

Intelligenza Artificiale e reti neurali: nozioni essenziali

di *Andrea Cossu*

Questo articolo introduce al funzionamento delle moderne Intelligenze Artificiali (IA), basate sul principio dell'apprendimento automatico. L'articolo si concentra su come le IA riescano a costruire la propria conoscenza a partire dai dati, sviluppando autonomamente meccanismi di riconoscimento a partire dalle regolarità presenti nei dati, senza regole esplicite fornite dagli esseri umani. Vengono poi approfondite le reti neurali artificiali, tra i più diffusi approcci di IA tramite apprendimento automatico, concentrandosi sulla loro capacità di risolvere un problema scomponendolo in sotto-problemi più semplici. Nonostante i notevoli successi applicativi, il comportamento di questi sistemi rimane spesso difficile da interpretare, poiché emerge dall'interazione di molte operazioni matematiche semplici, analogamente ai fenomeni dei sistemi complessi.

In conclusione, si sottolinea come l'IA contemporanea rappresenti un'intelligenza complementare a quella umana e come la sfida futura non sia imitare perfettamente l'uomo, ma sviluppare forme di collaborazione tra esseri umani e macchine capaci di ampliare le possibilità di entrambe queste Intelligenze.

1. Un'intelligenza Artificiale che apprende / 2. Le reti neurali artificiali / 3. Un'intelligenza aliena

1. Un'Intelligenza Artificiale che apprende

«*I understand that you want me to explain how Ava works, but I'm sorry. I'm not gonna be able to do that*» («Capisco che tu voglia che ti spieghi come funziona Ava, ma mi dispiace. Non posso farlo»).

Così Nathan, l'amministratore delegato di Blue-Book (la più importante azienda informatica al mondo), esprime al suo dipendente Caleb la difficoltà nel comprendere come il *robot* antropomorfo "Ava" funzioni realmente. Nel film *Ex Machina* (2015), diretto da Alex Garland, infatti, incontriamo un'Intelligenza Artificiale (IA) che riesce ad interagire in modo naturale con altri esseri umani, trasmettendo emozioni, desideri e (senza voler anticipare nulla del finale del film) perfino celando alcune delle sue intenzioni.

Pur essendo stata creata dagli esseri umani, però, il funzionamento di Ava sfugge alla loro comprensione. Sappiamo che Ava è un'IA che apprende dalle proprie esperienze ed è in grado di adattarsi a nuove situazioni. Sappiamo che è il *software* installato all'interno del suo corpo robotico (cioè del suo *hardware*) a dotarla delle avanzate capacità cognitive. Per quanto oggi nessun *software* di IA riesca ad ottenere gli strabilianti risultati del *software* ospitato da Ava, i principi fondanti del suo funzionamento sono già molto diffusi.

Ava *apprende* dalle proprie esperienze. Questa informazione consente di identificare Ava come un'IA di tipo "apprendimento automatico" (*machine learning*). L'apprendimento automatico costruisce un'IA che svolge compiti basandosi solamente su esempi di ciò che deve fare. L'essere umano non

fornisce esplicitamente le regole di comportamento per ciascuna delle possibili situazioni che l'IA incontrerà. In primo luogo, perché la varietà di tali situazioni è troppo ampia per essere predetta in anticipo. In secondo luogo, perché molto spesso il comportamento ottimale è difficilmente esprimibile con regole formali. Ad esempio, pensiamo a come noi esseri umani riconosciamo un gatto: se dovessimo elencare delle regole probabilmente diremmo che riconosciamo un gatto dai suoi tratti distintivi. Cerchiamo di individuare i caratteristici baffi, o le quattro zampe, o la coda. In pratica, però, siamo in grado di riconoscere un gatto da un'immagine anche senza la presenza di molte o tutte queste caratteristiche. Più in generale, per risolvere molti dei nostri compiti quotidiani ci affidiamo a conoscenza latente, tacita, che non sappiamo formalizzare. Per evitare questi problemi, un'IA basata sull'apprendimento automatico acquisisce la conoscenza sui gatti "semplicemente" osservando tanti esempi di gatti. Nell'apprendimento automatico il ruolo dell'umano è spesso confinato alla fase di raccolta degli esempi, che dovranno essere sufficientemente diversi tra loro da rappresentare in modo completo i concetti di interesse. Sarà l'IA stessa a creare i propri meccanismi di riconoscimento dagli esempi, senza che l'essere umano debba esplicitarli.

Il processo con cui un'IA acquisisce conoscenza di alcuni concetti tramite esempi è chiamato *apprendimento*. La fase di apprendimento presenta ripetutamente i dati all'IA (per esempio: l'immagine di un gatto), la quale fornirà una risposta (per esempio: "gatto", se tutto va bene). Inizialmente, però, l'IA non possiede alcuna conoscenza e le sue risposte saranno di conseguenza casuali. A seguito di ciascuna risposta, l'IA viene informata del risultato, e avrà quindi l'opportunità di cambiare la propria risposta quando, successivamente, le verrà ripresentato lo stesso esempio. L'apprendimento procede in questo modo, cambiando le risposte un poco alla volta, fino a che la risposta non diventa corretta per la maggior parte degli esempi (vedremo tra un attimo come la risposta viene generata e come è possibile cambiarla). Così facendo, l'IA costruisce le proprie rappresentazioni basandosi sulle regolarità presenti nei dati. Se tutte le immagini di gatti hanno qualcosa in comune, anche se è difficile definire chiaramente cosa, l'IA avrà l'opportunità di accorgersene e sfruttare questi schemi ricorrenti (*pattern*) per rispondere correttamente.

Ma qual è l'obiettivo dell'apprendimento? Rispondere correttamente a tutti gli esempi? Per quanto questo possa sembrare intuitivo, il vero scopo dell'apprendimento è leggermente diverso. Rispondere correttamente a tutti gli esempi ricevuti non assicura che l'IA abbia veramente compreso il problema, né che lo abbia imparato a risolvere in modo efficace. L'IA

potrebbe semplicemente memorizzare la risposta per ciascun esempio, evitando di identificare le regolarità utili per una sua vera soluzione. Nessun docente propone più volte lo stesso compito con le stesse domande agli stessi studenti. Tale compito perderebbe rapidamente qualunque utilità, visto che gli studenti riuscirebbero, dopo pochi tentativi, a ricordare esattamente le risposte corrette senza la necessità di acquisire alcuna competenza (tranne quella mnemonica). L'obiettivo della fase di apprendimento è dunque quello di costruire un'IA in grado di *generalizzare*, cioè di rispondere correttamente anche ad esempi leggermente diversi degli stessi concetti visti durante l'addestramento. È compito dell'essere umano che supervisiona la procedura di apprendimento (tipicamente un esperto del settore) assicurarsi che l'IA non stia scegliendo la strada della pura memorizzazione.

2. Le reti neurali artificiali

Fino a questo momento abbiamo ignorato una domanda rilevante: chi o cosa fornisce la risposta per ciascun esempio? In che modo? Nell'apprendimento automatico questo ruolo è svolto da un modello matematico: senza perderci in dettagli irrilevanti, è sufficiente considerare un modello matematico come una lunga sequenza di semplici operazioni che vengono applicate sull'esempio stesso fino a produrre la risposta finale. Per prima cosa quindi, l'esempio deve essere codificato in numeri. Ciò è solitamente immediato: le immagini diventano una lista di numeri che rappresentano il valore di ciascun *pixel*, le parole diventano una lista di numeri che ne indica la posizione in un vocabolario, e così via. Durante l'apprendimento, il modello matematico si adatta cambiando tipo di operazioni in modo tale che la risposta finale sia corretta. Analizziamo brevemente un esempio concreto. Se l'obiettivo è distinguere i gatti dai cani all'interno di immagini, il modello convertirà ciascuna immagine in una lista di numeri e applicherà la propria sequenza di operazioni. Supponiamo che moltiplichiamo prima tutti i numeri nella lista per 3, e poi sommiamo tutti i numeri tra loro. Il risultato è un unico numero finale che rappresenta la risposta. Se il numero finale è maggiore di zero, il modello risponderà "cane", altrimenti risponderà "gatto". Supponiamo che il modello riceva l'immagine di un gatto e risponda erroneamente "cane". Una volta informato dell'errore, il modello cambierà le operazioni per avvicinarsi alla risposta corretta, cioè a un numero negativo. Proverà quindi a moltiplicare tutti i numeri per (supponiamo) -2 anziché 3. Naturalmente, essendo il modello matematico unico per tutti gli esempi, cambiare le operazioni per un

esempio può portare a rispondere in modo errato a un esempio precedentemente risolto. Per questo motivo il processo di apprendimento richiede numerose iterazioni prima di trovare un buon compromesso in grado di generalizzare bene.

Il più diffuso modello matematico di apprendimento automatico è la rete neurale artificiale (chiamata così per una sua lontana ispirazione alle reti neurali biologiche, di cui non discuteremo qui). La proprietà principale di questo modello è la sua capacità di risolvere problemi scomponendoli autonomamente in sotto-problemi più semplici. Ad esempio, una rete neurale artificiale è in grado di riconoscere il contenuto delle immagini in quanto non tenta di risolvere il problema in un unico passaggio. Invece, il modello riconosce prima dei piccoli schemi ricorrenti (ad esempio, delle linee continue nell'immagine), poi combina questi schemi tra di loro per ottenerne di più complessi (diverse linee continue tra di loro formano ora un tratto che sembra un occhio) e così via fino a quando non diventa semplice risolvere il compito (avendo riconosciuto gli occhi, la bocca, il naso e i baffi la rete neurale artificiale risponde: "gatto"). Anche in questo caso, l'essere umano interviene nel processo principalmente raccogliendo i dati e curandone la qualità. Nessuno dice esplicitamente alla rete come decomporre il problema. Stiamo sempre parlando di apprendimento automatico, dopotutto.

Risolvere problemi tramite decomposizione automatica in sotto-problemi si è rivelato un paradigma di apprendimento molto potente. Negli ultimi 15/20 anni, le reti neurali artificiali sono passate dall'essere una tecnologia agli albori e prevalentemente confinata a pochi casi d'uso, a diventare il principale modello di IA, ampiamente utilizzato nella moderna IA generativa (*ChatGPT*, *Gemini*, *Claude* e simili), nella gestione delle immagini e dei video (solo per citare alcuni esempi: controllo di processi produttivi, diagnosi mediche, analisi delle prestazioni in vari sport), nell'analisi di segnali (conversione da audio a testo e viceversa, predizione dell'evoluzione di indici finanziari, etc.) e in molti altri settori.

Data la generalità delle sue applicazioni e dei suoi casi di successo, il paradigma di apprendimento automatico con reti neurali ci ha fatto scoprire come molte delle capacità ritenute esclusivamente umane fossero, in realtà, altamente automatizzabili. Non credo però che questi avanzamenti scientifici e tecnologici riducano l'ambito o la portata dell'intelligenza umana. Né che innalzino l'IA su vette irraggiungibili. L'IA riesce a portare a termine compiti per noi molto difficili (come si ripiega una proteina data la sua struttura lineare?) o addirittura impossibili (riconoscere rapidamente un evento anomalo tra decine di migliaia di eventi). Eppure, l'IA commette errori che noi non

commetteremmo mai. Per esempio, cambiando piccole porzioni di un'immagine è possibile indurre una rete neurale artificiale a restituire una risposta errata per un esempio prima correttamente risolto. Un essere umano, invece, non noterebbe nemmeno la differenza tra queste due immagini. Come si conciliano queste due realtà?

3. Un'intelligenza aliena

L'IA che utilizziamo oggi è un'intelligenza "aliena", complementare a quella umana, talvolta chiaramente più efficace, talvolta estremamente fragile. Così come Nathan non riesce a spiegare in modo soddisfacente il funzionamento di Ava, noi non riusciamo sempre a spiegare il comportamento delle IA moderne. Un possibile livello di interpretazione risiede nella sequenza di operazioni matematiche effettuate dal modello, ad esempio dalla rete neurale artificiale. Ma questa spiegazione di basso livello si perde nei dettagli di singole operazioni tra numeri e non cattura adeguatamente il significato semantico che porta alla predizione finale. Il fenomeno è simile a ciò che accade nei sistemi complessi in natura, dove tante semplici entità interagiscono tra loro: se guardiamo uno stormo muoversi in cielo, concentrarci sul movimento di una singola rondine non ci aiuta molto a comprendere come lo stormo crei le sue tipiche forme sinuose. Dall'altro lato, però, descrivere le forme nella loro interezza è una spiegazione altrettanto incompleta, che non chiarisce i meccanismi dietro la loro formazione. Allo stesso modo, anche tentare di spiegare il funzionamento dell'IA descrivendone soltanto le risposte non ha il sapore di una spiegazione completamente soddisfacente.

Sembra, quindi, che non ci sia un giusto livello di spiegazione. Come anticipato, questa frustrazione è comune nei sistemi complessi, dove il comportamento finale emerge dall'interazione tra le sue componenti. Nello stormo, studiare la relazione tra le singole rondini può aiutare a spiegarne le forme e i movimenti: ciascuna rondine cerca di volare vicino alle altre, per rimanere nel gruppo, ma sufficientemente lontana da non urtarle. L'applicazione di questa semplice regola da tutte le rondini contemporaneamente contribuisce a far emergere le dinamiche fondamentali del movimento dello stormo. Nelle reti neurali, tante operazioni matematiche interagiscono tra loro, ciascuna contribuendo alla risposta finale. È pur vero che al momento non sappiamo dire molto più di questo. Per quanto talvolta possiamo indurre le reti neurali a darci indizi sui motivi delle sue risposte (tramite il linguaggio, ad esempio, oppure chiedendo di identificare le regioni delle immagini

che sono state più utili per fornire la risposta), una spiegazione generale di come funziona un'IA rimane sfuggente. Molti comportamenti semplicemente emergono dalla sua complessità. Forse non a caso, negli ultimi anni il tema della complessità è diventato pervasivo nella nostra società.

Concludendo, le IA che apprendono dai dati sono ancora relativamente giovani. Servirà più tempo per capirne meglio le caratteristiche, per abituarci alla loro presenza e per intuire quando il loro utilizzo possa effettivamente aiutarci o rischi, invece, di essere deleterio. Per ora, l'IA è confinata prevalentemente in mondi virtuali. Ava non è qui tra noi. Il mito dei

robot antropomorfi che si comportano come gli esseri umani, pur essendo sicuramente più vicino alla realtà di quanto non lo fosse un secolo fa, rischia di distrarci da quelle che oggi sono le vere potenzialità dell'IA: non in competizione con gli umani, ma in una continua ricerca di proficua collaborazione. Più che sviluppare *robot* che seguano il mito dell'imitazione, le grandi sfide dei prossimi anni saranno incentrate intorno allo sviluppo di tecnologie capaci di ampliare le nostre possibilità di comprensione e di azione. In questo equilibrio dinamico tra due delle Intelligenze più avanzate del pianeta si gioca il futuro dell'Intelligenza Artificiale e del suo ruolo nella vita collettiva.