Macchine che vedono: prospettive e limiti della Visione Artificiale

Giovanna Castellano, Università degli Studi di Bari “Aldo Moro”

La visione è per noi una capacità naturale. Fin dalla nascita interpretiamo il mondo attraverso le immagini: riconosciamo volti, oggetti, colori, movimenti quasi senza rendercene conto. Questa apparente semplicità è in realtà il risultato di uno dei sistemi biologici più sofisticati che la natura abbia sviluppato. Il nostro sistema visivo inizia negli occhi, dove la luce viene trasformata in segnali elettrici dai fotorecettori della retina, ma la vera interpretazione avviene nella corteccia visiva del cervello. È lì che l’informazione visiva viene elaborata, organizzata e trasformata in significato. Vedere, dunque, non significa soltanto ricevere un’immagine: significa interpretarla, collegarla alla nostra esperienza e attribuirle un senso. Per noi esseri umani questo processo è immediato e quasi automatico. Guardiamo una scena e riconosciamo istantaneamente persone, oggetti, azioni. Tuttavia, ciò che per noi è naturale rappresenta una delle sfide più difficili per le macchine.

Ed è proprio da questa domanda che nasce la **visione artificiale** (*computer vision*): *è possibile dotare le macchine della capacità di vedere e comprendere il mondo visivo?*

La risposta, almeno in parte, è sì. Oggi esistono numerose applicazioni che dimostrano quanto questa tecnologia sia diventata potente. Pensiamo, ad esempio, alle auto a guida autonoma, che utilizzano sistemi di visione artificiale per riconoscere segnali stradali, individuare pedoni e identificare ostacoli sulla strada. Oppure ai sistemi di diagnostica medica assistita, che analizzano immagini radiologiche per supportare i medici nell’individuazione precoce di patologie. Anche nei nostri smartphone la visione artificiale è ormai presente, ad esempio nei sistemi di riconoscimento facciale o nelle applicazioni di realtà aumentata.

Nonostante questi straordinari progressi, la visione artificiale rimane una disciplina estremamente complessa. Paradossalmente, un sistema automatico può riconoscere migliaia di oggetti ma avere ancora difficoltà a distinguere, in alcune condizioni, un piccolo sasso da un foglio di carta. Un errore che per un essere umano sarebbe quasi impossibile.

Per comprendere perché questo accade bisogna considerare una differenza fondamentale tra il modo in cui noi vediamo e il modo in cui le macchine “vedono”. Quando un essere umano guarda un’immagine, percepisce immediatamente oggetti, relazioni e significati. Una macchina, invece, non vede oggetti: vede numeri. Più precisamente, un’immagine digitale è rappresentata come una matrice di valori numerici. Ogni punto dell’immagine, chiamato pixel, è descritto dalla combinazione di tre colori fondamentali: rosso, verde e blu. Per questo motivo un’immagine a colori è rappresentata da tre matrici di numeri, una per ciascun canale cromatico. Curiosamente, questa rappresentazione digitale riflette anche il funzionamento biologico del nostro occhio. I fotorecettori della retina sono infatti sensibili proprio a tre bande principali dello spettro luminoso: rosso, verde e blu.

Il problema nasce quando bisogna trasformare questi numeri in significato. In informatica questo problema è noto come semantic gap, ovvero la distanza tra la rappresentazione numerica a basso livello dei dati e la rappresentazione concettuale ad alto livello che utilizziamo noi esseri umani. Per molto tempo si è cercato di colmare questo divario attraverso sistemi basati su regole esplicite. Tuttavia il mondo visivo è troppo complesso e variabile perché possa essere descritto da un insieme rigido di regole. Anche gli esseri umani, del resto, non imparano a riconoscere il mondo in questo modo. Noi impariamo dagli esempi. Un bambino impara a riconoscere un gatto non perché qualcuno gli spiega formalmente cos’è un gatto, ma perché ne vede molti esempi. Dopo aver osservato gatti di forme, colori e dimensioni diverse, il cervello impara gradualmente a riconoscere le caratteristiche che definiscono quel concetto.

Questa stessa idea è alla base del machine learning, ovvero l’apprendimento automatico, una delle discipline fondamentali dell’intelligenza artificiale moderna. In questo approccio, invece di programmare

esplicitamente tutte le regole, si forniscono alla macchina grandi quantità di esempi, dai quali il sistema impara autonomamente a riconoscere schemi e regolarità. Ed è proprio qui che entrano in gioco i dati.

Uno dei progetti più importanti nella storia recente della visione artificiale è ImageNet, un enorme archivio di immagini creato nel 2009 da Fei-Fei Li, informatica e professoressa alla Stanford University. Il progetto nasce con un obiettivo ambizioso: costruire una gigantesca banca dati visiva che potesse essere utilizzata per addestrare algoritmi di riconoscimento delle immagini. ImageNet contiene oltre 15 milioni di immagini, organizzate in più di 22.000 categorie di oggetti. Questo dataset ha avuto un impatto enorme nello sviluppo dell'intelligenza artificiale moderna, perché ha permesso di addestrare modelli sempre più sofisticati basati su reti neurali artificiali.

Le reti neurali sono modelli computazionali ispirati, almeno concettualmente, alla struttura del cervello umano. Il nostro cervello è composto da circa 86 miliardi di neuroni, interconnessi da un numero ancora maggiore di connessioni sinaptiche. Le reti neurali artificiali cercano di imitare questo principio attraverso strutture composte da strati di unità computazionali che trasformano progressivamente l'informazione. Ogni neurone artificiale esegue una semplice operazione matematica, ma quando milioni di queste operazioni vengono organizzate in strutture profonde — le cosiddette deep neural networks — emergono capacità sorprendenti. Grazie a questi modelli, i sistemi di visione artificiale sono oggi in grado di riconoscere migliaia di categorie di oggetti, segmentare scene complesse e persino generare descrizioni testuali delle immagini.

Negli ultimi anni, inoltre, l'integrazione tra visione artificiale e modelli linguistici di grandi dimensioni (LLM) ha portato allo sviluppo di sistemi capaci di descrivere in linguaggio naturale il contenuto di un'immagine, generando spiegazioni dettagliate e contestuali.

Nonostante questi progressi straordinari, manca ancora qualcosa di fondamentale: il senso comune. Gli esseri umani interpretano il mondo non solo sulla base di ciò che vedono, ma anche sulla base di un enorme patrimonio di conoscenze implicite che derivano dall'esperienza quotidiana. Sappiamo, ad esempio, che un elefante non può stare sopra un albero o che una tazza rovesciata probabilmente verserà il suo contenuto. Questo tipo di conoscenza, apparentemente banale, è estremamente difficile da formalizzare e insegnare alle macchine. L'assenza di senso comune rende i sistemi di intelligenza artificiale ancora vulnerabili a errori e interpretazioni sbagliate. Per questo motivo oggi uno dei temi più importanti nella ricerca sull'IA è quello della affidabilità.

Quanto possiamo fidarci delle decisioni prese dai sistemi di intelligenza artificiale? Un caso emblematico è quello di Nijeer Parks, un uomo afroamericano ingiustamente condannato nel 2019 perché il giudice si fidò ciecamente del sistema di visione artificiale per il riconoscimento facciale che trovò una perfetta corrispondenza tra la foto del ragazzo memorizzata nei database della polizia e la foto sulla patente contraffatta ritrovata sul luogo del crimine. Il caso ha sollevato un ampio dibattito internazionale sui limiti e sui rischi dell'uso non critico di sistemi di visione artificiale in contesti sensibili.

Episodi come questo ricordano che l'intelligenza artificiale non è infallibile e che le decisioni automatizzate devono sempre essere accompagnate da controllo umano. Per questo motivo oggi sempre più ricercatori parlano di Symbiotic AI o Human-AI collaboration: un modello in cui l'intelligenza artificiale non sostituisce l'essere umano, ma lo affianca e ne amplifica le capacità. In questa prospettiva, le macchine possono analizzare grandi quantità di dati e individuare schemi complessi, mentre gli esseri umani mantengono la capacità di interpretazione, giudizio e responsabilità. Il futuro dell'intelligenza artificiale, quindi, non è quello di un conflitto tra uomo e macchina, ma quello di una cooperazione intelligente. Una relazione di reciproco arricchimento in cui la tecnologia rende i processi decisionali più rapidi ed efficienti, senza mai perdere di vista il ruolo centrale dell'essere umano. In altre parole, l'uomo rimane al centro, supportato da strumenti sempre più potenti, ma ancora guidati dalla nostra capacità di comprendere il mondo.