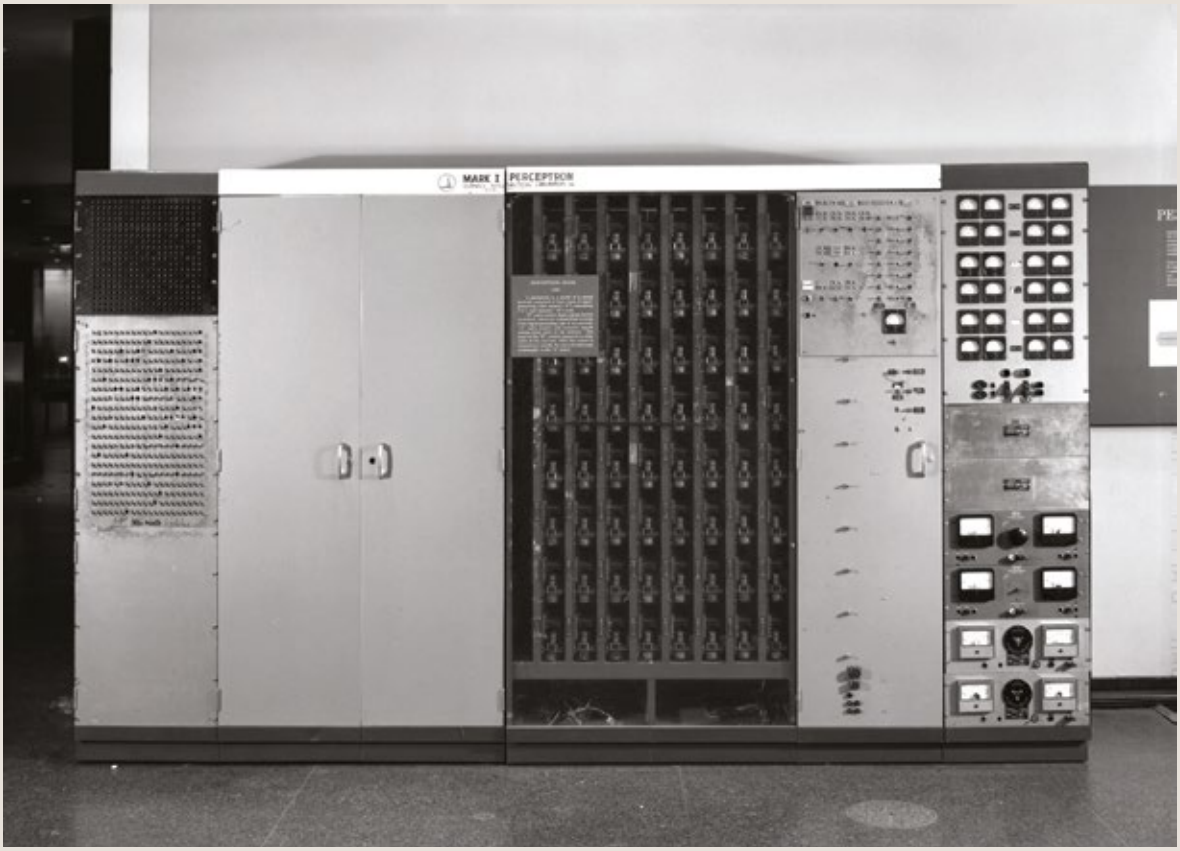
Rovescio vincente

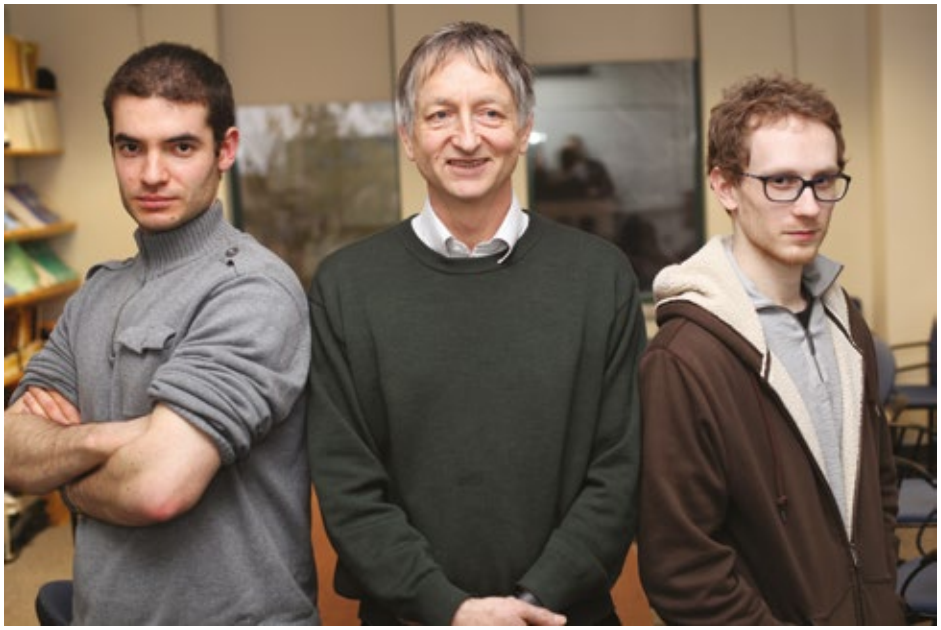
Backpropagation e reti neurali

di Matteo Osella



a.
Perceptron Mark I, la prima macchina costruita per mettere in pratica l'algoritmo del percettrone, il primo modello di neuroni artificiali. Era collegata a una fotocamera dotata di 400 sensori di luce (disposti in una griglia 20×20), in grado di catturare immagini semplici. Dietro le due ante nella parte sinistra della fotografia è alloggiato un pannello con cavetti, usato per collegare in modo diverso i segnali in ingresso e "insegnare" alla macchina a riconoscere modelli. Nella parte centrale, che assomiglia un po' a una griglia, si trovano dei "potenziometri" (piccoli componenti che permettono di regolare un potenziale elettrico), che servono a modificare i "pesi" dell'apprendimento, ovvero quanto ogni segnale influenzava la decisione finale.

Fin dagli anni '40, con la nascita del primo modello di neurone, il percettrone, si è fatta strada l'idea che funzioni cognitive complesse potessero emergere da reti di molte unità elementari interconnesse. In quest'ottica, le reti neurali rappresentano un esempio paradigmatico di sistema complesso: migliaia, milioni di unità semplici la cui dinamica collettiva dà luogo a comportamenti sofisticati, non riducibili alla somma delle parti.
Non sorprende quindi che la fisica statistica, con il suo bagaglio di concetti e tecniche sviluppate per modellizzare il comportamento collettivo di sistemi disordinati e interagenti, si sia presto interessata alle reti neurali. Esistono numerose connessioni formali e concettuali con la fisica statistica. Per esempio, nell'apprendimento supervisionato si conosce l'output desiderato per un insieme di esempi, e su questi si cerca di minimizzare l'errore commesso dalla rete. Questo errore viene rappresentato da una "funzione costo", analoga all'energia in un sistema fisico. Infatti, come in fisica statistica si cercano le configurazioni di minima energia di un sistema, così durante l'apprendimento cerchiamo i minimi della funzione costo, corrispondenti a configurazioni di parametri che meglio spiegano i dati. Tuttavia, come in molti sistemi disordinati, il panorama energetico puo' presentare una geometria complessa e frastagliata, che l'algoritmo di apprendimento deve esplorare alla ricerca di "buoni" minimi, ovvero configurazioni in grado di fornire predizioni accurate anche su dati non visti.
Negli anni '80, fisici come Giorgio Parisi, Marc Mézard, Bernard



b.
Il professore Geoffrey Hinton al centro, con i suoi studenti Ilya Sutskever, a sinistra, e Alex Krizhevsky, a destra, all'Università di Toronto (Canada) nel 2013.

Derrida e Haim Sompolinsky (per citarne alcuni), facendo leva su queste numerose analogie, iniziarono a costruire una "meccanica statistica" delle reti neurali. Negli stessi anni, nuovi modelli, esplicitamente ispirati alla fisica, per esempio dei vetri di spin (vd. in Asimmetrie n. 32 p. 12, ndr), vennero proposti. Un celebre esempio è il modello di memoria associativa di Hopfield (vd. approfondimento a p. 12, ndr).
Tuttavia, per lungo tempo, le reti neurali artificiali rimasero ai margini in termini di applicazioni pratiche. Erano, e sono tuttora, utilizzate nelle neuroscienze computazionali come possibile modello semplificato del funzionamento del cervello, anche se la maggior parte degli addetti ai lavori concordano sulla distanza abissale esistente tra queste due diverse incarnazioni di "intelligenza". Analogamente, seppur le reti neurali fossero un attivo ambito di ricerca in *computer science*, avevano ancora uno scarso impatto pratico. L'ingegneria e l'industria preferivano soluzioni più robuste e controllabili.
Il punto di svolta arrivò nel 2012, quando un lavoro di Geoffrey Hinton in collaborazione con Alex Krizhevsky e Ilya Sutskever (futuro cofondatore di OpenAI e figura chiave nello sviluppo di ChatGPT) mostrò che le reti profonde (o *deep*, da cui il termine *deep learning*), se addestrate con grandi quantità di dati e risorse computazionali, potevano superare ogni precedente metodo nel riconoscimento di immagini.
Gli ingredienti fondamentali erano già in circolazione: l'algoritmo di *backpropagation* per l'addestramento di reti profonde era stato sviluppato già negli anni '80, così come l'intuizione di costruire reti specializzate per l'analisi di immagini – le reti convoluzionali – che incorporavano il concetto di località. In analogia con la corteccia visiva del cervello, i neuroni nelle reti convoluzionali si specializzano nell'identificazione di *pattern* in regioni locali dell'immagine, contribuendo alla costruzione di una rappresentazione gerarchica che conserva le relazioni spaziali. Tuttavia, mancava ancora la giusta combinazione di questi elementi e l'hardware adeguato (l'utilizzo delle GPU) per arrivare a un modello efficiente e addestrabile su grandi quantità di dati.
Da allora, le reti neurali sono entrate nella nostra vita quotidiana: nei nostri telefoni, negli algoritmi di raccomandazione, nella diagnostica medica, nei processi industriali. Negli ultimi anni abbiamo assistito anche alla rapida ascesa dei modelli generativi di linguaggio, addestrati semplicemente a predire la parola successiva in una frase basandosi su enormi raccolte di testi (circa 300 miliardi di parole nel caso di ChatGPT-3), ma capaci di sviluppare funzioni complesse e in parte inattese.
Eppure, mentre l'ingegneria dell'apprendimento automatico progredisce a ritmi straordinari, la teoria resta indietro. Oggi costruiamo modelli sempre più grandi, energivori e sofisticati, ma non possediamo ancora una vera comprensione del perché funzionino. Una teoria dell'intelligenza artificiale con la solidità predittiva della meccanica o della termodinamica ancora non esiste. E il concetto stesso di "intelligenza" resta sfuggente, privo di una definizione operativa condivisa e rigorosa.
La situazione ricorda quella del XVIII secolo: l'utilizzo dei motori a vapore stava aprendo le porte alla rivoluzione industriale che avrebbe trasformato il mondo, ma la termodinamica non era ancora stata del tutto sviluppata. Solo con la formalizzazione teorica si poté comprendere il funzionamento di quei motori e guidare l'innovazione tecnologica verso forme più efficienti. Allo stesso modo, oggi ci troviamo ad applicare modelli potenti di apprendimento senza una teoria che ne spieghi a fondo il comportamento, che ne quantifichi i limiti, che ne guidi lo sviluppo futuro.
Esiste un ampio corpo di eleganti contributi teorici che arrivano principalmente dal mondo della matematica, ma spesso si concentrano su modelli fortemente semplificati e sui limiti delle loro performance in scenari "*worst-case*", ovvero validi per qualsiasi possibile dato di input, ma poco rilevanti per i casi reali. I dati concreti che ci interessano hanno struttura,

regolarità, correlazioni complesse – proprietà che raramente entrano nelle trattazioni teoriche tradizionali. Inoltre, le architetture moderne sono così sofisticate da essere, almeno per ora, inaccessibili agli strumenti matematici formali di cui disponiamo.
Così si procede per tentativi: metodi empirici, ottimizzazioni *ad hoc*, intuizioni ingegneristiche. Un progresso rapido, ma in parte opaco, quasi "alchemico" come è stato definito in una discussa relazione a una delle più importanti conferenze del settore (NeurIPS).
È in questo scenario che i fisici potrebbero giocare un ruolo cruciale. Per colmare il *gap* attuale tra teoria e pratica serve precisamente ciò che la fisica fa da sempre: costruire modelli fortemente semplificati, ma capaci di catturare gli ingredienti fondamentali e quindi di essere predittivi, anche a costo di rinunciare al rigore formale matematico. È il momento di cercare leggi generali domandandosi quali aspetti – la struttura dei dati, l'architettura della rete, la dinamica dell'addestramento, la funzione di errore – siano davvero cruciali e quali invece siano dettagli contingenti, frutto di scelte storiche o vincoli tecnologici.
In questo solco sarà forse possibile costruire una "teoria dell'intelligenza" che ci consenta non solo di costruire reti che funzionano, ma di capire anche perché funzionano, quando falliscono e come migliorarle.

c.
La situazione attuale dell'IA ricorda quella del XVIII secolo, quando l'utilizzo dei motori a vapore stava aprendo le porte alla rivoluzione industriale che avrebbe trasformato il mondo, ma la termodinamica non si era ancora sviluppata appieno.

Biografia
Matteo Osella si è laureato in fisica teorica all'Università degli Studi di Torino, dove ha proseguito gli studi con un dottorato di ricerca in sistemi complessi per la biologia. Ha lavorato alla Sorbonne Université di Parigi e ora è professore associato all'Università di Torino. È docente di reti neurali presso il Dipartimento di Fisica e i suoi temi di ricerca riguardano la fisica statistica, la fisica biologica e le reti neurali artificiali.

10.23801/asimmetrie.2025.39.3