**L’Edge AI: il nuovo paradigma dello sviluppo**

Negli ultimi anni, anche grazie alle grandi mutazioni sociali e tecnologiche che hanno interessato tutti i continenti, stiamo vivendo un’accelerazione nello sviluppo dell'intelligenza artificiale, la cosiddetta AI.

Grazie ai progressi degli algoritmi e alla potenza di calcolo applicata ai big data, il deep learning - il settore più brillante dell'IA - ha fatto grandi progressi in molti campi, che vanno dalla computer vision, al riconoscimento vocale e all'elaborazione del linguaggio naturale, fino alla robotica.

È ampiamente riconosciuto che applicazioni intelligenti stanno arricchendo in modo significativo lo stile di vita delle persone, cercando di migliorare la produttività umana e l'efficienza sociale.

Come fattore chiave per lo sviluppo dell'intelligenza artificiale, i big data hanno recentemente subito un radicale trasformazione della fonte di elaborazione dei dati, passando dai data center in cloud ai dispositivi finali sempre più diffusi, quali, ad esempio i dispositivi mobili e i dispositivi Internet-of-Things (IoT) anche di tipo industriale (IIOT).

Tradizionalmente, i big data nascevano e venivano archiviati principalmente nei grandi hyperscales, datacenters di grandissima capacità distribuita su scala mondiale. Tuttavia, con la proliferazione del mobile computing e dell'IoT, la tendenza sta mutando: infatti gli enormi volumi di dati che devono essere spostati dai dispositivi terminali al cloud per l’elaborazione centralizzata creano un sensibile aumento dei costi di gestione delle infrastrutture di collegamento e dei tempi di elaborazione.

Spingersi verso l'ecosistema edge non è tuttavia un’operazione banale.

La tendenza, infatti, era quella di trasportare le masse di dati dai dispositivi locali ai data center del cloud per l'analisi. Tuttavia, quando si sposta un'enorme quantità di dati attraverso la rete, sia il costo sia il ritardo di trasmissione possono essere proibitivi cosi come la perdita di privacy.

Negli ultimi anni è stato introdotto il concetto di analisi computazionale “on-device”, che esegue applicazioni di intelligenza artificiale sul dispositivo (l’”edge” appunto) per elaborare localmente i dati IoT. Tali devices, tuttavia, possono soffrire di scarse prestazioni ed efficienza energetica. Questo perché molte applicazioni di IA richiedono un'elevata potenza di calcolo che supera di gran lunga la capacità dei dispositivi IoT, che sono limitati dal punto di vista delle risorse e dell'energia.

Per realizzare il progetto sopra descritto la comunità tecnologica a introdotto l'edge-computing, che spinge i servizi computazionali dal nucleo della rete (cloud) ai bordi della rete (edge), più vicini ai dispositivi locali e alle fonti dei dati.

Un nodo edge può essere quindi un dispositivo finale locale, collegabile tramite comunicazioni device-to-device (D2D), un server collegato a un punto di accesso (ad esempio, un WiFi Access Point, una stazione radio base), un gateway di rete o micro datacenters locali (chiamati “edge datacenter”) disponibile per l'uso da parte dei dispositivi vicini. I nodi edge possono quindi avere dimensioni molto diverse: da un computer delle dimensioni di una carta di credito a un edge datacenter con numerosi rack di server.

La vicinanza fisica alle fonti di generazione delle informazioni è quindi la caratteristica più importante enfatizzata dall'edge-computing. In sostanza, la vicinanza fisica tra le fonti di elaborazione e di generazione delle informazioni permette vantaggi rispetto al tradizionale paradigma di elaborazione basato su cloud, quali la bassa latenza, l’efficienza energetica, la protezione della privacy, il consumo ridotto di larghezza di banda.

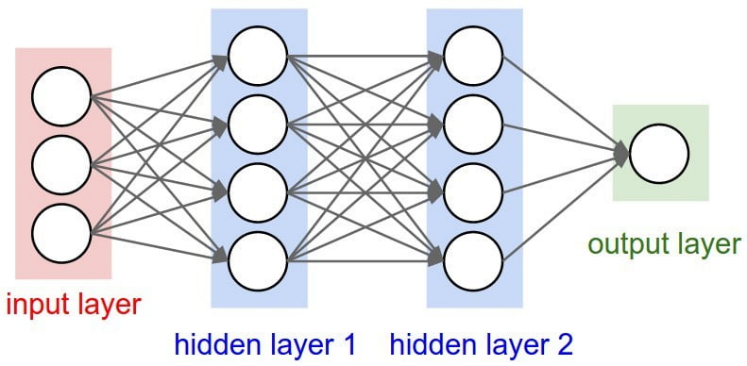


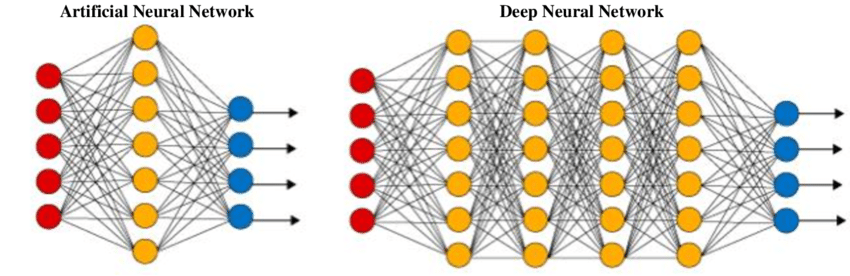
Il connubio tra edge computing e IA ha dato origine a una nuova area, la "**edge intelligence (EI)**" o "**edge AI**". Invece di affidarsi completamente al cloud, l'EI sfrutta le risorse edge diffuse per ottenere la potenza di calcolo desiderata.

Anche se l'IA è salita recentemente alla ribalta, non è un termine nuovo ed è stato coniato per la prima volta nel 1956: è un approccio per costruire macchine intelligenti in grado di svolgere compiti simili a quelli dell'uomo. Una branca dell’IA è costituita dall'apprendimento automatico (o “machine learning in modo abbreviato ML).

Molte metodologie di ML, come per esempio gli alberi decisionali, il clustering K-Means, le SVM (Support Vector Machines) e le reti neurali, sono state sviluppate per addestrare la macchina a produrre classificazioni e previsioni, sulla base dei dati ottenuti dal mondo reale.

Tra gli algoritmi di ML più impegnativi dal punto di vista computazionale abbiamo il deep learning, che, basandosi su reti neurali “profonde”, ha portato a prestazioni molto importanti in diversi campi, tra cui la classificazione delle immagini, il riconoscimento dei volti, la classificazione dei segnali audio e così via. Poiché le reti neurali profonde consistono tipicamente in una serie di layers, il modello è chiamato rete neurale profonda (DNN) in cui ogni strato di una DNN è composto da neuroni che sono in grado di generare uscite non lineari sulla base dei dati provenienti dall'ingresso del neurone.





Alimentando un gran numero di campioni di addestramento e ripetendo questo processo finché il tasso di errore non è inferiore a una soglia predefinita, si ottiene un modello di apprendimento con un'elevata precisione.

L'inferenza (vale a dire la predizione in base a nuovi dati) del modello avviene dopo l'addestramento. Ad esempio, per la classificazione di immagini, con l'alimentazione di un gran numero di campioni di addestramento, viene addestrata una DNN per imparare a riconoscere un'immagine; quindi, l'inferenza prende in input le immagini del mondo reale e ne ricava rapidamente le previsioni/classificazioni.

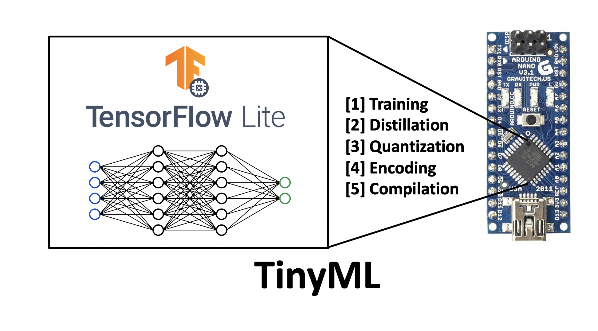
L'edge-computing è ultimamente diventato fondamentale per applicazioni che sfruttano i suddetti algoritmi per avere inferenze in tempo reale, dato che l'edge-computing mira a coordinare una moltitudine di dispositivi e server edge collaborativi per elaborare i dati generati in prossimità.

A seguito della proliferazione del numero e dei tipi di dispositivi mobili e IoT, grandi volumi di dati multimodali (ad esempio, audio, immagini e video) provenienti dell'ambiente vengono continuamente rilevati dal lato dei dispositivi. In questo contesto, l'Edge-AI risulta indispensabile per la sua capacità di analizzare rapidamente questi enormi volumi di dati e di estrarre da essi informazioni utili per prendere decisioni di alta qualità.

Il deep learning, infatti, offre la capacità di identificare automaticamente modelli e rilevare anomalie nei dati rilevati all’edge, come ad esempio il flusso del traffico nelle smart cities, la telemedicina, i dati ambientali, elettrici e audio per catturare le anomalie di funzionamento delle macchine, la guida autonoma.

Le predizioni estratte dal modello costruito sui dati rilevati vengono poi utilizzate per il processo decisionale predittivo in tempo reale in risposta agli ambienti in rapida evoluzione, aumentando l'efficienza operativa.

Come ulteriore miglioramento in questi ultimi due anni sta prendendo piede il paradigma dell'esecuzione di algoritmi di IA a livello locale su un dispositivo finale, con dati (dati di sensori o segnali) elaborati sul dispositivo, la cosiddetta “TinyML”, anche se tali elaborazioni sono limitate dalle prestazioni dei microprocessori utilizzati sui devices finali.



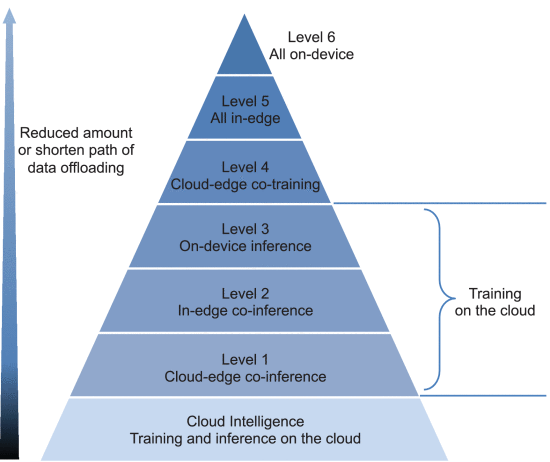
Ecco che dunque la Edge-Ai può essere bene rappresentata con il connubio computazionale fra edge data centers e devices finali, a seconda delle prestazioni che devono essere ottenute. E’ ciò che oggi permette la “democratizzazione dell'IA”, cioè la visione di "rendere fruibile l'IA per ogni persona e ogni organizzazione, ovunque".

L'ambito di applicazione dell'IE, tuttavia, non deve essere limitato all'esecuzione di modelli di intelligenza artificiale esclusivamente sul server o sul dispositivo edge. Infatti, come dimostrato da studi sulla materia, per i modelli DNN, l'esecuzione con la sinergia edge-cloud può ridurre sia la latenza end-to-end sia il consumo energetico rispetto all'approccio di esecuzione locale.

L'IE deve essere il paradigma che sfrutta appieno i dati e le risorse disponibili nella gerarchia dei dispositivi finali, dei nodi edge e dei data center del cloud per ottimizzare le prestazioni complessive dell'addestramento e dell'inferenza di un modello DNN. Tuttavia questo non significa necessariamente che il modello DNN sia completamente addestrato o inferenziato all'edge, ma che può funzionare in modo coordinato cloud-edge-device tramite l'offloading dei dati. In particolare, in base alla quantità e alla lunghezza del percorso di offloading dei dati, classifichiamo l'EI in sei livelli, come illustrato nella figura.

In particolare, è possibile definire sette livelli di EI[[1]](#footnote-1) costituenti un continuum tra il cloud computing degli hyperscales, gli edge datacenter e i device finali.

1. ***Livello 0:*** addestramento e inferenza completamente nel cloud
2. ***Livello 1***: addestramento nel cloud, ma inferenza in modo cooperativo edge-cloud; in questo caso, la cooperazione edge-cloud implica lo scarico parziale dei dati nel cloud
3. ***Livello 2:*** addestramento nel cloud, ma l’inferenza viene eseguita “in-edge”, il che può essere realizzato scaricando completamente o parzialmente i dati sui server edge o su dispositivi vicini
4. ***Livello 3:*** addestramento nel cloud e inferenza sul dispositivo finale "on-device": in questo caso non viene scaricato alcun dato sul server edge
5. ***Livello 4:*** addestramento e inferenza in modalità di cooperazione edge-cloud
6. ***Livello 5:*** addestramento e inferenza in modalità in-edge, cioè sui server edge
7. ***Livello 6***: addestramento e inferenza del modello DNN in modalità on-device

[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/5/8789751/8736011/chen7-2918951-large.gif)

All'aumentare del livello di EI, la quantità e la lunghezza del percorso di offloading dei dati si riducono. Di conseguenza, la latenza di trasmissione dell'offloading dei dati diminuisce, la privacy dei dati aumenta e il costo della larghezza di banda WAN si riduce. Tuttavia, questo risultato è ottenuto al costo di un aumento della latenza computazionale e del consumo energetico. Questo conflitto indica che non esiste un "livello migliore" in generale; piuttosto, il "livello migliore" di IE dipende dall'applicazione e deve essere determinato considerando congiuntamente più criteri come la latenza, l'efficienza energetica, la privacy e il costo della larghezza di banda WAN.

Le architetture di addestramento distribuito di DNN all'edge possono essere suddivise in tre modalità: centralizzata, decentralizzata e ibrida.

Nella modalità centralizzata il modello DNN viene addestrato nel datacenter del cloud. I dati per l'addestramento sono generati e raccolti da dispositivi finali distribuiti, come telefoni cellulari, automobili e telecamere di sorveglianza. Una volta ricevuti i dati, il datacenter cloud esegue l'addestramento DNN utilizzando questi dati. Pertanto, il sistema basato sull'architettura centralizzata può essere identificato in cloud intelligence di livello 0, livello 1, livello 2 o livello 3 a seconda del tipo di inferenza utilizzato.

Nella modalità decentrata, ogni server edge/dispositivo locale addestra localmente il proprio modello DNN con dati locali, preservando le informazioni private a livello locale. Per ottenere il modello DNN globale devono essere condivise le migliorie apportate all'addestramento locale; pertanto i nodi della rete comunicheranno tra loro per scambiarsi gli aggiornamenti del modello locale. Questa modalità corrisponde al livello 5 di EI.

L’ultima modalità è quella ibrida, in cui vengono combinate la modalità centralizzata e quella decentralizzata. In questo caso gli edge server possono addestrare il modello DNN sia tramite aggiornamenti decentralizzati tra loro, sia tramite l'addestramento centralizzato con il cloud datacenter. Questa modalità corrisponde al livello 4 e 5 di EI.

Come ho giá avuto modo di anticipare nel corso dell’articolo, i parametri caratteristici da monitorare nell’implementazione di un modello EI sono la banda e l’efficienza energetica, la latenza, il costo della comunicazione, l’affidabilità, la sicurezza. Per ricordarli è stato coniato l’acronimo ***BLERP*** (**B**andwith, **L**atency, **E**conomics, **R**eliability, **P**rivacy) da Jeff Bier, fondatore di EDGE AI e Vision Alliance.

Vediamo nel seguito come influiscono nella fase di addestramento.

L'addestramento del modello complesso è ad alta intensità di dati, poiché i dati grezzi o intermedi devono essere trasferiti tra i nodi. Intuitivamente, questo overhead di comunicazione aumenta l'energia, il consumo di banda e latenza di addestramento. L'overhead di comunicazione è quindi influenzato dalla dimensione dei dati di input originali, dalla modalità di trasmissione e dalla larghezza di banda disponibile.

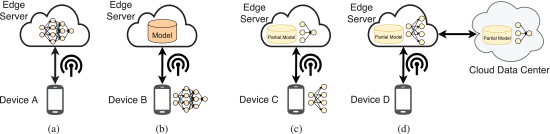
Quindi anche l’efficienza energetica può essere considerata un parametro legato alla banda (quantità di dati che devono essere trasmessi). Quando si addestra il modello tipo DNN in modo decentralizzato, per esempio, sia il processo di calcolo sia quello di comunicazione consumano molta energia. Tuttavia, per la maggior parte dei dispositivi finali, l'energia è limitata (alimentazione a batteria). Di conseguenza è necessario che l'addestramento del modello sia efficiente dal punto di vista energetico.

La latenza è probabilmente uno degli indicatori più importanti dell'addestramento distribuito poiché influenza direttamente il momento in cui il modello addestrato è disponibile per l'uso. La latenza del processo di addestramento distribuito consiste tipicamente nella latenza di calcolo e nella latenza di comunicazione. La latenza di calcolo dipende strettamente dalla capacità dei nodi edge (edge datacenters e nodi finali). La latenza di comunicazione può variare dalla dimensione dei dati grezzi o intermedi trasmessi e dalla larghezza di banda della connessione di rete.

La connettività di un device costa molto in termini economici. Più banda è richiesta, più il costo aumenta. Effettuare l’addestramento sull’edge (sugli edge datacenters o dove possibile sui devices locali) permette un risparmio economico notevole. In ogni caso l’impatto economico non deve essere trascurato perché non sempre è possibile risolverlo con l’introduzione dell’edge-computing.

Anche la privacy dei dati non deve essere trascurata. Limitare la connettività all’edge preserva la privacy dei dati, dato che non devono essere trasmessi attraverso la rete di comunicazione con il cloud. Per esempio, nel caso della video analisi, piuttosto che effettuare lo streaming video e audio ad un server remoto, può essere utilizzata l’”intelligenza” racchiusa in una telecamera predisposta per EI per elaborare gli algoritmi di apprendimento con DNN.

Vediamo ora le tipologie di architetture che possono essere utilizzate per l’inferenza.

[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/5/8789751/8736011/chen12abcd-2918951-large.gif)

a) ***Modalità edge-based***: il dispositivo A è in modalità **edge-based**, il che significa che il dispositivo riceve i dati di input e li invia al server edge. Quando l'inferenza viene eseguita dall'edge server, i risultati della predizione vengono restituiti al dispositivo. In questa modalità di inferenza, poiché il modello si trova sull'edge server, è facile implementare l'applicazione su diverse piattaforme mobili. Tuttavia, lo svantaggio principale è che le prestazioni di inferenza dipendono dalla larghezza di banda della rete tra il dispositivo e l'edge server.

b) ***Modalità basata sul dispositivo***: il dispositivo B è in modalità **device-based**. Il dispositivo mobile ottiene il modello dall'edge server ed esegue l'inferenza del modello a livello locale. Durante il processo di inferenza, il dispositivo mobile non comunica con l'edge server. Pertanto, l'inferenza è affidabile, ma richiede una grande quantità di risorse come CPU, GPU e RAM sul dispositivo mobile. Le prestazioni dipendono dal dispositivo locale stesso.

c) ***Modalità Edge-Device:*** il dispositivo C è in modalità **edge-device**. Nella modalità edge-device, il dispositivo per prima cosa partiziona il modello in più parti in base ai fattori ambientali del sistema connesso, come la larghezza di banda della rete, le risorse del dispositivo e il carico di lavoro del server edge. Quindi, il dispositivo esegue il modello fino a un livello specifico e invia i dati intermedi all'edge server. L'edge server esegue i livelli rimanenti e invia i risultati della previsione al dispositivo. Rispetto alle modalità edge-based e device-based, la modalità edge-device è più affidabile e flessibile.

d) ***Modalità Edge-Cloud:*** il dispositivo D è in modalità **edge-cloud**. È simile alla modalità edge-device ed è adatta al caso in cui il dispositivo sia fortemente limitato nelle risorse. In questa modalità, il dispositivo è responsabile della raccolta dei dati di input e il modello viene eseguito attraverso la sinergia edge-cloud. Le prestazioni di questo modello dipendono fortemente dalla qualità della connessione di rete.

Va sottolineato che le quattro modalità di inferenza edge-centriche sopra menzionate possono essere adottate simultaneamente in un sistema per eseguire compiti complessi di inferenza di modelli AI (ad esempio, gerarchia cloud-edge-device), utilizzando in modo efficiente risorse eterogenee tra una moltitudine di dispositivi finali, nodi edge e cloud.

Gli stessi parametri considerati per i processi di addestramento devono essere considerati per l’inferenza, soprattutto per i livelli EI che interessano i dispositivi edge più che i server edge.

I devices locali spesso catturano più dati della banda trasmissiva a disposizione. Immaginiamo un sensore intelligente che controlli le vibrazioni di una macchina industriale tramite l’acquisizione di segnali audio su cui effettuare l’inferenza. I dati raccolti possono essere milioni in brevissimo tempo. Cosa potrebbe succedere se non venissero elaborati tutti o il sensore perdesse dati fondamentali poco prima della rottura? Cosa potrebbe succedere se, visto l’altissimo numero di campioni raccolti, avvenisse la parziale perdita legata all’efficienza energetica del device?

Per quanto riguarda la latenza, per esempio, per alcune applicazioni mobili intelligenti in tempo reale (ad esempio, telemedicina, giochi mobili AR/VR e robot intelligenti), i requisiti sono molto stringenti, come ad esempio la latenza di 100 ms. Anche l’energia ha per l’inferenza stretti vincoli a causa dell’alimentazione a batteria dei device e dell’energia necessaria al calcolo.

Ad eccezione della modalità basata sul dispositivo, l'overhead di comunicazione influisce notevolmente sulle prestazioni di inferenza delle altre modalità

Anche l'utilizzo di memoria nell'esecuzione dell'inferenza soprattutto sui dispositivi mobili è un parametro da valutare attentamente. Da un lato, infatti, un modello DNN è accompagnato da milioni di parametri, il che può risultare molto impegnativo per le risorse hardware dei dispositivi mobili. D'altra parte, a differenza delle GPU discrete ad alte prestazioni dei data center, non esiste una memoria dedicata ad alta larghezza di banda per le GPU sui dispositivi mobili. Inoltre, le CPU e le GPU mobili sono tipicamente in competizione per la larghezza di banda della memoria condivisa e scarsa. Per l'ottimizzazione dell'inferenza DNN sul lato edge, l'ingombro della memoria è un indicatore non trascurabile. L'impronta di memoria è influenzata principalmente dalle dimensioni del modello DNN originale e dal modo in cui vengono caricati i tremendi parametri DNN.

Con questo articolo ho cercato di evidenziare le caratteristiche peculiari dell’Edge AI e come esso rappresenti il nuovo paradigma per lo sviluppo. Si apre quindi la strada a una vasta gamma di nuove applicazioni in cui è necessaria l’estrema reattività del sistema computazionale.

In sintesi, l'Edge AI sta rivoluzionando il mondo fino ad ora conosciuto facendo si che i dati in realtime diventino centrali nello sviluppo economico e tecnologico, dando vita a una nuova era, quella dell'intelligenza artificiale decentralizzata.

1. *cfr. Daniel Situnayake, Jenny Plunkett “AI at the Edge” O’Reilly e Zhou e altri: “Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing”, Proceedings of IEEE* [↑](#footnote-ref-1)