensare che se la versione inglese di Wikipedia comprende quasi 3 miliardi di parole, un set di dati visivi come, per esempio, potrebbe essere quello di Microsoft Common Objects in Context – meglio conosciuto come MS COCO – ne contiene appena 7 milioni. Combinare le due classi di dati diventa chiaramente molto difficile e impegnativo in termini di tempo.

I ricercatori americani, però, sono riusciti ad aggirare il problema con un **metodo di apprendimento supervisionato** capace di **adattare i dati in MS COCO alle dimensioni di Wikipedia**. Con il risultato di aver creato **un modello di linguaggio che supera quelli più all'avanguardia in alcuni dei test più difficili** utilizzati per valutare la comprensione del linguaggio AI. Dimostrando che se il modello oltre a imparare la parola gatto, la vede anche, sarà in grado di ragionare in maniera sempre più razionale ed efficace.

## Bert contro Gpt-3

I ricercatori hanno quindi utilizzato l'accoppiamento tra parole e immagini che hanno creato con MS COCO per addestrare il loro algoritmo e hanno riqualificato un modello di linguaggio open source sviluppato da Google, noto come Bert, che precede Gpt-3. Terminato il processo di aggiornamento, hanno sfidato il "nuovo BERT" su sei diversi test di comprensione linguistica, tra cui SQuAD, Stanford Question Answering Dataset, che chiede ai modelli di **rispondere a domande di comprensione della lettura** su una serie di articoli, e SWAG, che costringe gli algoritmi a **dimostrare di aver compreso il significato delle parole** e di non averle "semplicemente" imparate a memoria.

Un approccio così innovativo apre scenari tutti da esplorare. Certo, il vecchio Bert non potrà mai battere Gpt-3, ma la consapevolezza di poter addestrare un sistema così potente con parole e immagini ci fa capire quanto la tecnologia corra veloce e come le innovazioni siano conquiste quotidiane. Facendo un altro passo verso l'*artificial general intelligence*.

**Ufficio Stampa Indigo.ai: ddl studio**
indigo@ddlstudio.net
Mara Linda Degiovanni | +39 3496224812
Elisa Giuliana | +39 3386027361