TIM

INTELLIGENZA ARTIFICIALE

notiziariotecnico

3-2023

Il **Notiziario Tecnico** è un webmagazine, con taglio tecnico-divulgativo che presenta in modo ragionato l'evoluzione del settore delle tecnologie dell'informazione, dando particolare attenzione alle sinergie tra innovazione digitale e scenari di business.

Notiziario Tecnico
Anno 32 - Numero 3, Dicembre 2023
www.telecomitalia.com/notiziariotecnico

Proprietario ed editore
TIM S.p.a.

Direttore responsabile
Michela Billotti

Comitato di direzione
Gabriele Elia
Daniele Franceschini
Elisabetta Romano
Federica Romano

Web Director
Enrico Gallo

Photo
123RF Archivio Fotografico
Archivio Fotografico TIM

Redazione
Roberta Bonavita
Giampiero Rossi

Contatti
Via G. Reiss Romoli, 274, 10148 Torino
notiziariotecnico.redazione@telecomitalia.it

Registrazione
Presso il Tribunale di Torino n. 60 del 03/11/2021 - ISSN 2038-1921

Editoriale

È indubbio che l'Intelligenza Artificiale (IA) sia una delle soluzioni più innovative in continua crescita. Dappertutto si parla infatti di IA: ad oggi ben il 93% degli italiani ha già sentito parlare di intelligenza artificiale e il 55% afferma che è molto presente nella loro quotidianità.

In linea con questa tendenza, abbiamo voluto dedicare questo Notiziario Tecnico ad approfondire il nostro impegno sull'IA, consapevoli delle sfide tecnologiche, etiche e sociali che il suo impiego comporta.

Da quanto emerge da una ricerca **dell'Osservatorio Artificial Intelligence della School of Management del Politecnico di Milano**, in Italia, nel 2022, il mercato della IA ha raggiunto i 500 milioni di euro, con una crescita di ben il 32% in un solo anno, di cui il 73% commissionato da imprese italiane e il 27% rappresentato da export di progetti.

Il tasso di adozione nelle imprese evidenzia che la quota più indicativa rimane legata ai progetti di **Intelligent Data Processing** (34%) seguita dall'area che afferisce all'interpretazione del linguaggio, ovvero **Language AI** (28%).

L'adozione è in continua crescita, tanto che il 61% delle grandi imprese italiane ha già avviato almeno un progetto sull'IA, e tra queste, il 42% ne ha più di uno operativo. E noi di TIM siamo appunto tra queste! Il ruolo del Governo rimane centrale anche su questo tema, lo dimostra la nascita del *Comitato di coor-*

dinamento per aggiornamento della strategia nazionale per l'IA, che contribuirà a redigere la strategia nazionale sull'utilizzo dell'intelligenza artificiale.

Relativamente le Telco, la previsione di sviluppo dell'IA è molto rosea: secondo **Global Research** crescerà da 1,2 miliardi di dollari nel 2021 a quasi 40 miliardi di dollari entro il 2030.
Le soluzioni di intelligenza artificiale rappresentano, infatti, il futuro anche per migliorare i servizi di telecomunicazioni esistenti e per l'implementazione di assistenti digitali, e già da ora, da quanto si evince da un recente report di **McKinsey**, le Telco che hanno adottato le soluzioni di IA hanno registrato un CAGR dei ricavi a cinque anni di 2,1 volte superiore alle altre che non hanno fatto ricorso e un ritorno economico per gli azionisti di 2,5 volte maggiore.

L'accelerazione dell'uso di questa tecnologia solleva però importanti questioni etiche e di sicurezza, come evidenziato nella nostra intervista alla Professoressa Maria Rosaria Taddeo e ribadito dall'AI Act europeo di prossima attuazione.
Nuove funzionalità come il *living forever*, che fondono il Metaverso con l'IA, possiamo spingerci ad immaginare che una persona, attraverso il suo avatar dotato dei tratti della personalità di quello specifico individuo, potrà vivere all'infinito attraverso appunto il suo gemello digitale, che replicherà anche dopo la morte reale, in maniera indipendente, il modo di interagire con gli altri tipico di quella persona.

È dunque essenziale garantire che l'IA sia utilizzata in modo responsabile e conforme ai principi etici su due distinti piani: livello individuale, la sicurezza, la privacy e la protezione dei dati, e a livello sociale la giustizia, il lavoro e i diritti civili.

Vi auguro una buona lettura

Elisabetta Romano
Chief Network, Operations & Wholesale Office, TIM

Indice

- Un caffè con... Maria Rosaria Taddeo 8
- Il dibattito sulla responsabilità nell'uso dell'Intelligenza Artificiale 12
- Humane AI Net 24
- AI nei progetti internazionali 36
- Generative AI: la sfida di TIM per il futuro dell'IT 50
- Le opportunità offerte dall'AI ad un operatore Telco 66
- Large Language Model per il processo documentale TIM 72
- AI & Machine Learning per la rete TIM 82
- Verso una Greener AI 94
- Intelligenza Artificiale per la Compressione Video 108
- Glossario 120

TIM

Un caffè con...
Maria Rosaria Taddeo

A cura di Michela Billotti

Quando si parla di Intelligenza Artificiale è oggi sempre più inevitabile associarvi anche una riflessione di natura etica; prof.ssa Taddeo, ci vuole definire meglio di cosa l'etica digitale si occupi?

L'etica digitale nasce con l'avvento dell'ICT, quindi è una disciplina accademica che ha la sua storia decennale e indaga le questioni etiche, sociali e legali che le tecnologie digitali, inclusa l'Intelligenza Artificiale predittiva e generativa, comportano. Per farlo studia temi legati alla raccolta, cura, elaborazione e trasmissione dei dati; problemi inerenti al design, sviluppo ed uso degli algoritmi, per arrivare fino alle implicazioni di deontologia professionale legate all'uso delle tecnologie digitali.

Venendo adesso all'AI. È indiscusso che l'AI abbia un alto potenziale nello svolgere compiti, magari ripetitivi e che sia estremamente efficace nel "muoversi" velocemente nel mare magnum dei dati per dare output utili a raggiungere obiettivi complessi. È proprio questa la sfida che abbiamo di fronte: sfruttare l'enorme potenziale dell'AI per raggiungere obiettivi molto complessi a cui da soli non riusciremmo ad arrivare. Mi riferisco, ad esempio, all'uso dell'AI per predire e comprendere i cambiamenti climatici o per supportare ricerche in ambito biomedico, come per esempio la genomica.
Ma per raggiungere obiettivi così complessi ed importanti per l'umanità è necessario che si comprendano fino in fondo i rischi che l'AI può comportare, in modo da poterli poi mitigare. Uno dei rischi forse più conosciuti riguarda il così detto bias e le discriminazioni ingiustificate. Questo rischio deriva dai dati che usiamo per addestrare gli algoritmi di AI. Se i dati riflettono pregiudizi socio-culturali, allora l'AI produrrà output inficiati dallo stesso pregiudizio. Pensiamo al famoso caso di qualche anno riportato da Sweeney, in cui i servizi online per verificare le fedine penali apparivano più spesso nei risultati di ricerca per nomi identificabili con la comunità afroamericana rispetto ai risultati per ricerche con nomi identificabili come le altre comunità.
Questi sono rischi concreti che minacciano diritti umani e civili fondamentali. Ecco perchè l'etica dell'AI è cruciale. Senza, rischiamo che questa tecnologia mini le fondamenta della nostra società.

Quale è il problema del controllo dell'AI generativa?

Vede, se abbiamo un output problematico, magari con elementi falsi o discriminatori, come frutto di un processo in cui è intervenuto un AI, è difficile investigare quale criterio specifico abbia determinato quella risposta. Facciamo un esempio: se chiedo alla banca un muto, ma questo mi viene negato e la risposta mi viene data a valle di un esame fatto sul "mio caso" da un AI, sarà molto difficile se non impossibile sapere se il diniego sia legato al mio reddito basso a ri-

schio insolvenza del mutuo, oppure sia dovuto ad una discriminazione di genere e sono stata scartata perché donna.
L'AI è di fatto una tecnologia con poca trasparenza e che genera output che non possiamo predire con assoluta certezza. Questo tema della predicibilità è legato al problema del controllo. Entrambi i temi sono centrali quando si considera l'uso dell'AI per prendere decisioni di alto impatto. Il controllo di questa tecnologia è proporzionale alla certezza con cui prevediamo i suoi comportamenti. Questo dovrebbe essere un fattore tenuto in massima considerazione quando si pensa all'adozione di AI e quali compiti delegare a questa tecnologia.
Altra questione germana a quella del controllo è l'attribuzione della responsabilità morale (per certi versi anche legale) per le azioni di un sistema di AI. Questo per due ragioni, da un lato l'AI viene sviluppata in modo distribuito con gruppi di ingegneri, sviluppatori che lavorano per aziende o in posti diversi, riutilizzando pezzi di tecnologia prodotti da altri. Dall'altro lato, nella misura in cui l'AI apprende autonomamente come interagire con l'ambiente, può sviluppare comportamenti non previsti e non intesi dai sui programmatori o utenti e questo rende difficile attribuire la responsabilità per questi comportamenti.

Sul tema della responsabilità morale/legale ci sono ancora ampi margini di intervento, ma l'Europa con l'AI Act di prossima approvazione per la prima volta è apripista rispetto a tutti gli altri Paesi. Prof.ssa Taddeo, un suo commento in merito?

AI Act è il primo framework al mondo che guarda alla regolamentazione dell'AI direttamente, per misurare, mitigare e controllare il "rischio" di cui parlavamo poco fa, andando ad individuare modelli di rischio abbastanza "granulari" in questa prima fase, ma che sicuramente evolveranno con l'evolvere della AI. Reputo questo provvedimento necessario oltre che urgente. La parte che più mi interessa dell'AI Act è quella che si riferisce al conformity assessment. Credo che sia un'idea molto importante. Questo tipo di assessment è una sorta di audit dell'AI per capirne i rischi. È un'idea che finalmente cambia la prospettiva, passiamo da quella un pò banale per cui se progettiamo l'AI bene, allora non ci sono rischi etici, ad una per cui è importante monitorare e verificare l'uso dell'AI per capire di volta in volta quali rischi questa ponga.

Parliamo di AI e mondo del lavoro: se oggi è usuale assegnare ai sistemi intelligenti compiti ripetitivi per guadagnare tempo da dedicare ad altre attività, quali gli scenari nei prossimi anni?

Premetto che bisogna guardare con cautela alle stime che affermano che i sistemi di AI sono una minaccia per l'occupazione, perché queste stime si basano su modellizzazioni di intere professioni (per esempio l'avvocatura, il personale sanitario, gli autisti) che non possono che essere approssimative. Spesso queste stime confondono compiti, per esempio guidare un mezzo di trasporto pesante, con le professioni, per esempio essere un autista. Spesso deleghiamo all'AI un compito non un'intera professione.
Credo, sempre più che nei prossimi anni i professionisti saranno aiutati dai sistemi di AI nello svolgimento di attività pesanti, pericolose, monotone o che richiedono velocità esecutiva. Oggi ci sono già team ibridi di macchine AI ed esseri umani che lavorano insieme: un esempio il pilota automatico che collabora con il pilota umano nel decollo ed atterraggio di un aereo. L'importante è sfruttare al meglio la potenzialità dei sistemi di AI.

Un'ultima domanda, prof.ssa Taddeo: come vede il ruolo delle Telco con le AI sempre più pervasive?

Di fatto è un binomio inevitabile, nel senso che le AI per operare hanno bisogno di dati il che significa Data Center, bassa latenza e alta sicurezza. Il cloud computing diventerà indispensabile. Se non ci sono questi ingredienti la trasformazione digitale non si attua; ed è anche per questo che le Telco avranno un ruolo fondamentale, sia per i servizi B2B e progressivamente anche in quello B2C.
Le Telco possono essere gli abilitatori di questa trasformazione digitale.■

Autori

Michela Billotti *michela.billotti@telecomitalia.it*

Giornalista, direttore responsabile del Notiziario Tecnico TIM, è passata dal mondo delle lettere classiche, in cui si è laureata nel 1993, al settore delle telecomunicazioni. Da oltre vent'anni in Azienda ha dapprima collaborato all'organizzazione di eventi nazionali e internazionali, poi gestito i rapporti con i media interessati all'evoluzione dell'ICT; ora coordina i vari aspetti della comunicazione tecnica. È autrice di articoli e di libri sull'evoluzione del mondo delle telecomunicazioni scritti per un pubblico di "non addetti ai lavori". ■

Maria Rosaria Taddeo

Maria Rosaria Taddeo è Professor of Digital Ethics and Defence Technology presso l'Oxford Internet Institute dell'Università di Oxford, inoltre è Defense Science and Technology Fellow presso l'Alan Turing Institute. Il suo lavoro si concentra principalmente sull'analisi etica dell'intelligenza artificiale, della sicurezza e della difesa nazionale, dei conflitti informatici e dell'etica dell'innovazione digitale. La sua ricerca è stata pubblicata su importanti riviste come Nature, Nature Machine Intelligence, Science e Science Robotics.
È membro dell'Ethics Advisory Panel per il Ministero della Difesa inglese; inoltre, è membro della Scientific Advisory Board della Leonardo Foundation e del Scientific Advisory Board di Hi! Paris (l'istituto di ricerca avanza sull'intelligenza artificiale del Politecnico di Parigi) e membro Ethics and Data Governance Board de Italian Research Center on High-Performance Computing, Big Data and Quantum Computing.
Nel 2020 è stata elencata come una delle 100 donne più influenti nella tecnologia del Regno Unito da Computer Weekly. ■

Il dibattito sulla responsabilità nell'uso dell'Intelligenza Artificiale

Manuela Bargis, Giacomo Conti

Tutti i prodotti, compresi programmi ed applicazioni software, devono essere associati ad una responsabilità per i danni che possono causare a seguito di malfunzionamenti.
Il tema della responsabilità risulta particolarmente complesso per i sistemi di Intelligenza Artificiale (IA): chi risponde infatti per un'azione dannosa a seguito di decisioni di un sistema di IA? La causa del danno deriva da dati errati su cui si è basata la decisione di IA, da un malfunzionamento dell'algoritmo, da un uso improprio del sistema o da una sua evoluzione non prevista? Quando un sistema di IA generativa genera testo, audio o immagini a partire da elementi protetti da diritto d'autore si tratta di semplice ispirazione o di violazione di copyright?
I sistemi basati sull'Intelligenza Artificiale devono necessariamente gestire in modo appropriato la questione per poter godere della fiducia degli utilizzatori e dare avvio ad un pieno sviluppo ed utilizzo della tecnologia in diversi campi, compreso quello delle telecomunicazioni, e per innumerevoli applicazioni a beneficio di cittadini, imprese e amministrazione pubblica.
Il concetto è esplicitamente riconosciuto dalla Commissione Europea nella relazione sulle implicazioni dell'Intelligenza Artificiale [1], dove è sottolineata l'importanza di un quadro chiaro in materia di sicurezza e di responsabilità.

Le complessità derivanti dall'Intelligenza Artificiale

Vi sono alcune caratteristiche intrinseche dei sistemi di Intelligenza Artificiale che rendono particolarmente complesse la comprensione delle possibili cause di un danno e l'applicazione in generale delle norme sulla responsabilità.
Di conseguenza le vittime di danni causati dai sistemi di IA potrebbero andare incontro a difficoltà e costi nel risalire alla causa del danno per il quale richiedere il risarcimento e nel dimostrarlo, rischiando quindi di non essere adeguatamente risarcite.

I sistemi di Intelligenza Artificiale presentano sfide uniche in termini di responsabilità che emergono principalmente dalle peculiarità rappresentate in Tab.1 relative a: autonomia avanzata, opacità degli algoritmi, clausole di esclusione di responsabilità, molteplicità di attori, rischi di abuso.

L'importanza della responsabilità

Stabilire delle garanzie in caso di danni e problemi derivanti dall'utilizzo di sistemi basati sull'Intelligenza Artificiale contribuisce a creare un clima positivo sia per gli investimenti da parte dell'industria, sia per l'adozione della tecnologia da parte di cittadini ed imprese.

Secondo un sondaggio svolto da Ipsos Belgium e iCite per la Commissione Europea [2], pubblicato a luglio 2020, la responsabilità per i danni causati dall'IA rappresenta uno dei principali ostacoli esterni per le imprese europee nel proprio contesto di business e, restringendo il campo alle sole impre-

se che non hanno ancora adottato soluzioni di IA, è ritenuta il primo ostacolo in assoluto.
In particolare, come mostrato in Fig.1, costituiscono una sfida importante nell'adozione dell'IA da parte delle imprese ostacoli di tipo finanziario ed a livello di standardizzazione tecnica per lo scambio dei dati oltre che ostacoli legali legati in particolar modo alla responsabilità dei danni.

L'approccio europeo e le attività legislative in corso

In ambito europeo sono state avviate da alcuni anni e sono tuttora in corso iniziative legislative mirate a stabilire un quadro armonizzato di norme, con l'obiettivo di favorire l'evoluzione tecnologica garantendo al contempo il rispetto dei diritti degli individui. In particolare, per far fronte alle sfide derivanti dalla diffusione dell'Intelligenza Artificiale, è stato proposto dalla Commissione Europea ad aprile 2021 un nuovo regolamento sull'Intelligenza Artificiale, denominato AI Act, che mira a prevenire i danni e ridurre i rischi garantendo a cittadini ed imprese sicurezza e fiducia nell'utilizzo dell'IA grazie a regole flessibili e proporzionate a seconda dei rischi specifici posti dai sistemi di IA [3].
Considerato che le regole dell'IA Act, seppur mirate a minimizzare i rischi, non consentiranno di eliminare completamente i danni causati dall'uso di sistemi di IA, sono parallelamente in corso attività specifiche in tema di responsabilità per la revisione e l'adeguamento del quadro legislativo esistente. Le norme attuali sono basate da un lato sull'applicazione della direttiva sulla responsabilità per danno da prodotti difettosi [4] che risale a quasi 40 anni fa e che regola a livello europeo la responsabilità oggettiva del produttore di prodotti difettosi, e dall'altro sull'applicazione di norme nazionali sulle azioni e verifiche necessarie da parte del danneggiato per ottenere il risarcimento.

Tale attività di adeguamento delle norme sulla responsabilità ha portato all'adozione da parte della Commissione Europea a settembre 2022 di due proposte legislative attualmente in discussione presso le istituzioni europee [6]:

Tabella 1: Principali caratteristiche dei sistemi di Intelligenza Artificiale che aggiungono complessità al risarcimento per danni subiti

Peculiarità dei sistemi IA	Descrizione della complessità introdotta sulla responsabilità
Autonomia avanzata	I sistemi di Intelligenza Artificiale possiedono un'elevata capacità di autonomia: non si limitano a eseguire semplici comandi, ma possono adattarsi, evolversi e modificare le loro funzionalità basandosi sull'analisi e l'elaborazione di nuovi dati. Ciò rende la determinazione della responsabilità complessa, in quanto possono agire in modi non previsti inizialmente dai loro creatori.
Opacità degli algoritmi	Una delle sfide più significative dell'IA è l'opacità o "black box", che indica la difficoltà nel comprendere e prevedere completamente come un algoritmo di IA prenda determinate decisioni. Questo rende difficile determinare la causa di eventuali errori o malfunzionamenti e, di conseguenza, stabilire responsabilità. Le vittime di un danno potrebbero non essere in grado di far valere il diritto al risarcimento sia per mancanza di competenze tecniche e capacità analitiche sia per l'impossibilità di accedere all'algoritmo ed ai dati senza la collaborazione del potenziale soggetto responsabile.
Clausole di esclusione di responsabilità	Molti fornitori di servizi basati su IA includono clausole di sgravio di responsabilità nei loro termini di utilizzo. Queste clausole spesso escludono o limitano la responsabilità del fornitore in caso di malfunzionamenti o errori, spostando il peso della responsabilità sugli utenti o su terze parti.
Molteplicità di attori	L'ecosistema dell'IA è complesso e coinvolge una vasta gamma di attori, sviluppatori, fornitori di tecnologie e soluzioni, utilizzatori, utenti finali e fornitori di dati. Questa pluralità di parti interessate rende difficile identificare un singolo punto di responsabilità quando si verificano problemi.
Rischi di abuso	Gli utenti potrebbero sfruttare in modo improprio le capacità avanzate dei sistemi di IA per scopi illeciti o dannosi, creando dilemmi etici e legali, anche in relazione al copyright, su chi debba essere ritenuto responsabile per le azioni compiute dalla macchina.

Figura 1: Principali ostacoli esterni alle aziende nell'uso dell'Intelligenza Artificiale in UE secondo lo studio Ipsos-ICite "European enterprise survey on the use of technologies based on Artificial Intelligence"

External obstacles	Overall	Adopters	Non-adopters	Plan to use
Lack of public or external funding	36%	35%	38%	42%
Strict standards for data exchange	33%	34%	33%	37%
Liability for damage caused by artificial intelligence	33%	27%	38%	43%
The need for new laws or regulation	29%	29%	29%	34%
Lack of trust amongst citizens	28%	26%	30%	34%
Lack of access to high quality private data	27%	26%	27%	31%
Lack of access to or availability of public data	21%	20%	22%	24%
Reputational risks linked to using artificial intelligence	17%	16%	18%	19%

- la revisione della Direttiva sulla responsabilità per danno da prodotti difettosi [7] che aggiorna le regole della responsabilità adeguandole all'era digitale specificando la loro applicazione anche a software e sistemi di Intelligenza Artificiale;
- la proposta di una nuova Direttiva sulla responsabilità civile extracontrattuale dell'Intelligenza Artificiale (denominata AI Liability Directive) [8] che mira ad assicurare una migliore tutela per i danni causati dall'Intelligenza Artificiale intervenendo sull'onere della prova nelle richieste di risarcimento.

Quest'ultima proposta legislativa è espressamente legata alla responsabilità per i sistemi di Intelligenza Artificiale e mira ad armonizzare a livello europeo le norme sulla responsabilità civile extracontrattuale con l'obiettivo di rafforzare la fiducia nella tecnologia e garantire un regime di responsabilità efficiente in cui le richieste di risarcimento siano onorate.

Sul fronte economico sono promessi sostanziali benefici, con un potenziale aumento del valore del mercato dell'IA compreso tra 500 milioni e 1,1 miliardi di euro [8].

Approcci normativi sull'Intelligenza Artificiale extra UE

La concezione, l'utilità e l'utilizzo, nonché la regolamentazione dei sistemi di Intelligenza Artificiale sono profondamente diversi a seconda dei paesi del mondo, determinati da logiche storiche e geopolitiche.

Cina e Stati Uniti si trovano oggi agli antipodi di un possibile spettro normativo che simboleggia il modo di regolamentare l'Intelligenza Artificiale, mentre l'Unione Europea cerca di offrire una soluzione intermedia [11].

In Cina la tecnologia è sottoposta al controllo statale, ed è alla base di interessi primariamente economici e geopolitici: il governo cinese ha recentemente introdotto la necessità per le imprese di ottenere licenze specifiche nel momento in cui queste vogliano fornire servizi utilizzando IA generativa, ed offrire garanzie di integrità, sicurezza e ordine [12].
Il Governo intende assicurarsi che le informazioni generate dalle IA siano compatibili con i suoi interessi, ma non sembra voler frenare questo progresso, al contrario mira ad utilizzarlo strategicamente per potenziare la propria importanza a livello commerciale e tecnologico. In tema di responsabilità vi sono state discussioni sulla natura giuridica dell'IA, volte a valutare se trattasi effettivamente di un prodotto, cioè un bene commercializzato sul mercato e quindi soggetto alla normativa cinese sulla responsabilità per prodotti difettosi. In tal caso, l'onere della prova sarebbe a carico del ricorrente, che dovrebbe provare il difetto del prodotto, e dell'eventuale danno risponderebbe, per responsabilità oggettiva, il produttore [13].
Questa interpretazione è abbastanza vicina a quella europea, ma non considera la difficoltà, per l'utente finale, di provare il malfunzionamento del sistema di IA da lui utilizzato.

Negli USA, la tendenza è stata quella di evitare una regolamentazione diretta dell'IA attraverso normative centrali. Anche se predomina un sentimento di regolamentazione minima a livello federale, alcune agenzie come la Federal Trade Commission (FTC) hanno assunto un approccio proattivo verso specifiche sfide dell'IA, concentrandosi specialmente sulla trasparenza e sulle pratiche ingannevoli che influenzano i consumatori, in particolare nelle loro decisioni finanziarie. La FTC punta a sanzionare le imprese che forniscono informazioni fuorvianti sui loro prodotti [14], intervenendo ex-post nel mercato anziché attraverso normative preventive. In assenza di un quadro federale completo, i singoli Stati hanno affrontato singolarmente temi regolamentari legati all'IA. Questo approccio si è tradotto in un panorama normativo diversificato e frammentato: proprio ciò che l'UE intende evitare. Ad esempio, la città di New York impone verifiche sui pregiudizi per gli strumenti automatizzati di decisione sull'assunzione di personale, mentre in Colorado vigono leggi che proteggono dalle pratiche assicurative sleali guidate dall'IA.
Si evidenzia tuttavia che negli Stati Uniti, in ambito istituzionale a livello federale, sono state recentemente avviate iniziative ed azioni volte ad uno sviluppo ed utilizzo responsabile dell'IA a tutela dei cittadini e contro ogni discriminazione [15] ed il paese si sta mostrando incline a lavorare in modo collaborativo sull'IA, con una tendenza crescente ad allinearsi agli approcci internazionali, in particolare dell'Unione Europea.

Conclusioni

La diffusione di sistemi basati sull'Intelligenza Artificiale pone nuove sfide anche in termini di responsabilità, considerato che alcune peculiarità dell'Intelligenza Artificiale comportano maggiore difficoltà e complessità per il risarcimento dei danni causati dalla tecnologia. Il tema deve essere affrontato in modo da garantire lo stesso livello di protezione concesso in caso di danni relativi a tecnologie tradizionali ed allo stesso tempo ci siano le condizioni per lo sviluppo e l'innovazione tecnologica.

Nel contesto innovativo di impiego sempre più spinto dell'Intelligenza Artificiale, la Cina segue un approccio fortemente basato su un controllo statale ex ante, mentre gli USA un approccio basato su interventi ex-post e su regole frammentate per i singoli stati. In Europa si punta a trovare un equilibrio tra la protezione dei cittadini dai rischi legati all'IA e la promozione dell'innovazione da parte delle imprese; tramite l'elaborazione di un quadro legislativo armonizzato specifico per l'IA, che preveda anche una revisione delle norme in materia di responsabilità, si mira a favorire lo sviluppo e l'impiego della tecnologia nei diversi settori compreso quello delle telecomunicazioni.■

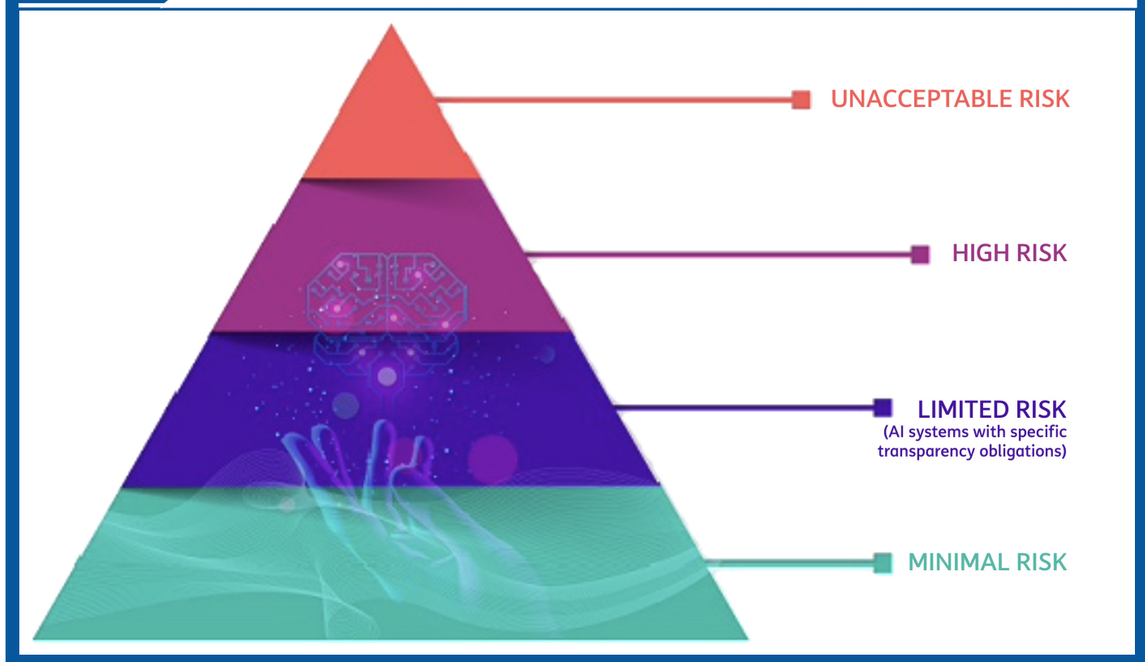
Nuovo AI Act: principali indicazioni e stato dei lavori

L'utilizzo dei sistemi di Intelligenza Artificiale nell'UE sarà regolamentato dall'AI Act, norma che affronta i rischi dell'IA e mira ad un ruolo di primo piano dell'Europa a livello globale sull'argomento.

Come già anticipato dal Libro Bianco sull'Intelligenza Artificiale nel 2020 [5], l'intento della Commissione è di consentire uno sviluppo affidabile e sicuro dell'Intelligenza Artificiale in Europa, nel pieno rispetto dei valori e dei diritti dei cittadini dell'UE.

La Commissione Europea si prefigge quindi l'ambizioso obiettivo di trovare un giusto equilibrio tra il bisogno di regolamentare il rapido sviluppo della tecnologia, soprattutto in termini di impatto e conseguenze sulla vita dei cittadini, e la necessità di non frenare l'innovazione. L'AI Act, un regolamento orizzontale (non settoriale), direttamente applicabile negli Stati membri senza bisogno di essere recepito nelle normative nazionali, detta le regole per lo sviluppo, l'immissione sul mercato e l'utilizzo di sistemi di Intelligenza Artificiale nell'Unione europea, adottando un approccio basato sul rischio. Sono inoltre stabilite misure per favorire l'innovazione, come ad esempio la creazione di spazi di sperimentazione normativa (sand boxes) o alcune agevolazioni per startup e PMI. L'AI Act definisce quattro categorie di sistemi di Intelligenza Artificiale, sulla base della finalità di utilizzo dell'applicazione di IA e dell'entità del potenziale danno e della sua probabilità, le quali saranno sottoposte a regimi differenti a seconda del rischio che presentano: rischio inaccettabile, rischio elevato, rischio limitato e rischio minimo (come mostrato in Fig. A). Saranno banditi i sistemi di IA che potrebbero essere utilizzati in modo intrusivo e discriminatorio, con rischi inaccettabili per i diritti fondamentali dei cittadini, la loro salute, la loro sicurezza o altre questioni di interesse pubblico, come ad esempio il riconoscimento facciale e l'identificazione biometrica, la manipolazione comportamentale, la polizia predittiva, il riconoscimento delle emozioni.

Figura A: Livelli di rischio definiti nell'AI Act – fonte della figura: webpage della Commissione Europea Regulatory framework proposal on artificial intelligence | Shaping Europe's digital future (europa.eu)



Le applicazioni considerate ad alto rischio saranno consentite subordinatamente al rispetto di determinati requisiti obbligatori (ad es. sistema di gestione del rischio, robustezza, trasparenza, controllo umano, etc.) e ad una valutazione di conformità ex ante e sono oggetto principale delle norme del regolamento. La proposta di AI Act definisce le aree in cui ci sarà una presunzione di rischio elevato, mentre la puntuale identificazione dei sistemi ad alto rischio è ancora oggetto di intenso dibattito tra le istituzioni.

Le applicazioni a rischio limitato saranno consentite ma sono soggette ad obblighi di trasparenza, mentre per le applicazioni a rischio minimo non è previsto alcun obbligo ex ante, anche se la Commissione incentiva l'adozione di codici di condotta per l'applicazione volontaria dei requisiti fissati per i sistemi di IA ad alto rischio.

Più recentemente, il dibattito si è concentrato sui cosiddetti General Purpose AI ed in particolare sull'AI generativa (es. ChatGPT), che secondo le istituzioni meritano delle regole ad hoc, tuttora in discussione.

Il percorso legislativo è iniziato ad aprile 2021 con la proposta della Commissione, seguita poi dalla posizione del Consiglio del dicembre 2022 e da quella del Parlamento di giugno 2023. I negoziati tra le istituzioni europee (cosiddetti triloghi) sono attualmente in corso e sono diversi i punti del testo ancora da definire. Si prevede che il regolamento possa essere approvato tra la fine del 2023 ed i primi mesi del 2024. Le nuove norme saranno applicabili dopo 24/36 mesi dall'entrata in vigore del regolamento.

Poiché la normativa potrebbe non essere applicabile prima del 2027, il Commissario per il Mercato interno Breton spinge le aziende del settore dell'IA ad impegnarsi alla compliance anticipata, con l'adesione volontaria ad un AI Pact, ad oggi in fase di definizione.

claudia.gerbino@telecomitalia.it

L’onere della prova e le proposte della nuova Direttiva AI Liability

Per “onere della prova” si intende la necessità di dimostrare certi fatti in un procedimento giuridico. Nel contesto dell’IA, riguarda la capacità della vittima di provare che un danno è stato causato da un sistema IA. L’“alleggerimento” dell’onere facilita le rivendicazioni dei danneggiati, ma dà oneri nei confronti dei produttori; al contrario, inasprire l’onere della prova rende più difficile per i danneggiati dimostrare le loro ragioni, ma consente un più ampio spazio di libertà imprenditoriale per i produttori.

Dall’indagine svolta dalla CE [9] contestuale alla proposta di AI Liability Directive è emerso che l’onere della prova in caso di malfunzionamenti dei sistemi di intelligenza artificiale è un tema fortemente dibattuto insieme al sistema di responsabilità da applicare. Tuttavia, è stato sottolineato dalla stessa Commissione che in questo momento di forte innovazione ed evoluzione tecnologica, risulta complesso e difficile, anche per gli esperti, discernere se un sistema di IA abbia funzionato male o se siano intervenuti altri fattori. Un quadro delle posizioni degli stakeholder è riportato in Tab.A.

Cercando una sintesi tra le parti, la proposta della CE suggerisce un approccio armonizzato che assicuri che qualsiasi tipo di danneggiato (persone fisiche o imprese) abbia un’equa possibilità di risarcimento qualora abbia subito danni causati da colpa o omissione di un fornitore, di uno sviluppatore o di un utente dell’IA.

Tabella A: Quadro delle posizioni degli stakeholder interessati su onere della prova e responsabilità

Stakeholder	Onere della prova	Responsabilità oggettiva
Cittadini, Organizzazioni di Consumatori, Istituzioni Accademiche	Forte supporto per qualsiasi misura legislativa che semplifichi l’onere della prova	Forte sostegno a norme armonizzate in tutta l’UE basate su responsabilità oggettiva eventualmente associata ad assicurazioni obbligatorie
Imprese produttrici di sistemi IA	Supporto per l’armonizzazione delle norme, ma scettici su un cambiamento completo dell’onere della prova	Pareri diversi; scetticismo su responsabilità oggettiva, percepita come “sproporzionata”

Di conseguenza la proposta di Direttiva si è concentrata sulla semplificazione del processo giuridico per le vittime di danni causati da sistemi di IA agendo in modo coordinato con la proposta di AI Act che prevede il rispetto di specifici requisiti a seconda del rischio associato al sistema.

E’ proposto un alleggerimento dell’onere della prova per il danneggiato introducendo, sotto determinate condizioni, la cosiddetta presunzione di causalità, grazie alla quale i danneggiati non dovranno spiegare in dettaglio come il danno sia stato causato da una determinata colpa o omissione, e prevedendo un diritto di accesso agli elementi di prova di imprese o fornitori, quando si tratta di IA ad alto rischio [10].

Bibliografia

1. Relazione COM(2020) 64 final della Commissione Europea al Parlamento Europeo, al Consiglio e al Comitato Economico e Sociale europeo sulle implicazioni dell'Intelligenza Artificiale, dell'Internet delle cose e della robotica in materia di sicurezza e di responsabilità - febbraio 2020 - disponibile al link EC Report 10022020 (europa.eu)
2. Studio condotto da Ipsos e iCite per la Commissione Europea pubblicato in data 28 luglio 2020 e disponibile al link European enterprise survey on the use of technologies based on artificial intelligence | Shaping Europe's digital future (europa.eu)
3. Proposta COM(2021) 206 final di Regolamento del Parlamento Europeo e del Consiglio che stabilisce regole armonizzate sull'Intelligenza Artificiale (legge sull'Intelligenza Artificiale) e modifica alcuni atti legislativi dell'Unione del 21.4.2021 disponibile al link EUR-Lex - 52021PC0206 - EN - EUR-Lex (europa.eu)
4. Direttiva 85/374/CEE del Consiglio del 25 luglio 1985 relativa al ravvicinamento delle disposizioni legislative, regolamentari ed amministrative degli Stati Membri in materia di responsabilità per danno da prodotti difettosi disponibile al link EUR-Lex - 31985L0374 - IT (europa.eu)
5. LIBRO BIANCO sull'Intelligenza Artificiale - Un approccio europeo all'eccellenza e alla fiducia pubblicato dalla Commissione Europea in data 19.2.2020 e disponibile al link White Paper on Artificial Intelligence: a European approach to excellence and trust (europa.eu)
6. Comunicato stampa CE "Nuove norme in materia di responsabilità per i prodotti e l'IA per proteggere i consumatori e promuovere l'innovazione" del 28 settembre 2022 disponibile al link Nuove norme in materia di responsabilità per i prodotti e l'IA per proteggere i consumatori (europa.eu)
7. Proposta COM(2022) 495 final del 28.9.2022 di Direttiva del Parlamento Europeo e del Consiglio sulla responsabilità per danno da prodotti difettosi disponibile al link EUR-Lex - 52022PC0495 - EN - EUR-Lex (europa.eu)
8. Proposta COM(2022) 496 final del 28.09.2022 di Direttiva del parlamento Europeo e del Consiglio relativa all'adeguamento delle norme in materia di responsabilità civile extracontrattuale all'Intelligenza Artificiale (direttiva sulla responsabilità da Intelligenza Artificiale) disponibile al link EUR-Lex - 52022PC0496 - EN - EUR-Lex (europa.eu)
9. Report conclusivo e contributi alla consultazione CE su "Adapting Civil Liability Rules to the Digital Age and Artificial Intelligence" (18 ottobre 2021 – 10 gennaio 2022) disponibili al link Product Liability Directive - Adapting liability rules to the digital age, circular economy and global value chains (europa.eu)
10. Domande e risposte sulla direttiva sulla responsabilità da Intelligenza Artificiale pubblicate dalla CE in data 28 settembre 2022 e disponibili al link Domande e risposte: direttiva sulla responsabilità da Intelligenza Artificiale (europa.eu)
11. Trustible, How Does China's Approach to AI Regulation Differ from the US and EU?, Forbes, 18 luglio 2023, https://www.forbes.com/sites/forbeseq/2023/07/18/how-does-chinas-approach-to-ai-regulation-differ-from-the-us-and-eu/
12. PCCN.com, Interim Measures for the Management of Generative Artificial Intelligence Services officially implemented, TMTnewsflash, Tiang&Partners, agosto 2023, https://www.pwccn.com/en/tmt/interim-measures-for-generative-ai-services-implemented-aug2023.pdf
13. R.Cai, X. Chen, S. Zhang, Legal Implications of Artificial Intelligence in China, The LegalTech Book, 20 luglio 2020. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119708063
14. Section (5)(a) dell'FTC Act, https://www.ftc.gov/legal-library/browse/statutes/federal-trade-commission-act
15. Annuncio del 4 maggio 2023 dell'Amministrazione Biden-Harris su "New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety" disponibile al link FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety | The White House

Autori

Manuela Bargis *manuela.bargis@telecomitalia.it*

Laureata in Ingegneria Elettronica presso il Politecnico di Torino ed in Ingegneria delle Telecomunicazioni presso la Universidad Politecnica di Madrid, nel 1999 entra a far parte dell'azienda conseguendo nel 2001 il master ICT organizzato da Telecom Italia in collaborazione con COREP/SSGRR. Ha svolto attività su architetture di rete e protocolli di segnalazione nell'ambito di enti di standardizzazione internazionali (ETSI ed ITU) e di progetti internazionali e si è occupata fin dal 2002 di aspetti tecnici della regolamentazione. Attualmente si occupa della coerenza tecnico-normativa di aspetti innovativi legati all'evoluzione delle reti, è il riferimento di progetti svolti in collaborazione con il Politecnico di Torino inerenti l'etica e la governance di tecnologie innovative ed è delegata in 5GAA su aspetti normativi legati a scenari di mobilità connessa. ■

Giacomo Conti *giacomo.conti@polito.it*

Laureato in Giurisprudenza e appassionato di informatica, riveste attualmente il ruolo di Project Manager presso il Centro Nexa del Politecnico di Torino. Il suo ambito primario di competenza è la responsabilità nell'uso del software, lo studio della sua natura, ed i problemi di privacy derivanti dalle nuove tecnologie digitali. Amante dei videogiochi, è fondatore del sito di critica e approfondimento MMO.it.
È iscritto all'Ordine dei Giornalisti del Piemonte dal 2016. ■

Humane AI Net

Raffaele de Peppe

L'Intelligenza Artificiale (AI) è destinata ad avere un forte impatto su tutti i settori dell'economia compreso quello delle Telecomunicazioni. Nell'ambito del progetto europeo *Human AI Net* ed in collaborazioni con altri enti quali ETNO e l'Istituto nazionale tedesco per la ricerca sull'AI, la GSMA ha prodotto una *Research Agenda* per la industry che contiene i temi di ricerca sull'AI che potranno massimizzarne l'impatto sia nel nostro settore che in quelli di applicazione delle Telecomunicazioni . Il documento è stato prodotto dalla Task Force GSMA "AI for Impact" partecipato da TIM ed altri operatori all'avanguardia sull'adozione dell'AI quali Telefonica, Telenor, Orange, Telia, Telstra e Vodafone.

L'AI produrrà dei benefici sostanziali sul nostro settore se verrà introdotta in maniera etica e responsabile in maniera regolamentata tale da minimizzare effetti e derive indesiderati.
L'*AI Act* prodotto dalla Commissione Europea in fase di approvazione sarà basato su un approccio *risk based*, ovvero tenderà a regolamentare le applicazioni dell'AI a seconda del loro grado di rischio.
Le applicazioni con grado di rischio inaccettabile, come il *social scoring*, saranno espressamente vietati mentre applicazioni ad alto rischio saranno soggetti a verifica di conformità ex ante.
Le applicazioni con rischio medio avranno solo obblighi di trasparenza verso il cliente ed infine le applicazioni a basso rischio saranno comunemente accettate.
Altri aspetti relativi all'AI sono oggetto di attenzione da parte di enti sovranazionali come l'OECD ed il G7.

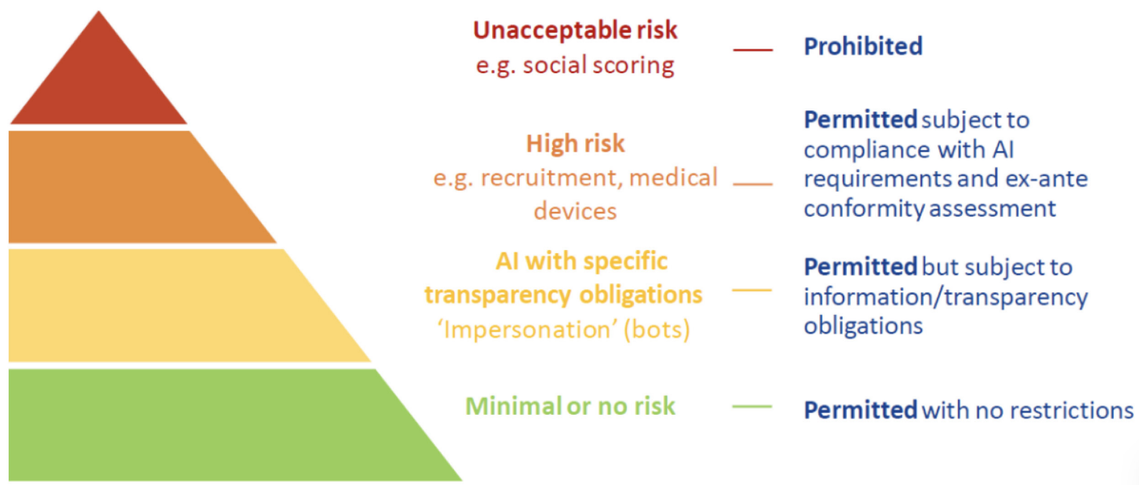
L'AI per le telecomunicazioni mobili, una realtà già presente

L'AI viene sperimentata dagli operatori mobili per migliorare la loro efficienza operativa sia in termini di **prestazioni di rete** che di **relazioni con il cliente**. Questi due campi di applicazione sono quelli a maggior impatto economico (fig.2).
Oltre a vantaggi di natura tattica sul mercato, gli operatori stanno valutando anche un posizionamento più strategico sull'AI con lo sviluppo di nuovi business per intercettarne il valore di mercato previsto in forte crescita.

Le Applicazioni di Rete

Il progetto Humane AI ha messo in evidenza i principali campi di applicazioni dell'AI da parte degli operatori di telecomunicazioni.
Le **applicazioni di rete** trovano spazio sia nell'ingegneria di rete che nella ge-

Figura 1: Risk based approach dell'AI Act



stione operativa dell'infrastruttura fissa e mobile. Un esempio nel mobile è dato dalla **pianificazione di rete**, dove l'AI viene utilizzata per la predizione dei livelli di traffico per adeguare capacità e copertura delle reti mobili - ciò permetterà agli operatori di fare investimenti solo dove necessari ("smart capex") generando al contempo un miglioramento della customer experience legata ai servizi.
Nell'ambito della **network optimisation** l'AI viene usata per analizzare i dati di traffico, per individuare quei fattori che impattano negativamente sulle prestazioni e permettere all'operatori di ottimizzare le configurazioni di rete anche in maniera predittiva.

Sempre più rilevante per gli operatori è risultato l'ambito **energetico**. L'AI viene già utilizzata sia per ottimizzare l'uso di energia nelle reti che per determinare la miglior fonte energetica da utilizzare.
Sempre sul versante delle reti, l'AI è sempre più decisiva nella lotta al **cybercrime**, permettendo di individuare con maggior accuratezza i pattern di attacco verso le reti di telecomunicazioni tramite l'analisi delle anomalie nel traffico, consentendo agli operatori di mettere in campo le contromisure più efficaci nella lotta alle frodi.

Le applicazioni di Marketing
Sul fronte **marketing** l'AI trova sempre più spazio per migliorare l'interazione con il cliente e quindi aumentarne l'engagement. Gli operatori stanno esplorando l'AI per implementare il paradigma NDA – *Next Best Action* – in cui imparando dalla storia passata del cliente l'operatore può determinare un ventaglio di azioni di marketing e quindi scegliere quella più efficace per il singolo cliente (ad esempio aiutando gli operatori a capire le domande dei clienti sui diversi canali ed il loro contesto per fornire loro la miglior risposta possibile).
L'evoluzione naturale del NBA è il *Next Best Experience* (NBE) che mira ad offrire la miglior experience con il proprio operatore attraverso l'analisi dei dati della customer journey per ogni cliente.
Altro campo di applicazione marketing dell'AI messo in evidenza dal progetto è quello del *churn prediction*.
La capacità dell'AI di imparare dalla storia pregressa del cliente permette di individuare quei pattern comportamentali che danno tipicamente origine al cambio di operatore, permettendo all'operatore di intraprendere preventivamente possibili azioni di retention.

Lo *smart pricing* costituisce un'evoluzione del campo di intervento dell'AI nel marketing - l'analisi del customer *behaviour* associato a quello dei market trends e all'analisi dei competitors permette di ottimizzare il pricing per determinati segmenti di cliente, per favorire un *pricing dinamico* piuttosto che statico sino ad offrire un *personalized pricing*.

Il *credit scoring* per la valutazione del rischio di insolvenza di un cliente è una interessante applicazione dell'AI – l'analisi di utilizzo dei servizi di telecomunicazioni permette di individuare pattern che possono dare indicazioni di *creditworthiness* di un utente. L'uso congiunto di tecniche di AI quali il

VALORE ECONOMICO DELL'INTELLIGENZA ARTIFICIALE

Al 2022 il mercato globale dell'AI era stimato a 87 miliardi $. Si prevede crescere con un CAGR del 36,2% sino al 2027 quando raggiungerà i 407 miliardi $, quintuplicando il suo valore in soli 5 anni. Il mercato dell'AI si compone di quello del Hardware (chipset, processori, schede di rete) e del Software (applicazioni, APIs, machine learning) che rappresentano circa 1/3 e 2/3 del valore complessivo mentre la componente servizi professionali è trascurabile.
L'AI in cloud rappresenta circa il 75% del valore della parte SW, la big industry adopter è predominante nell'adozione rispetto al mondo SME che invece è quello a maggior crescita.

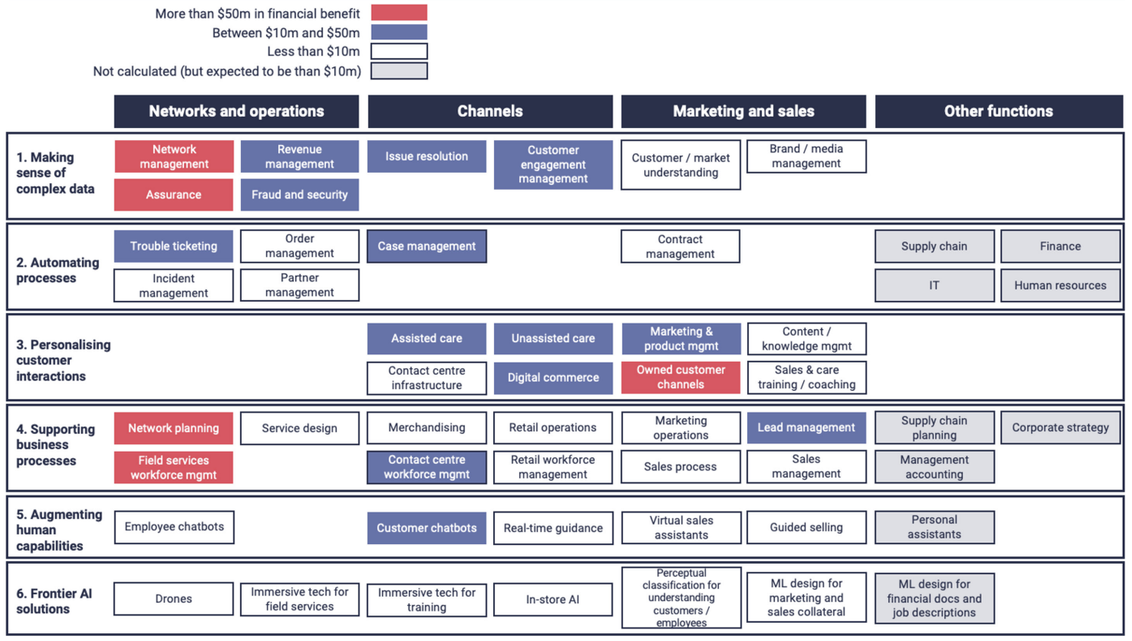
A livello di funzioni di business, è l'area di Marketing & Sales il maggior adopter dell'AI mentre le restanti funzioni registrano un livello di adozione comparabile, a dimostrazione che l'AI costituisce un potente abilitatore di *digital transformation* delle aziende.

L'AI viene adottato trasversalmente su tutti i settori industriali - i settori con maggior spesa sull'AI sono il settore Finance (banche, istituti finanziari, assicurazioni), il Retail (eCommerce) e il settore Healthcare

In Italia il valore di mercato al 2022 era stimato a 1,1 miliardi di euro e si prevede crescere sino a 6,6 miliardi di euro al 2027.

Fonte: Markets & Markets (2022)

Figura 2: STL Partners "AI & automation for telcos: Mapping the financial value



reinforcement learning ed il *deep learning* permettono di raccomandare una sostituzione del device con il modello più adatto al cliente al momento più opportuno. Il *collaborative filtering* ed il *content based filtering* sono metodologie basate su AI che permettono di imparare dalla storia di acquisto di un cliente e dalle sue preferenze per azioni di *service & product reccomandation* che va oltre la raccomandazione del singolo device.

Infine, il *Generative AI* è considerata la nuova frontiera per i *chatbots* ed i *digital assistants* già largamente in uso nella industry per offrire un customer caring 24x7. Le applicazioni AI di maggior successo nel campo del caring sono l'ottimizzazione delle FAQs ed i servizi di risposta personalizzati in base ai dati del cliente – l'AI permette di analizzare rapidamente una richiesta di un cliente su un canale, fornire una prima risposta in linguaggio naturale e solo nel caso di incomprensione di scalare la richiesta ad un operatore di un customer care.

Monetizzazione esterna dell'AI

Come messo in evidenza dal riquadro sul valore economico dell'AI, è possibile pensare ad un posizionamento più strategico degli operatori sull'AI orientato allo sviluppo di nuovi business.
Ad oggi non è stato definito un business model consistente per la monetizzazione esterna dell'AI in attesa del consolidamento di un quadro normativo di riferimento. In questo contesto il progetto indica nelle partnership pubblico private la miglior opportunità di monetizzazione esterna dell'AI.
L'operatore di telecomunicazioni potrà offrire prodotti AI (es. in cloud) da abbinare a servizi di analytics basati sui dati generati dalle reti che gli enti pubblici possono valorizzare a supporto di policy in tema quali la sostenibilità ambientale legate al cambio climatico, salute, trasporti pubblici e protezione civile.
Gli operatori possono aggregare ad anonimizzare i dati dei propri clienti, che i loro terminali creano ad un ritmo di 200 - 400 data points al giorno per fornire insights accurati che l'AI può generare nei diversi campi di pubblica utilità precedentemente menzionati.

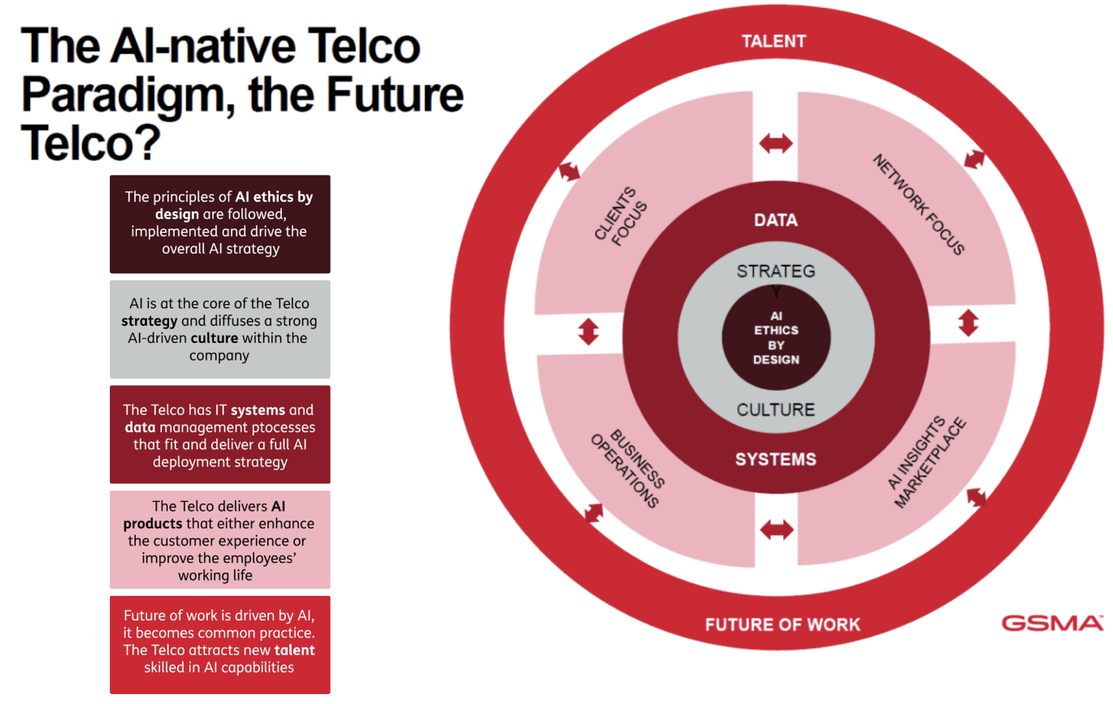
Nuovi orizzonti dell'AI per le Telecomunicazioni

La *research agenda* definita in GSMA indica i temi di ricerca per individuare nuovi campi di applicazione dell'AI per il settore delle telecomunicazioni. Secondo una survey tra le aziende partecipanti è emerso che entro 5 anni l'AI sarà adottato su vasta scala tra i telco diventando parte integrante di tutti i processi di business di un operatore ("AI everywhere").
L'AI cambierà anche il modo di lavorare ed è destinata ad essere la principale piattaforma di innovazione per i telco secondo il paradigma *AI native telco* (Fig.3).

Al centro del paradigma AI native telco la GSMA mette l'AI *Ethics by design*, ovvero l'incorporazione di regole etiche nell'uso dell'AI come peraltro già recepito da diversi codici etici delle aziende partecipanti al progetto.
A partire da principi etici, l'AI dovrà diventare parte integrante della cultura aziendale e quindi influenzarne le strategie di business.

Ma per poter mettere a terra il concetto di AI native telco, gli operatori dovranno investire maggiormente nelle loro *data foundations* ovvero il loro impianto di dati che saranno il materiale grezzo su cui innestare gli algoritmi AI.
I dati per l'addestramento dei sistemi AI dovranno essere di *alta qualità*, *diversificati* e *rilevanti* oltre ad avere una *governance* che assicuri privacy, trasparenza ed accountability dei sistemi AI tali da poter capire e spiegare il loro funzionamento e quindi eliminare effetti indesiderati quali il bias[1].
Un data foundation consistente si ottiene in base ad un insieme di processi comunemente definiti – *data collection, data cleaning, data integration, data annotation, data governance, storage & management*.
Molti di questi processi sono gestiti ancora in maniera manuale e quindi dovranno essere oggetto di automazione da parte dei telco al fine di rendere l'enorme mole di dati prodotti dalle reti e dai sistemi di CRM *actionable* per i sistemi AI.

Figura 3: AI Native Telco (GSMA)



Note

(1) Il problema del bias e le strategie di mitigazione viene trattato più avanti

Il progetto mette anche in evidenza il problema dei silos tra dati di rete e di CRM e quindi la necessità di maggior collaborazione tra i teams aziendali che li gestiscono.

I dati per addestrare sistemi AI sono particolarmente critici dal punto di vista della **privacy** e quindi della protezione dei dati personali del cliente per la quale è prevista la compliance alle varie regolamentazioni locali e regionali. L'adempienza a principi di privacy sarà critica e potrà essere ottenuta grazie ad algoritmi di anonimizzazione.

Il tradeoff tra rispetto della privacy e utilità dei dati è uno degli elementi di ricerca più importanti emersi nel progetto per avere in futuro sistemi AI *human centric*. La ricerca sarà focalizzata sull'uso di **dati sintetici** ovvero dati creati artificialmente (anche dai stessi sistemi AI) che simulano i dati reali. Gli ambiti di ricerca più rilevanti emersi su questo tema sono la *rappresentatività*, l'eliminazione di *bias* insiti nella loro generazione e quindi le *tecniche di valutazione* della loro bontà.

Un numero crescente di operatori tra cui la stessa TIM ha in esercizio sistemi AI ma il campo di applicazione rimane ancora limitato ai casi di utilizzo previamente descritti. Per rendere pervasivo l'uso dell'AI in tutti gli ambiti di applicazione sarà necessario determinare azioni di **scaling** sull'AI. A tal fine il progetto ha individuato nella *standardizzazione* degli use case AI più comuni e nell'applicazione del concetto di *DevOps* all'AI (*MLOps*) le vie percorribili per scalare l'uso dell'AI in tutti gli ambiti aziendali di un telco.

Il progetto indica nuove applicazioni dell'AI in ambito rete che hanno bisogno di ricerca. Ad esempio, nella gestione della **5G core** e **RAN** si dovrà fare più ricerca sull'automazione di alcuni tasks che limitino l'intervento umano e quindi sulla *pedictive maintenance* basata su AI che possa ridurre sensibilmente i guasti e gli interventi di riparazione.
Il progetto indica nella configurazione automatica di rete (*Self Organizing Network*) per limitare la necessità di configurazioni umane, nel *traffic optimisation* per l'overprovisioning di rete grazie all'allocazione di risorse di rete just in time i temi di ricerca prioritari.
La **network automation** sarà complessivamente una area di ricerca sull'AI a se stante, in ottica *zero touch management* per ridurre al minimo l'intervento umano nella gestione delle rete ed ottimizzarne gli OPEX.

Lo **smart energy management** è emerso come importante tema di approfondimento per l'AI che può diventare una arma per fare saving su una voce di OPEX sempre più critica per i telco. In questo contesto si dovrà approfondire l'uso dell'AI per predire i consumi energetici in base alle predizioni di traffico e quindi operare predittivamente sulla configurazione di rete.
L'ottemperanza al quadro regolatorio fornisce ulteriori temi di ricerca.
L'**explicability** dell'AI è un tema che molti operatori dovranno affrontare in particolare sui mercati europei. Uno dei temi più rilevanti sarà il passaggio dalla **correlazione** (relazioni tra variabili) alla **causation** (influenza causale tra variabili) nel machine learning, dovendo poter inferire sulla causalità tra variabili per aumentare l'explicability dei sistemi AI. Il progetto indica nella ricerca in campi quali *controlled experiments* (si cambia una variabile e si vede l'effetto su una altra), il *time order* (causalità temporale tra variabili), il *mechanism explanation* (trovare meccanismi plausibili di causalità tra variabili), la *reverse causality* (se c'è causalità inversa tra variabili) e l'*additional data* (aggiungere dati per individuare causalità) l'individuazione di possibili soluzioni.
Molti dei temi di ricerca individuati si possono inquadrare come sviluppo di **digital twins** delle reti, con gli algoritmi AI applicati ad una copia digitale prima ancora che fisica della rete. Il digital twin di una rete potrà essere usata nei vari ambiti di applicazioni precedentemente esaminati, come network optimisation, predictive maintanance ed il network planning.
Infine, sempre in ambito rete un tema di ricerca fortemente attenzionato dagli operatori è quello del **Network as a Service** (NaaS) in cui funzionalità di rete possono essere virtualizzate e accesse tramite APIs. Servizi di NaaS già in essere riguardano servizi molto utilizzati come VPN o SDN ma che potranno essere migliorati attraverso l'uso dell'AI.

Ma la ricerca sullo sviluppo dell'AI per i telco non riguarderà solo la rete ma toccherà anche il **marketing** e le **operations** di business. In questo campo un tema di ricerca prioritario riguarderà il paradigma del **real time AI** ovvero l'applicazione dell'AI ai dati in tempo reale ovvero al tempo della produzione stessa dei dati, per poter intraprendere azioni tempestive come offerte in real time in base alle esigenze estemporanee del cliente. Il Machine Learning verrà esplorato nella dimensione dell'**ottimizzazione** dei processi di business – in questo contesto la ricerca riguarderà temi quali il *model based optimisation* in cui ad essere ottimizzato sarà un modello del sistema, l'*heuristic optimisation* per ottimizzare la ricerca operative nella soluzione di problemi, il *reinforcement learning*, il *bayesian optimisation* sino ad arrivare al *multi objective optimisation* in cui ad essere ottimizzati sono diversi obiettivi di business secondo un trade off tra gli stessi obiettivi.
Nel campo del **CRM** si tenderà a migliorare in particolare le **chatbots**. Lo sviluppo dell' AI riguarderà la **proattività nell'interazione** (per anticipare i needs del cliente) e la **capacità di dialogo**.
Sul versante della proattività si tratterà di definire il giusto bilanciamento tra personalizzazione della relazione e spam, per poter fornire in anticipo al cliente le informazioni utili ma senza sovraccaricarlo di messaggi. Per quanto riguarda la capacità di dialogo invece la ricerca riguarderà soprattutto l'applicazione di modelli di Generative AI che possano andare oltre il modello FAQ ed instaurare un vero dialogo con il cliente in linguaggio naturale.
L'applicazione di generative AI verrà valutata anche per fornire chatbots di tipo *multilanguage* per servire clienti nel segmento etnico o i clienti in roaming.

Una nuova dimensione della ricerca sull'AI applicata al mondo telecomunicazioni riguarderà l'**Ethical & Responsible Business** nell'ambito della *Corporate Social Responsability* (CSR). L'AI può avere sia aspetti *positivi*, le-

gati alle applicazioni di forte impatto sociale come la salute o l'ambiente, ma anche aspetti fortemente *negativi* come il rischio di perdita della privacy o la discriminazione dovuti al bias insiti in sistemi di ML. La GSMA ha prodotto a riguardo un *AI Ethics Playbook*[2] come manale d'uso per i telco per un uso etico dell'AI ma anche un *Self Assessment Questionnaire* (SAQ) per una autovalutazione sulla compliance raggiunta sui temi etici.
La ricerca riguarderà in particolare le modalità di valutazione del rischio associato a determinati caso d'uso coerentemente con i dettami dell'AI Act europeo. Un fattore di rischio severo è quello della violazione della *privacy* – la ricerca sarà orientata a sviluppare sistemi AI e di ML che minimizzano il rischio di violare la privacy quando i sistemi vengono allenati con i dati dei clienti.
Altro fattore con forte impatto negativo per un operatore è quello del *bias* (polarizzazione) dei sistemi AI che può essere di diversi tipi: *bias di rappresentazione* (i dati con cui viene addestrata la macchina AI riguarda una popolazione diversa da quella su cui verrà applicato il sistema), *bias di misura* (quando i dati raccolti contengono insito già un bias), *bias dell'algoritmo* (quando l'algoritmo AI è progettato con un bias insito).
La ricerca in questo campo riguarderà quindi l'eliminazione o la mitigazione di qualsiasi forma di polarizzazione dei sistemi AI basato su tre metodologie: il *pre processing* con il quale i dati di addestramento sono pre processati prima di essere usati per l'addestrato per eliminare in anticipo forme possibili di bias, *in-processing* in cui la depolarizzazione avviene introducendo funzioni di *fairness constrain optimisation* contestualmente all'addestramento ed infine il *post processing* in cui il bias viene corretto solo a posteriori e non durante l'addestramento della macchina.
La *explanability* dei sistemi AI è un tema di ricerca trasversale e necessita di diversi approcci a seconda del tipo di sistema AI.
I sistemi di tipo *black box* presentano alte prestazioni negli outputs a fronte di una bassa trasparenza sul modello di funzionamento, i sistemi *white box* presentano al contrario maggiori qualità di explanability a fronte di prestazioni in uscita minori. Esistono anche modelli AI intermedi di tipo *grey box* con diversi tradeoff tra prestazioni ed explanability. Metodologie come LIME o SHAP permettono un certo grado di interpretabilità di sistemi di tipo black box che sono quelli più critici.

Infine, sarà necessaria molta ricerca sul tema della **sostenibilità** dell'AI considerando che ad oggi il consumo energetico dei grandi sistemi AI/ML comportano un consumo di energia proibitivo. Il tema di ricerca, il *Green AI*, è molto vasto ed esula dal campo della TLC.
SI è valutato che l'addestramento di un modello GPT ha le stesse emissioni CO2 del ciclo di vita di 5 autovetture con motore termico.

La ricerca riguarderà modelli di chipset dedicati a minor consumo ed algoritmi AI con minimo consumo energetico *by design*.
Dall'altro lato l'AI può supportare la sostenibilità in altri settori – **AI for Sustainibility** – contribuendo in maniera positiva alla sostenibilità ambientale, sociale ed economica. Su questo tema sono attive iniziative come AI4I (AI for Impact) della GSMA o AI4Good (AI for Good) delle Nazioni Unite.■

Note

(2) https://www.gsma.com/betterfuture/wp-content/uploads/2022/01/The-Mobile-Industry-Ethics-Playbook_Feb-2022.pdf

Acronimi

AI	Artificial Intelligence
API	Application Programming Interface
CRM	Customer Relationship Management
CSR	Corporate Social Responsibility
ETNO	European Telecommunications Network Operators' Association
FAQ	Frequently Asked Questions
GSMA	GSM Association
HW	Hardware
ML	Machine Learning
NDA	Next Best Action
NDE	Next Best Experience
OECD	Organization for Economic Co-operation and Development
SAQ	Self Assessment Questionnaire
SW	Software
ZTM	Zero Touch Management

Autori

Raffaele de Peppe *raffaele.depeppe@telecomitalia.it*

Raffaele de Peppe si è laureato presso la facoltà di Ingegneria “La Sapienza” nel 1996 ed ha ottenuto un Executive MBA presso la Business School IE di Madrid nel 2002.
Ha lavorato per l’area Internazionale di TIM presso le controllate estere in Serbia, Brasile e Spagna dal 1997 al 2002 nell’ambito dello sviluppo del business sui mercato esteri. Successivamente ha operato in diversi dipartimenti di Innovation e Strategy rappresentando TIM presso organismi internazionali come la GSMA e nel Board delle associazioni europee per la ricerca nelle telecomunicazioni mobili come 5GIA e 6GIA nell’ambito delle varie partnership pubblico private con la Commissione Europea.
Ha gestito un programma di liaison industriale con il MIT di Boston dal 2004 al 2019 a supporto della trasformazione del business oltre i servizi tradizionali attraverso l’esplorazione e la valutazione strategico dei nuovi paradigmi digitali. ■

AI nei progetti internazionali

Jovanka Adzic, Mauro Boldi, Roberto Fantini

In questo articolo si esaminano le attività sull'Intelligenza Artificiale (AI) nell'ambito dei progetti europei finanziati a cui TIM partecipa. Attualmente molti di questi progetti sono finanziati nel programma quadro Horizon Europe, all'interno del quale è stata creata una cosiddetta "Joint Undertaking" (JU), ovvero una collaborazione congiunta pubblico-privata, dedicata alle attività su "Smart Networks & Services" (SNS). TIM è membro della componente privata, facendo parte dell'associazione 6GIA, che si interfaccia con il direttorato DG-Connect della Commissione Europea per i progetti della JU SNS. Tra i progetti finanziati dalla JU SNS in Horizon Europe molti studiano l'uso delle metodologie AI/ML in quanto tali, oppure in quanto funzionali alla definizione delle future generazioni di sistemi di telecomunicazione, come ad esempio il 6G.

L'introduzione di metodologie di Intelligenza Artificiale (AI) nelle reti è avvenuta ormai da tempo, a partire dall'ampia diffusione di queste metodologie a più ampia scala: studi e valutazioni di prestazioni si sono avviati per determinare se tali metodologie potessero portare vantaggi anche nell'ambito del 5G ed ora del 6G. Si veda ad esempio quanto pubblicato recentemente nel libro che definisce i principali filoni su cui si evolverà il 6G, dove l'AI è considerata un pilastro fondamentale della architettura del nuovo sistema.
In particolare, la "Network Automation" è considerata come un elemento fondamentale per le future reti, sia per ragioni di soddisfacimento dei requisiti (KPI, Key Perfomance Indicators), sia per soddisfare nuovi "valori" (KVI, Key Value Indicators) che le nuove reti devono introdurre, quali la maggiore efficienza, la sostenibilità complessiva, la più generale conformità ai bisogni dell'utenza. Nel dettaglio, si riportano le attività svolte nei progetti Hexa-X ed Hexa-X-II, che sono considerati i progetti capofila ("flagship") per la definizione del sistema 6G, e nel progetto AI@EDGE, che studia più nel dettaglio soluzioni di AI/ML al cosiddetto "Edge" della rete, ovvero in posizione decentralizzata e più vicina all'utente del servizio.

L'AI nella definizione del sistema 6G

Il progetto europeo Hexa-X (Fig.1) ha rappresentato un punto di riferimento cruciale per lo sviluppo del futuro sistema di comunicazione 6G. Finanziato dal programma Horizon 2020 dell'Unione Europea, questo progetto ha gettato le basi per la definizione delle reti di comunicazione 6G, affrontandola con approcci all'avanguardia.
Il Work Package 4 (WP4) di Hexa-X, intitolato "AI driven communication and computation co-design", si è concentrato sull'applicazione dell'Intelligenza Artificiale (Artificial Intelligence - AI) all'interno del sistema 6G,

Figura 1: Il progetto Hexa-X

focalizzandosi sulla progettazione di una nuova interfaccia radio basata sull'AI. Il progetto ha identificato alcune *aree tematiche*, su cui si sono focalizzate le attività dei diversi partner, evidenziate nei riquadri azzurri presenti in Fig.2 (dove si vede anche l'architettura complessiva).

Le cinque aree tematiche per l'AI/ML nel 6G identificate da Hexa-X sono:

- **miglioramento delle prestazioni della rete tramite AI/ML nel 6G**. Le attività proposte in Hexa-X in questo ambito sono finalizzate a migliorare l'affidabilità delle comunicazioni, aumentare la velocità della trasmissione dei dati, garantire una maggiore efficienza nell'uso dello spettro e ottimizzare la progettazione delle reti. Dal punto di vista dell'architettura queste soluzioni sono connesse sia al livello di infrastruttura che al livello di servizio di rete, in quanto mirano a ottimizzare gli algoritmi di elaborazione radio per migliorare le prestazioni delle comunicazioni, introdurre soluzioni energeticamente efficienti e supportare nuovi concetti per l'accesso radio come il Distributed MIMO (D-MIMO) e le Reconfigurable Intelligent Surfaces (RIS);
- **Gestione e Orchestrazione End-to-End Intelligente**. Le tecniche basate sull'AI/ML possono incrementare le prestazioni complessive di rete anche tramite una più accurata gestione e orchestrazione delle reti. In questo ambito il progetto Hexa-X ha presentato degli abilitatori tecnologici basati sull'AI/ML che in Fig.2 si posizionano nel livello trasversale dedicato alla gestione e l'orchestrazione di rete (M&O). Sono state presentate soluzioni predittive per migliorare gli aspetti dell'orchestrazione, e soluzioni decentralizzate per migliorare la scalabilità dei componenti architetturali;
- **il 6G come una piattaforma per l'AI**. Obiettivo cruciale è rendere il 6G una vera e propria "piattaforma per l'Intelligenza Artificiale", in grado di gestire diverse applicazioni di Intelligenza Artificiale, di varia natura. Ciò implica la definizione di servizi di rete e strutture dati che possano essere sfruttati dalle applicazioni di AI, la capacità di assegnare in modo flessibile le risorse computazionali, e di individuare il posizionamento ottimale dei carichi di lavoro per l'AI all'interno dell'architettura di rete. L'architettura di rete 6G deve poter garantire elevata affidabilità agli applicativi basati su AI che dovrà ospitare, facilitare la condivisione della conoscenza e dei dati all'interno della rete, e al contempo offrire soluzioni per il risparmio energetico. Gli abilitatori tecnologici per raggiungere questo obiettivo sono posizionati in Fig.2 all'interno del livello di servizio di rete, con interfacce verso il livello applicativo, nonché verso le altre funzionalità basate su AI interne alla rete stessa;
- **l'AI/ML come abilitatore per la sostenibilità delle reti 6G**. L'AI/ML ha le potenzialità per migliorare l'efficienza energetica delle reti di comunicazione. La sostenibilità delle soluzioni basate su AI/ML è un tema a cui prestare particolare attenzione, in quanto questi approcci spesso portano a realizzare grandi reti neurali, estremamente complesse, che devono essere eseguite in tempo reale, e che possono determinare un importante consumo di risorse. Hexa-X ha evidenziato che in molti casi è possibile realizzare architetture AI più semplici sfruttando la conoscenza del problema che si cerca di risolvere. L'AI/ML può inoltre migliorare l'efficienza energetica risolvendo proble-

Figura 2: Collocazione delle aree tematiche per gli abilitatori tecnologici basati su ML/AI all'interno dell'architettura definita dal progetto Hexa-X

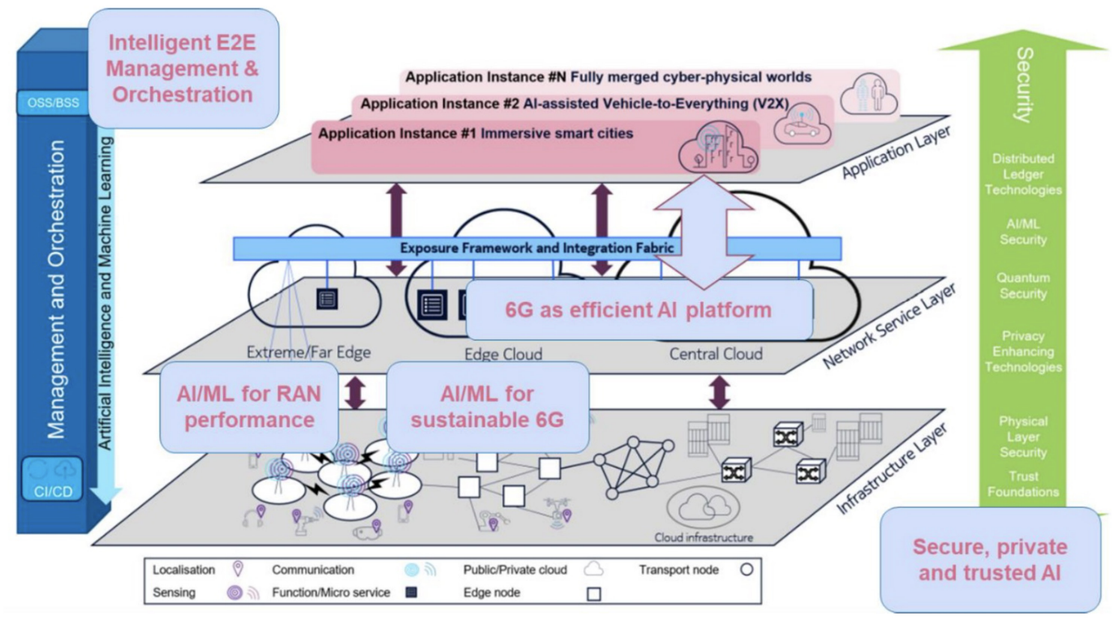
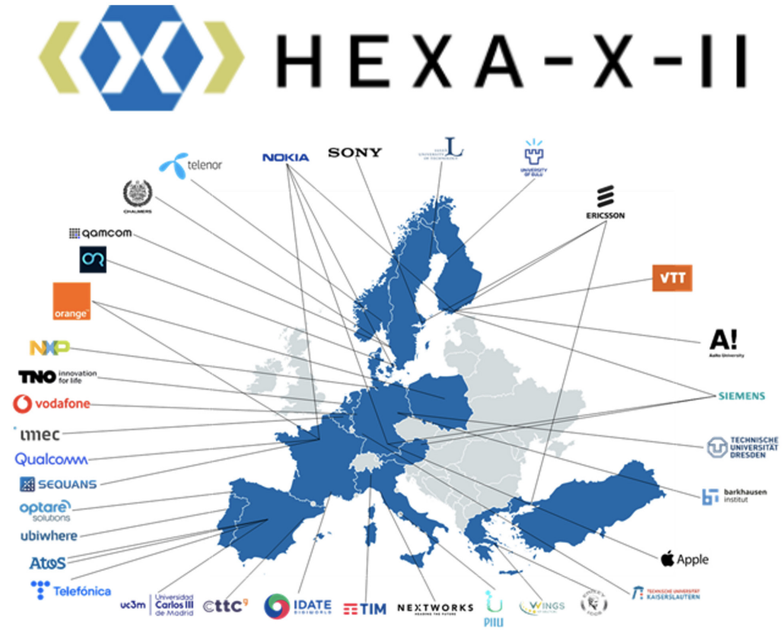


Figura 3: Il Progetto Hexa-X-II



mi di ottimizzazione che risultano intrattabili con i metodi convenzionali;
- **privacy, sicurezza e affidabilità nell'AI abilitato dal 6G**. L'uso di AI su enormi quantità di dati comporta la necessità di affrontare minacce alla privacy e alla sicurezza. Un obiettivo primario degli abilitatori AI è progettare e sviluppare sistemi che siano trasparenti, affidabili ed equi, garantendo al contempo la privacy e la sicurezza dei dati.

Il progetto Hexa-X è terminato a giugno 2023, ma ora è in corso un proseguimento nell'ambito di Hexa-X-II (Fig.3), il progetto della 6GIA JU SNS avviato a inizio 2023. Il progetto svilupperà la visione del 6G, i concetti di base, e le potenziali tecnologie chiave abilitanti, tra cui l'applicazione di AI/ML.
L'obiettivo di Hexa-X-II in questo ambito è quello di investigare l'applicazione dell'AI/ML per la progettazione e l'ottimizzazione dell'interfaccia radio 6G.

Le soluzioni di AI/ML vengono impiegate per apprendere automaticamente alcune specifiche funzionalità al trasmettitore, al ricevitore, o in modo congiunto operando ad entrambi i lati della catena radio. All'interno del progetto sono state proposte soluzioni per applicare AI/ML alla trasmissione MIMO, al beamforming, alla selezione degli utenti in ambito MU-MIMO, e in generale all'allocazione delle risorse radio.
Sono inoltre proposte soluzioni per apprendere in modo automatico la forma d'onda da utilizzare, per ottimizzare modulazione e codifica, per compensare le non-linearità al ricevitore, per migliorare la stima del canale e comprimere il feedback che viene generato dal ricevitore.

Si vuole dimostrare che integrando l'AI/ML nel design dell'interfaccia radio, il sistema 6G potrà beneficiare di maggiore flessibilità, maggiore efficienza, migliori prestazioni e migliore adattabilità alla molteplicità di condizioni ambientali e scenari in cui verrà dispiegato.

L'AI all'EDGE

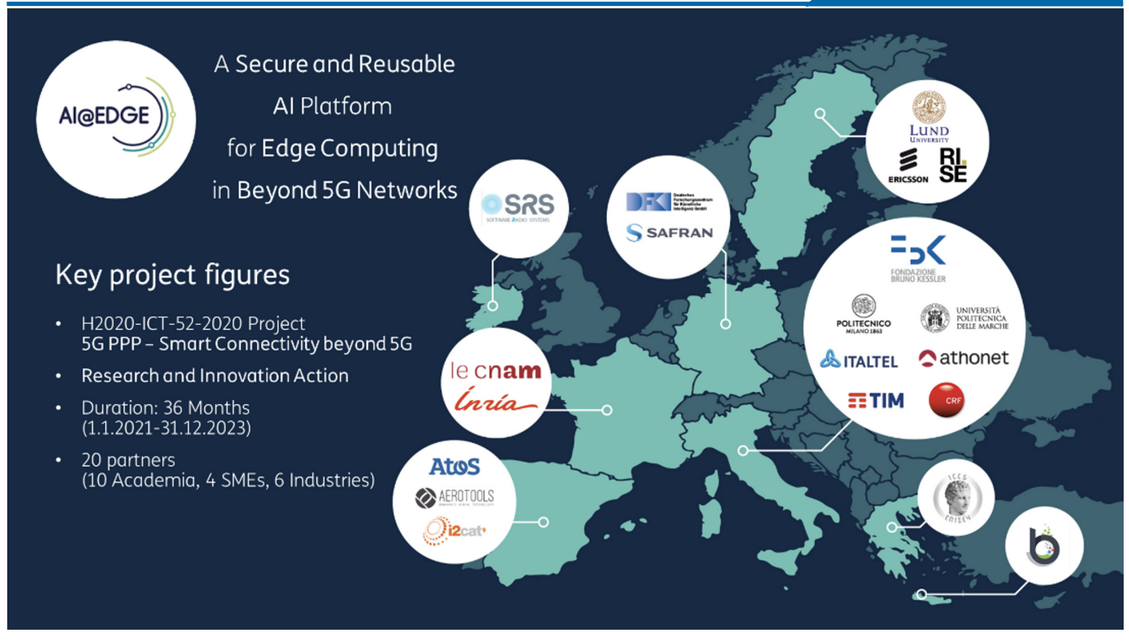
La possibilità di adottare metodi di AI/ML per la gestione delle reti del futuro in combinazione con l'Edge Computing risulta particolarmente promettente per soddisfare requisiti come reattività, dinamicità, gestione ottimizzata dei dati, sicurezza e tutela della privacy.
Edge Computing racchiude diversi tipi di dispositivi capaci di calcolo che non fanno parte delle infrastrutture centralizzate in Cloud. Collocando le risorse di calcolo e di trasmissione "ai bordi della rete" si può incrementare la velocità di risposta dei sistemi.
L'elaborazione dei dati avviene il più vicino possibile là dove vengono generati tali dati, permettendo la gestione, la fusione e l'analisi di dati in tempo reale o quasi.
A beneficiarne saranno in particolare servizi che necessitano maggiore privacy, maggiore affidabilità e minore latenza come, ad esempio, servizi nei settori automotive e smart mobility.

Il progetto EU AI@EDGE (Fig.4) - *"A secure and reusable Artificial Intelligence platform for Edge computing in beyond 5G Networks"* - è stato avviato nel gennaio 2021 con l'obiettivo di studiare, progettare e sperimentare la piattaforma che integra Intelligenza Artificiale ed Edge Computing, abilitando così nuovi servizi insieme all'automazione nelle reti.

Le reti future B5G/6G e la crescita esponenziale dei dispositivi connessi rappresentano la sfida per una gestione efficiente e affidabile delle risorse di trasmissione e di elaborazione.
A questo si aggiunge la complessità di gestione di servizi sempre più avanzati.
Il progetto AI@EDGE punta a fornire proposte e risposte ad alcune di queste esigenze, ponendo focus sui seguenti ambiti di ricerca:
- AI/ML per l'automazione della rete e dei servizi "a circuito chiuso";
- tutela della privacy, apprendimento automatico per ambienti multi-stakeholder;
- piattaforma convergente di elaborazione e comunicazione distribuita e decentralizzata;
- Provisioning e Orchestation di applicazioni AI/ML, definite come AIFs – AI Functions;
- Piattaforma Serverless con accelerazione hardware per le applicazioni AI/ML;
- cross-layer, multi-connectivity, disaggregated radio access.

Per portare l'Intelligenza Artificiale "ai bordi della rete" è stata definita la piattaforma AI@EDGE complessiva (Fig.5), descritta in dettaglio nel *"Deliverable D2.3 - Consolidated system architecture, interfaces specifications, and techno-economic analysis"*, dove TIM ha avuto la responsabilità ed il ruolo di Lead Editor.
L'architettura del sistema AI@EDGE, composta da Connect-Compute Platform (CCP) e Network and Service Automation Platform (NSAP), prevede le funzionalità a supporto di:
- "In-Platform AI" - Network and Service Automation intelligence (red head in Fig.5) - che consente un migliore utilizzo delle risorse infrastrutturali ed alte prestazioni della piattaforma convergente di elaborazione e comunicazione (CCP);

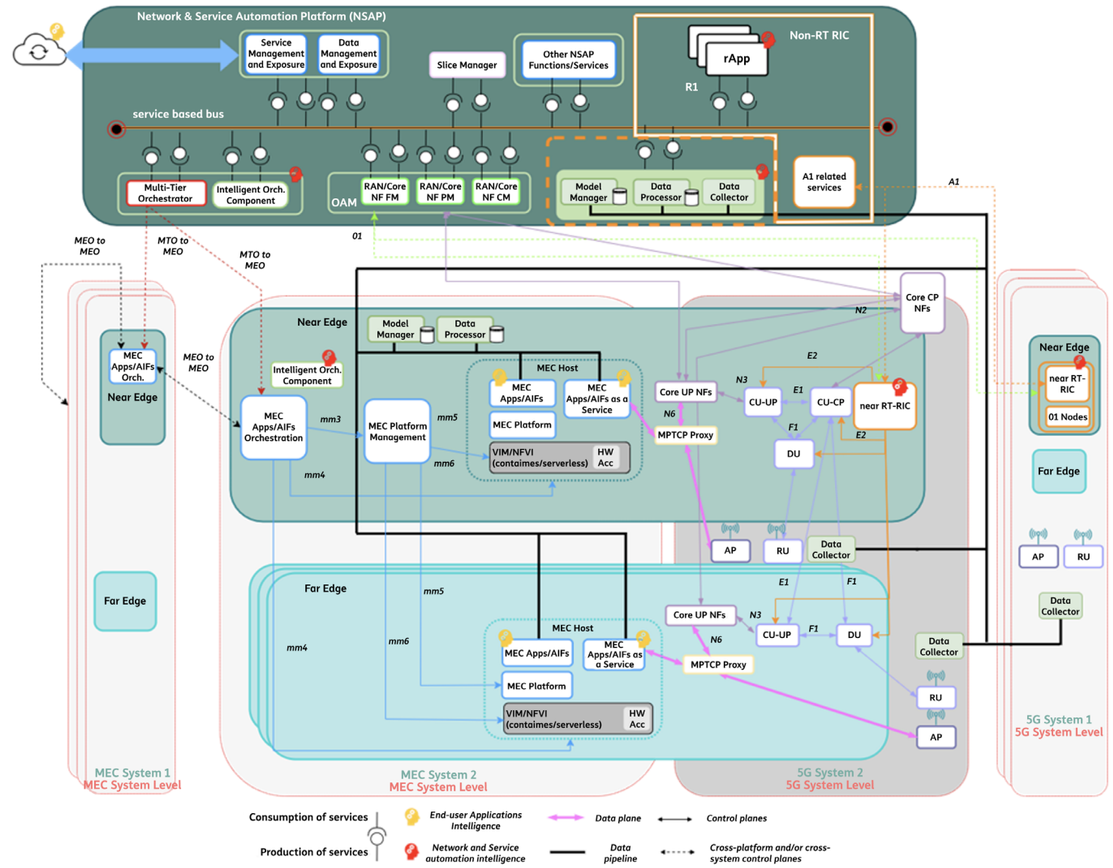
Figura 4: Il Progetto AI@EDGE



- “On-Platform AI” - End-user Application intelligence (yellow head in Fig.5) - con dispiegamento semplice e flessibile delle applicazioni di terze parti per una migliore qualità dell’esperienza dei servizi per l’utente finale in vari domini applicativi;
- Data Pipelining e Data Governance per garantire la privacy e la sicurezza dei dati in un ambiente multi-stakeholder;
- Orchestrazione e gestione End2End del sistema, comprensiva della gestione dei workflow e workload specifici di AI/ML, integrata con l’orchestrazione e la gestione della rete e delle risorse di elaborazione, consentendo ecosistemi dinamici, distribuiti e aperti.

La NSAP risulta basata sui servizi (Service Based Architecture - SBA), prevede la federazione dei domini (Core, RAN ed Edge) e l’integrazione delle componenti di terze parti. La Connect-Compute Platform (CCP) combina cloud-edge computing, virtualizzazione, accelerazione hardware (GPU, FPGA e CPU) e la RAN disaggregata in un’unica piattaforma che presenta i vantaggi dei paradigmi cloud-native consolidati.

Figura 5: AI@EDGE Consolidated System Architecture



La CCP supporta le funzionalità per l’Intelligenza Artificiale estendendo la distribuzione e l’orchestrazione delle applicazioni MEC agli aspetti specifici relativi all’Intelligenza Artificiale come: gestire e supportare workload AI-intensive; gestire le informazioni sui dati utilizzati per creare e aggiornare i modelli AI/ML; gestire le politiche del ciclo di vita AI-specific al fine di monitorare le prestazioni dei modelli AI/ML, la loro evoluzione e sostituzione.

La piattaforma AI@EDGE sarà validata utilizzando quattro casi d’uso con requisiti specifici che non possono essere soddisfatti dalle attuali reti, particolarmente in termini di supporto per le applicazioni dinamiche, sensibili alla latenza e abilitate all’Intelligenza Artificiale.

- UC1: Validazione virtuale della percezione cooperativa tra veicoli - ricreare lo scambio e analisi dei dati a livello di piattaforma necessari per la costruzione di una percezione cooperativa tra veicoli emulati e un veicolo guidato da esseri umani in contesto della rotonda stradale;
- UC2: Orchestrazione sicura e resiliente di grandi reti (I)IoT – dispiegare l’Intelligenza Artificiale per la sicurezza (Intrusion Detection) a livello di dispositivo IoT e a livello della rete; validare inoltre il dispiegamento sicuro dell’Intelligenza Artificiale (Adversarial Machine Learning);
- UC3: Monitoraggio, assistito da Edge AI, di infrastrutture critiche con droni in operazioni BVLOS – validare utilizzo delle funzionalità della piattaforma durante la scansione eseguita dai droni, combinando le capacità di calcolo integrate nei droni, nei dispositivi edge dedicati ed in cloud;
- UC4: “Smart content & data curation” per servizi di intrattenimento a bordo degli aerei - fornire contenuti selezionati ai passeggeri delle compagnie aeree su un’infrastruttura cloud edge a bordo degli aerei.

L’architettura AI@EDGE descritta nella Fig.5 non è dedicata specificatamente ai quattro use case da validare, ma anzi presenta elementi come Data Collector, Data Processor e (AI/ML) Model Manager riutilizzabili da molteplici casi d’uso, con l’obiettivo di superare l’approccio “AI-Silos” con pipeline di data/model management dedicate per ciascuna applicazione di Intelligenza Artificiale.
Ciò significa che i nuovi casi d’uso dovrebbero riutilizzare non solo i dati (elementari e/o pre-processati), ma anche i modelli AI/ML (addestrati o no).

Conclusioni

Le nuove reti dovranno al tempo stesso possedere un’architettura nella quale ospitare elementi di AI e nuovi algoritmi dovranno essere studiati e applicati, per rendere tangibili i benefici potenzialmente abilitati dall’AI.
Questi benefici, in ultimo, saranno materialmente realizzati attraverso l’abilitazione di nuovi servizi e applicazioni che le reti “intelligenti” potranno consentire.
In questo articolo si sono analizzate soluzioni architetturali e potenziali applicazioni che l’AI potrà consentire, in un riferimento temporale di qualche anno, facendo leva sulla virtualizzazione della rete e sulle Network Functions (NFs) in grado di implementare algoritmi di AI/ML.
La piena adozione di queste tecnologie dovrà transitare, come sempre, attraverso il processo di standardizzazione delle stesse, come accennato nel box dedicato.■

AI
L'associazione 6GIA e il ruolo di TIM

6G IA è un'associazione no profit che rappresenta l'industria europea delle telecomunicazioni in una partnership pubblico privata con la Commissione Europea denominata SNS JU. 6GIA ed SNS JU sono rispettivamente la continuazione di 5GIA e 5GPPP per lo sviluppo del 6G.

Il ruolo di 6GIA é quello di fornire guidance industriale alla ricerca finanziata sul 6G in Europa. La guidance viene impressa tramite un Workprogram emesso a scadenza biennale a cui partecipano membri volontari del Board di 6GIA. Il workprogram serve ad indirizzare le call finanziate la partecipazione è regolata da un meccanismo IKOP che favorisce i membri dell'associazione e che ha fatto lievitare la membership di 6GIA sino a quasi 300 soci. Altra novità rilevante che caratterizza SNS JU è il meccanismo IKAA, ovvero il meccanismo di accounting dei contributi della industria privata alla ricerca sul 6G al di fuori dei progetti finanziati europei – 6GIA deve essere accountable per un contributo certificato "in kind" pari ai finanziamenti della Commissione Europea ovvero 900 Mil€ sino al 2030.

La governance di 6GIA è garantita da un Board eletto ogni 2 anni che vede la partecipazione di TIM assieme ad altri incumbents europei (Orange, DT, Telenor) ed ai vendors di pima fascia (Ericsson, Nokia e Huawei). TIM ha assunto la Vice Chairmanship e mantiene la leadership sulle attività di Verticals Engagement, partecipando ad una CSA denominata SNS ICE con il ruolo di promotore del 6G presso i principali settori industriali europei. Sotto la guida di TIM, 6GIA ha realizzato numerose partnership strategiche tramite protocolli di intesa con associazioni quali 5GAA (Automotive), 5G ACIA (Smart Manufacturing), ESA (Spazio), 6GHI (Health), PSCE (Public Safety), ECSO (Cybersecurity), ERTICO (Transportation) a cui si è aggiunto di recente 5G MAG (Media) facente parte di EBU.
La governance di SNS JU è garantita da un Governing Board che vede una partecipazione paritetica tra Commissione Europea e industry privata rappresentata da 6GIA. Le operations saranno gestite da un Office con sede a Bruxelles che ha nominato Erzsébet Fitori nuova Executive Director, in carica dal 1 Settembre 2023 per 4 anni.

Bibliografia

1. SNS JU - https://smart-networks.europa.eu/
2. 6GIA - https://6g-ia.eu/

Acronimi

6G IA	6G Smart Networks and Services Industry Association
DT	Deutsche Telekom
IKAA	In Kind Additional Activities
IKOP	In Kind Operations
SNS JU	6G Smart Network & Services Joint Undertaking

Standard per AI nelle reti di telecomunicazioni

L'introduzione dell'Artificial Intelligence (AI) applicata alle reti di telecomunicazioni ha visto negli ultimi anni un incremento di interesse cui stanno seguendo sviluppi negli standard internazionali. In questo ambito, l'adozione di tecniche di Machine Learning (ML) permette il potenziamento delle funzionalità di rete e dei processi operativi di gestione della rete stessa e dei servizi

Per garantire l'interoperabilità e facilitare l'industralizzazione di soluzioni aperte, tutti gli standard di riferimento prevedono degli elementi abilitanti comuni:

- interfacce e API per l'acquisizione, il processamento, l'archiviazione, e l'esposizione dei dati generati dai nodi di rete, dai terminali e dalle piattaforme di servizio, in modalità automatica secondo il paradigma cosiddetto "DataOps". Ciò include la federazione di dati da altri domini (per esempio: meteo, palinsesti video, abitudini degli utenti) per rinforzare la capacità inferenziale degli algoritmi AI/ML;
- funzionalità per l'orchestrazione dell'intero ciclo di vita degli algoritmi che comprendono la raccolta dei dati, l'analisi preliminare dei dati, l'addestramento e il testing dell'algoritmo, il trasferimento nell'ambiente operativo target, l'attivazione, il monitoraggio delle prestazioni durante la sua fase inferenziale e il versionamento dei modelli di ML secondo il paradigma MLOps.

Nel seguito si descrivono le attività su AI/ML di alcuni enti di standardizzazione cui TIM contribuisce.

O-RAN Alliance

Uno dei principali obiettivi di O-RAN Alliance [RIF-A] è l'evoluzione dell'architettura dell'accesso radio 5G per supportare nativamente AI/ML, tramite l'introduzione di due diversi RAN Intelligent Controller, piattaforme programmabili che introducono funzionalità di ML a supporto rispettivamente dei processi di gestione non real-time a livello Service Management and Orchestration (SMO) e delle procedure di controllo near real-time della rete [RIF-B].
Le API esposte da tali piattaforme, la cui definizione sarà finalizzata nei prossimi mesi, faciliteranno lo sviluppo di applicazioni ML-based di configurazione, monitoraggio, efficientamento e ottimizzazione automatica della rete, anche grazie all'impulso delle community, quali ONAP e O-RAN SC, che sviluppano componenti open source secondo i requisiti e le specifiche O-RAN.

Broadband Forum

Il Broadband Forum, fondato nel 1994, è l'ente di riferimento per gli standard di accesso fisso inclusi i domini Edge e Customer Premises. Dal 2018, sono state sviluppate le specifiche per Automated Intelligent Mangement (AIM).
In particolare il TR-436 [RIF-BBF1] definisce il framework e specifica i requisiti architetturali e funzionali delle soluzioni AIM. IL framework recepisce i lavori sviluppati in ITU-T [RIF-C], ETSI GANA ed ETSI ENI e li armonizza con le architetture di rete moderne fondate su tecnologie SDN, NFV e Cloud. Inoltre è definito in modo agnostico rispetto ai domini di rete coinvolti e/o federati e soprattutto rispetto alle specifiche applicazioni di AI/ML (per es.: manutenzione proattiva, efficienza energetica, ottimizzazione dell'esperienza del cliente).
L'AIM Suite è completata dal WT-486 (che sarà pubblicato a fine 2023) [RIF-BBF2] che specifica le interfacce fra i sottosistemi logici, quelle di orchestrazione delle pipeline di ML e di comunicazione interna ed esterna delle pipeline.

3GPP

Nell'ambito del 3GPP numerose sono le attività di definizione di funzionalità a supporto dell'introduzione diffusa di algoritmi basati su tecniche di AI/ML in grado di svolgere un ruolo fondamentale sia in ambito previsionale (delle prestazioni di rete e dei servizi, di possibili fault) sia in ambito di ottimizzazione delle prestazioni della rete stessa.
Il passaggio dal paradigma SON (Self-organizing Network) del 4G alle metodologie basate su AI/ML nel 5G ha spostato l'attenzione sui dati prodotti dalla rete e utilizzati per la gestione del ciclo di vita di un algoritmo AI/ML.
Inoltre, la softwarizzazione della rete ha permesso di poter distribuire tali algoritmi in tutte le Network Function con conseguente coinvolgimento di quasi tutti i gruppi tecnici: SA5 per gli aspetti di gestione, SA2 per le funzionalità di Network Data and Analytics (NWDAF) ed esposizione dei dati [Rif. 3GPP4], SA3 per gli aspetti di sicurezza [Rif 3GPP5] e RAN3 per il coinvolgimento dei nodi NG-RAN [Rif. 3GPP6]). Espandendo il concetto di distribuzione, sono state definite procedure per abilitare il Federated Learning in zone in cui i dati delle Network Function non sono rese disponibili a livello centralizzato per ragioni di privacy [Rif. 3GPP7].

Alcuni degli Study Item in corso di definizione per la Release 19 dello standard riguardano la gestione dei dati e del ciclo di vita degli algoritmi AI/ML e prevedono la collaborazione tra il 3GPP e altri gruppi di standard quali ETSI ZSM, ETSI SAI e TMForum.

simone.bizzarri@telecomitalia.it
andrea.buldorini@telecomitalia.it
mauro.tilocca@telecomitalia.it
antonio.varvara@telecomitalia.it

Riferimenti

1. [RIF-A] Notiziario Tecnico TIM 2-2019 - "Enabling Software Intelligence all over the Wireless Access: The O-RAN Initiative"
2. [RIF-B] O-RAN AI/ML workflow description and requirements 1.03 (10/2021)
3. [RIF-C] ITU-T Y.3172 - Architectural framework for machine learning in future networks including IMT-2020 (09/2019)
4. [RIF BBF1] BBF TR-436 - Access & Home Network O&M Automation/Intelligence (02/2021)
5. [RIF-BBF2] BBF WT-486 - Interfaces for AIM (pubblicazione prevista alla fine del 2023)
6. [Rif. 3GPP1] TR 28.908 "Study on Artificial Intelligence / Machine Learning (AI/ML) management"
7. [Rif. 3GPP2] TS 28.104 "Management Data Analytics (MDA)"
8. [Rif. 3GPP3] TS 28.105 "Artificial Intelligence / Machine Learning (AI/ML) management"
9. [Rif. 3GPP4] TR 23.700-80 "Study on 5G system support for AI/ML-based services"
10. [Rif. 3GPP5] TR 33.898 "Study on security and privacy of Artificial Intelligence/Machine Learning (AI/ML)-based services and applications in 5G"
11. [Rif. 3GPP6] TR 37.817 "Study on enhancement for Data Collection for NR and EN-DC"
12. [Rif. 3GPP7] TR 23.700-81 "Study of Enablers for Network Automation for 5G System (5GS)"

Bibliografia

1. https://nowpublishers.com/article/BookDetails/9781638282389
2. https://hexa-x.eu
3. HEXA-X D4.3 - AI-driven communication & computation co-design solutions
4. HEXA-X D1.4 - Hexa-X architecture for B5G/6G networks - final release
5. https://hexa-x-ii.eu/
6. HEXA-X-II D2.1 – Draft foundation for 6G system design
7. https://aiatedge.eu/
8. AI@EDGE_D2.3_Consolidated-system-architecture-interfaces-specifications-and-techno-economic-analysis_v1.0.pdf (https://aiatedge.eu)
9. ETSI - Multi-access Edge Computing - Standards for MEC

Acronimi

5G ACIA	5G Alliance for Connected Industry and Automation
5GAA	5G Automotive Association
6G IA	6G Smart Networks and Services Industry Association
6GHI	6G Health Institute
AI	Artificial Intelligence
AIF	AI Function
B5G	Beyond 5G
BVLOS	Beyond Visual Line of Sight
CCP	Connect-Compute Platform
CPU	Central Processing Unit
CSA	Collaboration Support Action
DG	Directorate General
DT	Deutsche Telekom
EBU	European Broadcasting Union
ECSO	European Cybersecurity Organization
ERTICO	European Road Transport Telematics Implementation COordination
ESA	European Space Agency
ETSI	European Telecommunications Standards Institute
EU	European Union
FPGA	Fixed Programmable Gate Array
GPU	Graphics Processing Unit
IKAA	In Kind Additional Activities
IKOP	In Kind Operations
IoT	Internet of Things
JU	Joint Undertaking
KPI	Key Perfomance Indicators
KVI	Key Value Indicators
MEC	Mobile Edge Computing
MIMO	Multiple Input Multiple Output
ML	Machine Learning
MU-MIMO	Multi-User MIMO
NF	Network Function
NSAP	Network and Service Automation Platform
PSCE	Public Safety Conference Europe
RAN	Radio Access Network
RIS	Reconfigurable Intelligent Surface
SBA	Service Based Architecture
SNS	Smart Network and Services
SNS ICE	SNS International Cooperation Ecosystem
SNS JU	6G Smart Network & Services Joint Undertaking

Autori

Jovanka Adzic *jovanka.adzic@telecomitalia.it*

Laureata in Informatica nel 1996, l'anno dopo consegue il Master in Informatics and Telecomunications del COREP – Politecnico di Torino. Successivamente entra in Azienda e si occupa di soluzioni innovative e sviluppo sistemi basati sulle tecnologie di Data Warehouse, Business Intelligence, Data Analytics e Data Mining, gestendo anche lo sviluppo del sistema di Advanced Analytics a supporto dell'antifrode su rete radiomobile. Oggi si occupa di progetti di ricerca su Advanced Analystics e AI, incluse le collaborazioni universitarie e la partecipazione ai progetti Europei nell'ambito del programma Horizon Europe, come AI@EDGE. ■

Mauro Boldi *mauro.boldi@telecomitalia.it*

Laureato con Lode in Ingegneria Elettronica al Politecnico di Torino nel 1997, è entrato in Azienda occupandosi di soluzioni Radio over Fiber e più in generale dell'accesso mobile. Dal 2010 segue i temi di efficienza energetica come vice-chairman del gruppo ETSI EE. Gestisce per TIM le partecipazioni ai progetti Europei nell'ambito del programma Horizon Europe. ■

Roberto Fantini *roberto.fantini@telecomitalia.it*

Ingegnere delle telecomunicazioni, è entrato in Azienda nel 2002, occupandosi dell'evoluzione dello standard 3GPP dal 3G fino al 5G, sia sviluppando piattaforme simulative per valutare le prestazioni di tali tecnologie, sia partecipando a trial in laboratorio e in campo. Ha partecipato a diversi progetti europei, tra cui METIS e METIS-II, in cui sono state poste le basi per il futuro 5G, ed è attualmente parte di HEXA-X-II, il progetto "flagship" della comunità europea per la definizione del 6G, per cui segue le attività relative all'applicazione dell'AI alla trasmissione radio. ■

Generative AI: la sfida di TIM per il futuro dell'IT

Mario Bonnet, Dario Mana, Clara Oliva, Pier Carlo Paltro

In questo articolo passiamo in rassegna le principali modalità mediante cui i Large Language Model e le altre tecnologie GenAI potranno avere un impatto, a 360 gradi, dal Marketing al Caring, dalla Rete alle funzioni di Staff.
A seguire verrà descritta la metodologia che si sta seguendo in TIM per indirizzare in modo organico l'adozione di soluzioni basate sull'Intelligenza Artificiale generativa. Seguirà uno step di valutazione di impatto dei case candidati, al fine di determinare una short list di casi che saranno oggetto di PoC e di roll-out. Lo scopo è quello di muoversi in modo armonico, al fine di cogliere le opportunità che la tecnologia offre, senza lasciarsi trascinare da proposte estemporanee e vertical.

Il fenomeno della GenAI

L'Intelligenza Artificiale generativa offre, a partire dal turning point dell'introduzione di ChatGPT, a fine 2022, una varietà apparentemente infinita di opportunità per le aziende per rendere più efficienti e più efficaci i propri processi.
I grandi gruppi consulenziali fanno a gara per alimentare l'hype che circonda questa nascente tecnologia: Forrester sostiene che il 10% delle Fortune 500 genererà contenuti usando la GenAI entro la fine di quest'anno; Accenture che, entro il 2028, ci sarà un incremento del 30% nella produttività degli impiegati grazie all'automazione e alla sintesi di insight.

A queste mirabolanti previsioni si accompagnano investimenti concreti che aziende grandi e piccole stanno facendo nell'ambito della GenAI: è di gennaio la notizia che Microsoft ha investito 10 miliardi di dollari in OpenAI; a luglio erano più di 450 le startup al lavoro sui temi della GenAI; recentemente Amazon ha annunciato investimenti per 4 miliardi in Anthropic, l'azienda che ha sviluppato il modello Claude; la stessa SK Telecom ha investito 100 milioni sempre in Anthropic. Dal punto di vista dell'offerta dei modelli cosiddetti foundation, gli Hyperscaler si sono tutti messi in gioco: Microsoft, rivendendo le soluzioni OpenAI in contesto enterprise; Google, offrendo soluzioni chiavi in mano, come Vertex AI Search & Conversation, basato sul suo modello LLM Palm e promettendo l'arrivo a breve di un nuovo "super" modello, Gemini; Amazon segue una strada diversa e cerca di far leva sulle risorse della community e con Bedrock raggruppa soluzioni provenienti da Hugging Face, Anthropic, Stability, etc. Anche Meta, attualmente leader nelle soluzioni open source con il suo LLaMA, ha annunciato lo sviluppo di un nuovo modello in grado di competere con quelli dei rivali. Apple è attesa anche lei nella partita e pare stia investendo molto nel settore.
Una delle grandi promesse della GenAI è quella di fungere da assistente, co-pilot, per i c.d. Knowledge Worker/White Collar, aumentandone la produttività e liberandone quota parte del tempo, che potrà essere dedicata alle attività di più alto livello strategico.
La funzione della GenAI in qualità di assistente e non di sostituta del lavoratore: resterà comunque in capo al lavoratore stesso la responsabilità di adottare o meno la soluzione suggerita dalla GenAI, liberando quest'ultima dall'onere di fornire risultati corretti nel 100% dei casi, cosa che al momento non è in grado di garantire.

La frontiera della sperimentazione è data da veri e propri agenti [MS] che sono in grado di coniugare la conoscenza proveniente da documentazione aziendale con informazioni prelevate in real-time da Internet.
Gli ultimi prototipi sono in grado, in totale autonomia, di generare codice in grado di reperire informazioni dalla rete, eseguirlo, eventualmente debuggandolo in caso di errore e, infine, integrare le informazioni reperite con la conoscenza estratta dalla documentazione per rispondere alle domande degli utenti.

GenAI e Telco

Le possibili applicazioni della GenAI coprono uno spettro molto ampio e la quasi totalità

degli use case proposti dagli analisti si declinano contemporaneamente su più industry.
A questi casi cross si aggiungono una serie di use case specifici del settore telco.

Opportunità strategiche
Case a maggior diffusione

Conversational AI su KB – a.k.a. Search & Summarize
Si tratta dello use case di gran lunga più diffuso e prevede la creazione di un chatbot (o Conversational AI) in grado di dialogare con un esperto di dominio sul contenuto di un set di documenti di riferimento (manualistica, contratti, documentazione interna di Privacy & Compliance, etc.) [AWS].
L'esperto, a valle dell'"addestramento" iniziale del bot, può porgli delle domande e ottenere risposte. Le domande possono anche richiedere step di ragionamento da parte del bot, come l'aggregazione di informazioni o il confronto di punti di vista. Tecnicamente le soluzioni stanno ancora evolvendo, cercando di creare schemi di prompting capaci di "far ragionare" il LLM facendogli suddividere task ad oggi per lui troppo complessi in sotto-task più semplici. L'esecuzione di questi sotto-task permette la raccolta delle informazioni necessarie per rispondere all'utente.
Su questo fronte, è bene tener d'occhio anche le offerte evolutive delle applicazioni office: Microsoft Office 365 con Copilot, Google Duet AI, che promettono di abilitare capacità di interfacciamento conversazionale con i propri documenti.

Co-pilot per Customer Care, Vendite e attività di Operation dei tecnici; ChatBot
Questo case, che è una specializzazione del precedente, prevede la disponibilità per gli Operatori e per i tecnici di rete di interfacce conversazionali che li supportino da un lato nella risposta alle esigenze dei clienti [O2] [Acc] e dall'altro nel trovare soluzioni e guide durante fasi di delivery/manutenzione dei servizi [AWS].
In questi contesti la GenAI può essere utilizzata anche per generare in automatico riassunti di telefonate o di conversazioni piuttosto che e-mail di recap/follow-up da inviare come reminder ai clienti che hanno contattato l'assistenza o ai prospect che si sono informati su caratteristiche di prodotti o servizi.
Un case molto popolare nel mondo Telco [AWS] [Acc] è quello dei chatbot, che prevede di esporre direttamente al cliente finale un'interfaccia conversazionale, come già si fa da anni. La differenza sta nel fatto che oggi il motore che genera le risposte del bot può essere un LLM e quindi capace di migliorate capacità di comprensione ed efficacia nelle risposte, al prezzo di un rischio di generazione di risposte non sempre corrette.

Code Generation for IT
In questo caso modelli LLM specializzati nella generazione del codice, come Codex di OpenAI, usato da GitHub Copilot, assistono i programmatori in vari task legati allo sviluppo e alla manutenzione del codice: code generation, debugging, refactoring, spiegazione e documentazione del codice, generazione di test automatici [AWS].
Le previsioni più ottimistiche sbandierano un vertiginoso 100% di incremento nella produttività ("twice as fast") di chi scrive codice [McK]. Va osservato che, anche se venissero raggiunte queste vette, tuttavia le attività IT nel perimetro delle grandi aziende solo in parte si declinano in task di scrittura/evoluzione/manutenzione del codice e comprendono, invece, e forse per la maggior parte, attività sicuramente più difficili da automatizzare legate alla gestione delle risorse, al coordinamento, alla stesura ed evoluzione delle specifiche e dei sistemi legacy e integrazioni che ne derivano.

Enhanced Data Analysis
Questo caso prevede che la GenAI venga usata per assistere le attività dei Data Analyst. Ad oggi, tipicamente, i Data Analyst hanno a disposizione degli ambienti analitici, come SAS, dove impostano delle query in linguaggi specializzati al fine di ottenere risposte a domande di business e insight di valore.
La promessa della GenAI, in questo ambito, è quella di poter chiedere le risposte che si desiderano in linguaggio naturale, ottenendo risposte in forma di narrazione, di grafici o direttamente di report completi.
Esistono plugin per ChatGPT specializzati nella generazione di SQL a partire da domande poste in linguaggio naturale, così come vari strumenti di BI ormai da tempo offrono capability di interrogazione in linguaggio naturale.

Generazione di contenuti creativi e iper-personalizzati
In questo caso i modelli generativi, non solo del linguaggio, ma soprattutto di immagini e video, quali DALL-E, Stable Diffusion e MidJourney, vengono utilizzati per creare immagini e contenuti testuali (copy) adatti alle esigenze di marketing o di caring [Acc]. I contenuti possono anche essere adattati alle caratteristiche del singolo cliente.
Va osservato che questo tipo di attività porta con sé il rischio di incorrere in problemi di copyright infringement. Sono ormai all'ordine del giorno le cause intentate da famosi artisti a OpenAI per violazione dei diritti di copyright.

AI for AI
Si tratta di use case dove le capability di comprensione e processamento di Large Language Model vengono utilizzate al servizio di altri algoritmi di Intelligenza Artificiale, quali clustering, classificazione, recommendation, Natural Language Processing, Sentiment Analysis, Anonymization, pulizia dei dati.

Learning and re-skilling
Questo use case mira a fornire ai dipendenti strumenti a supporto dell'acquisizione

Tabella 1: Mappatura degli UC più diffusi su Benefici/Ostacoli

	BENEFICI					OSTACOLI				
USE CASE	Ottimizzaz. effort/costi	Revenue Generation	Customer Satisfaction & QoS	Risposte standard	Acquisizione Know-How	Allucinazioni, logica fallace, Prompt Injection	Compliance, Security & Privacy	Costi di Consumption, Markup e Maintenance	Rischi di Copyright Infringement	Bias & Ethics
Search & Summarize	X		X	X	X	*		*		
Co-pilot & ChatBot	X	X	X	X	X	*	*	*		*
Code Generation	X				X					
Enhanced Data Analysis	X	X	X			*				
Gen. contenuti creativi e personalizzati	X	X				*	*	*	*	*
AI for AI	X	X	X					*		
Learning	X			X	X	*		*		
Generazione dati sintetici	X		X				*	*		

di nuove competenze o aggiornamento di knowledge già appresa. Mediante l'interfaccia conversazionale un argomento da apprendere può essere affrontato in maniera personalizzata: l'utente può fare domande dal proprio punto di vista e anche dichiarando esplicitamente quali sono le proprie competenze di partenza; ad esempio: "Tieni conto che sono uno sviluppatore software esperto nel linguaggio Java e sto imparando Python; come faccio, in Python, a realizzare un programma parallelo? Quali librerie conviene usare? Quali sono gli idiomi tipici?". Oppure l'IA può simulare il comportamento di un cliente che chiede assistenza su un certo argomento per abilitare nuove forme di training per gli operatori [Acc].

Generazione di dati sintetici

Questo caso d'uso prevede la generazione di dati finti, ma aventi le stesse caratteristiche statistiche di un certo dataset di partenza. In questo modo, i dati sintetici generati sono privi della presenza di dati personali e/o sensibili e possono essere utilizzati per addestrare modelli di Machine Learning con vincoli di privacy e compliance meno stringenti. Può essere il caso di modelli addestrati a partire dalle note degli Operatori piuttosto che dai dati di traffico dei clienti sulla rete.

Benefici

Ottimizzazione effort e costi

Il principale beneficio che viene sbandierato da tutti i sostenitori della rivoluzione della GenAI consiste nell'aumento della produttività dei lavoratori, nello specifico dei cosiddetti Knowledge Worker.

Come conseguenza gli operatori e tecnici potranno dedicare il proprio tempo alla risoluzione delle problematiche più complesse e che non siano catturate nell'ambito delle procedure codificate, consentendo una miglior Customer Satisfaction dal punto di vista del cliente e una miglior Quality of Service della Rete.
Ci si attende che le versioni future dei pacchetti di Office Automation (Teams, Office, Google Suite, etc.) integreranno strumenti di GenAI per abilitare scenari quali il riassunto automatico dei meeting, la ricerca conversazionale sui documenti presenti nella propria area personale oppure sull'intera Intranet, velocizzando procedimenti e ricerche che, ad oggi, risultano essere molto time-consuming.

Revenue Generation

La capacità di adattarsi alle caratteristiche dei clienti anche nel linguaggio e nel modo di porsi possono facilitare operazioni di upselling, cross-selling e Next Best Offer. La dimensione dell'incremento dei ricavi è impattata positivamente anche dall'aumentata capacità di produrre analisi e/o modelli analitici performanti.

Incremento della Customer Satisfaction e della Quality of Service

Clienti serviti in modo più efficace saranno più soddisfatti. Inoltre, gli strumenti abilitati dalla GenAI a supporto dell'operatività dei tecnici sulla rete ne determinerà un aumento della QoS di Rete. Anche la CS e la QoS, come la Revenue Generation, beneficiano della accelerata disponibilità di analisi e modelli ottenuta grazie all'applicazione della GenAI.

Standardizzazione delle risposte verso i Customer

Un effetto collaterale positivo dell'adozione di questi strumenti, per fornire risposte ai clienti, anche se non esclusivo della GenAI, è la standardizzazione delle risposte date ai clienti. Ad oggi, laddove i clienti si interfacciano con operatori human per la risoluzione delle proprie problematiche relative a prodotti e servizi dell'azienda, possono talvolta incappare in risposte che variano a seconda dell'operatore, per questioni di varietà di conoscenze e di skill.
La generazione di risposte verso il cliente da parte di strumenti di GenAI produrrà risposte maggiormente standardizzate e univoche dal punto di vista concettuale, anche se diverse di volta in volta.

Acquisizione di Know-How

Utilizzando strumenti della AI generativa "addestrati" su documentazione, manualistica e anche, dove applicabile, su passate interazioni operatore/cliente o lavorazioni di Operations, i lavoratori potranno acquisire più velocemente il know-how necessario ad affrontare le varie situazioni lavorative in modalità personalizzata ed adeguata al proprio livello di conoscenza.

Ostacoli all'adozione

Allucinazioni, logica fallace e rischi di Prompt Injection

Sono celeberrimi i casi di allucinazioni di ChatGPT: il modello, infatti, è stato addestrato per rispondere nel modo più probabile alle domande degli utenti, con, ovviamente, un pizzico di varietà. Il modello non è stato, invece, addestrato per essere corretto o fattuale nelle risposte che fornisce. Da qui scaturiscono i casi in cui le riposte generate, pur essendo molto verosimili e convincenti, tuttavia, a una più attenta analisi, si rivelano false.
Ne discende la necessità di adottare le soluzioni di GenAI in ambiti vincolati e monitorati, al fine di misurare e prevenire/mitigare i casi di allucinazioni. La fase di monitoring non può che essere svolta da persone esperte del dominio.

Va osservato che, nel caso in cui si sviluppino applicazioni customer-facing, senza mediazione tra quanto prodotto dalla GenAI e il cliente finale, esiste un rischio di generazione di allucinazioni e, quindi, un rischio reputazionale per l'azienda.

Ad oggi, la stessa OpenAI suggerisce di non sviluppare applicazioni direttamente affacciate verso il cliente, ma di prevedere sempre l'intermediazione di un esperto [OAI].

I modelli LLM sono molto bravi nel fare delle semplici deduzioni a partire da pochi semplici fatti. Questo è un loro punto di forza grandissimo, che abilita la capacità di rispondere a domande non banali. Tuttavia, quando la difficoltà delle domande aumenta e inizia a richiedere deduzioni complesse, aggregazioni di dati o computazioni, allora si manifestano abbastanza spesso episodi in cui la capacità di reasoning dell'LLM risulta fallace. Esistono delle tecniche per mitigare questo effetto (es.: Chain-of-Thought) e sono possibili integrazioni con strumenti esterni per sopperire, ad esempio, alle scarse capacità di computazione degli LLM.
Anche in questo caso è richiesto un rigoroso e costante controllo human, ancorché necessariamente a campione, per comprendere i limiti della tecnologia e tarare il giusto equilibrio tra le sue capability e quanto si offre/vende ai clienti o agli utilizzatori interni.
Nel caso di applicazioni customer-facing esiste la concreta possibilità che utenti malintenzionati cerchino di piegare il comportamento dei modelli LLM e costringerli a bypassare i vincoli di buona condotta imposti dai fornitori dei foundation model (cosiddetto hijacking).
In questi casi, esiste un rischio reputazionale dell'azienda che potrebbe vedere i propri chatbot generare contenuti scorretti o drammaticamente inadeguati.

Compliance , Security & Privacy
Molti degli use case di maggior valore richiedono che comunicazioni dei clienti con l'azienda e/o testo prodotto dai dipendenti e trascrizioni di riunioni siano elaborati da parte dei modelli LLM. La fattibilità normativa di tali elaborazioni va concordata con le funzioni di Security, Privacy e Compliance, anche tenendo conto del fatto che si tratta di servizi forniti esclusivamente in cloud dai vari Microsoft, Google, etc.
Attenzione, inoltre, all'imminente AI Act [EU], che regolamenterà l'uso dell'AI in ambito europeo, con finalità di protezione della privacy dei cittadini e di utilizzo etico della tecnologia.

Costi di Consumption, Markup e Maintenance
Ovviamente i servizi dei modelli di GenAI hanno un costo, tipicamente basato su misure di consumo da corrispondere ai fornitori di foundation model. Oltre a tali costi, nel caso in cui ci si affidasse a soluzioni costruite on top degli LLM di base, ci sarebbe da corrispondere il markup richiesto dal fornitore di servizi di turno. [Squ]
Inoltre, occorre provvedere all'organizzazione e alla spesatura di monitoraggi continui, a campione, e di maintenance per far fronte a cambiamenti che sicuramente avverranno nel tempo a livello di comportamento dei modelli foundation: i vendor si stanno attrezzando per offrire Long Term Support per almeno un sott'insieme dei LLM, ma gli orizzonti temporali visibili ad oggi sono di circa un anno di disponibilità garantita [Azu]; dopodiché il modello di base cambierà e tutte le applicazioni costruite on top ne potrebbero risentire.

Rischi di copyright infringement
Come già accennato, esiste il rischio che il contenuto generato dai modelli GenAI possa violare i diritti d'autore. In molti casi, per ora in ambito USA, i fornitori di foundation model sono stati chiamati in causa in tribunale per difendere il proprio operato [Wir].

Bias & Ethics
Altro rischio di tipo reputazionale per l'azienda, pur non essendo esclusivo dei modelli di GenAI, consiste nella possibilità che emergano nei contenuti prodotti i cosiddetti bias dei modelli foundation, legati a differenze, presenti nei documenti usati per il training iniziale dei modelli stessi, collegate al genere, al perimetro culturale di appartenenza, alle opinioni politiche, etc.
Anche questo rischio può essere monitorato, campionando le conversazioni con le intelligenze artificiali ed eventualmente adoperandosi in modo proattivo per far emergere le problematiche per poi indirizzarle.
Va riconosciuto che gli stessi fornitori di modelli foundation si adoperano da tempo per ingabbiare il comportamento dei modelli stessi in regole/meccanismi volti ad evitare comportamenti non fair.

Il challenge della GENERATIVE AI in ambito IT enterprise

Per analizzare le sfide che si pongono nell'adozione della AI generativa all'interno del dominio e delle applicazioni IT e nell'ambito dei dati proprietari e privati dell'azienda, vale la pena ripercorrere lo sviluppo di alcuni elementi tecnologici peculiari, per poi inquadrare la metodologia di selezione e di sviluppo delle soluzioni architetturali abilitanti e dei casi d'uso più interessanti per TIM.
E' necessario premettere che l'hype cycle delle nuove tecniche di Artificial Intelligence sta ponendo una pressione operativa su tutte le organizzazioni aziendali ed i dipartimenti IT, a livello mondiale, per i loro impiego e valorizzazione. Per questo motivo è importante avviare programmi di "early adoption" che sfruttino e sperimentino lo stato dell'arte oggi disponibile e consentano di misurare le tecnologie sulle aspettative di business, pur essendo consapevoli che i risultati ottenuti potrebbero essere superati in breve tempo dalle evoluzioni della ricerca e dei prodotti di mercato.

I driver tecnologici della Gen AI

La nascita della "Generative AI" è indissolubilmente legata all'ideazione dell'architettura "Transformer", una rivoluzionaria tecnologia concepita nel 2018 da Google e tutt'ora in costante evoluzione presso tutti i più importanti attori del mondo industriale e accademico.
L'architettura "Transformer" è realizzata da un encoder e un decoder costruiti collegando fra loro molteplici configurazioni di reti neurali profonde. Tra queste è presente **un meccanismo innovativo di apprendimento chiamato "Self-Attention"** capace di riconoscere pattern di correlazione molto sofisticati all'interno dei dataset di training.
Questo meccanismo è il principale responsabile dell'espressività creativa che si riscontra nei contenuti prodotti dalla IA, in quanto è capace di ricostruire e apprendere in modo molto dettagliato il contesto del dato in input, misurando il peso semantico delle sue componenti. In questo senso le parole di un testo, oppure gli elementi grafici elementari all'interno di un'immagine, vengono correlate considerando ogni possibile angolazione. Da questa sofisticata ricostruzione contestuale nasce la conoscenza profonda del modello e la sua capacità di produrre per inferenza statistica un nuovo contenuto più ricco ed espressivo dell'esempio iniziale, quest'ultimo indicato tecnicamente con il termine "prompt".
Ma la meraviglia dei "Transformer" non si limita a queste proprietà predittive. Infatti, la soluzione **si pone attualmente come ottimale anche in termini di consumo energetico, in quanto la sua architettura è intrinsecamente scalabile sulle moderne infrastrutture hardware oggi erogate dai principali cloud provider (GPU e TPU).** Questa caratteristica è fondamentale per il deployment delle soluzioni in campo, poiché velocizza i tempi di training e di serving, riducendo le risorse impiegate di diversi ordini di grandezza rispetto alle precedenti tecnologie RNN e CNN ormai ritenute obsolete ("Recursive e Convolutional Neural Network").
In definitiva, grazie alla conoscenza contestuale profonda e all'intrinseca scalabilità, l'architettura "Transformer" è alla base dei **"Large Language Model" (LLM), cioè dei modelli neurali che vengono pre-addestrati con enormi corpus di informazione e resi disponibili sul mercato enterprise dai più importanti vendor, tra cui spiccano Open AI, Google e Meta con i rispettivi ChatGPT, Palm2 e LLAMA2.**
Il processo di addestramento e tuning degli LLM di mercato è oneroso e complesso ma risulta differenziante per i vendor che investono notevoli risorse per aggiornare, pulire e arricchirne costantemente la conoscenza intrinseca dei loro modelli. Questo motiva le numerose partnership scaturite tra i protagonisti della AI e i principali cloud provider, tra cui la più nota OPEN AI con Microsoft, ma non

meno importanti sono quelle di altri attori del mondo AI con GCP e AWS.
Tuttavia, il pre-training non copre in modo sufficiente nessuna delle conoscenze strategiche fondamentali per l'impiego delle tecnologie in realtà corporate come TIM o altre Industry, in quanto il corpus di addestramento per definizione è costruito con dati "open" non vincolati da alcun IPR.
Emerge **dunque il terzo e fondamentale aspetto peculiare della tecnologia generativa in ambito enterprise, cioè la possibilità e la necessità di aggiungere autonomamente strati di nuova conoscenza proprietaria, specifica del dominio, al di sopra di quella nativa dei modelli pre-addestrati.**
Tutti i fornitori si stanno rapidamente allineando nel rendere disponibili tre diverse tecniche per abilitare questa stratificazione fornendo la possibilità di "Fine Tuning", di "Parameter Efficient Fine Tuning" piuttosto che di "Reinforcement Learning with Human Feedbacks" (FT, PEFT e RLHF). Non è possibile entrare qui nel dettaglio, ma è sufficiente dire che con queste tecniche lo strato di conoscenza addizionale rimane confinato nel perimetro enterprise in cui viene svolto, semplificando, inoltre, la fase di "prompt engineering" delle soluzioni complessive. Per contro, si tratta di tecniche generalmente complesse da applicare con investimenti addizionali in risorse computazionali, in up-skill del personale e di ridisegno dei processi DevOps, in modo da mantenere accurata la risposta fornita anche a fronte del congenito disallineamento evolutivo dei dati e/o del contesto nel tempo ("data e context drift").
E' dunque importante ed urgente, per le organizzazioni IT, attrezzarsi per approfondire e sfruttare in modo ottimale i driver sopra descritti, applicandoli in modo bilanciato e controllato, così da ottenere e verificare i benefici ed il ritorno dagli investimenti derivanti dall'adozione delle tecnologie nel dominio enteprise, sia in termini di produttività che di qualità e consistenza del risultato.

Questo passaggio non è per nulla scontato, in quanto, sebbene nel settore delle telecomunicazioni i potenziali casi di applicazione della generative AI siano molteplici e impattino l'organizzazione nel suo complesso, per il loro successo risulta cruciale impostare una governance agile e rigorosa, che guidi l'adozione dell'innovazione in modo che sia orientata alla profittabilità e sia conforme ai principi dell'etica e della compliance, oltre che alla sicurezza dei dati e delle soluzioni preposte al loro trattamento.

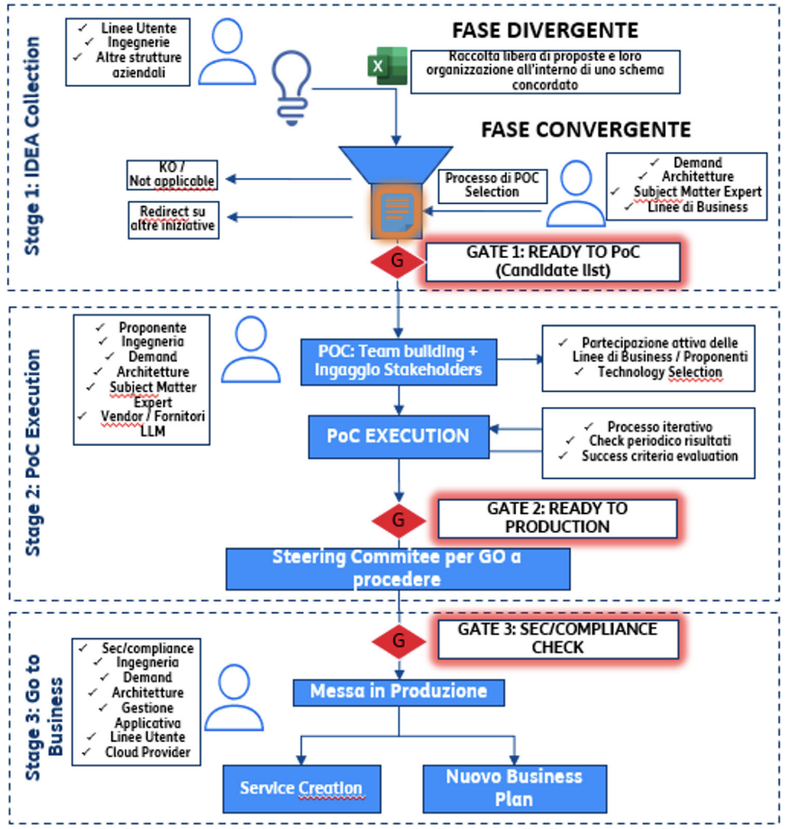
Il funnel di selezione e i pattern architetturali prioritari

L'IT di TIM, sulla base delle priorità' aziendali, ha avviato un processo di raccolta delle proposte e delle esigenze espresse internamente, per applicare la generative AI all'ottimizzazione dei rispettivi processi.
Su questi casi d'uso è elaborata una valutazione di opportunità, per ottenere un ranking di priorità delle iniziative da perseguire, identificando le aree di trasformazione dei processi impattati ed i relativi KPI/KPO che misureranno l'efficacia dell'adozione della generative AI nel tempo.
A seguito della fase di selezione, è previsto un framework operativo di "messa a terra" del funnel attraverso la predisposizione di una "Generative AI-FARM", che rappresenta sia un centro di osservatorio e presidio tecnologico che un nucleo di prototyping agile che sperimenta nuove soluzioni in campo con un ciclo di progettazione e realizzazione rapido, così da garantire l'acquisizione e dissemination della conoscenza in parallelo alla sperimentazione delle soluzioni di GenAI attraverso PoC ed applicazioni sui casi di utilizzo più rilevanti, ed alimentando un processo incrementale e progressivo di adozione ed industrializzazione via via più vasto (Fig.1).

Tra i primi risultati conseguiti dalla Generative AI-FARM, due in particolare sono rilevanti:
- assessment e monitoraggio delle soluzioni disponibili rispetto alla costante evoluzione della AI in ambito enterprise e consumer;
- pattern architetturali prioritari da realizzare per rispondere e supportare gli use case espressi dalle esigenze interne.

Sul primo punto si è constatato come le opzioni di deployment della Gen AI siano piuttosto complesse e in parte funzionalmente sovrapposte fra loro; per avere un

Figura 1: Il funnel della GENERATIVE AI-FARM di TIM

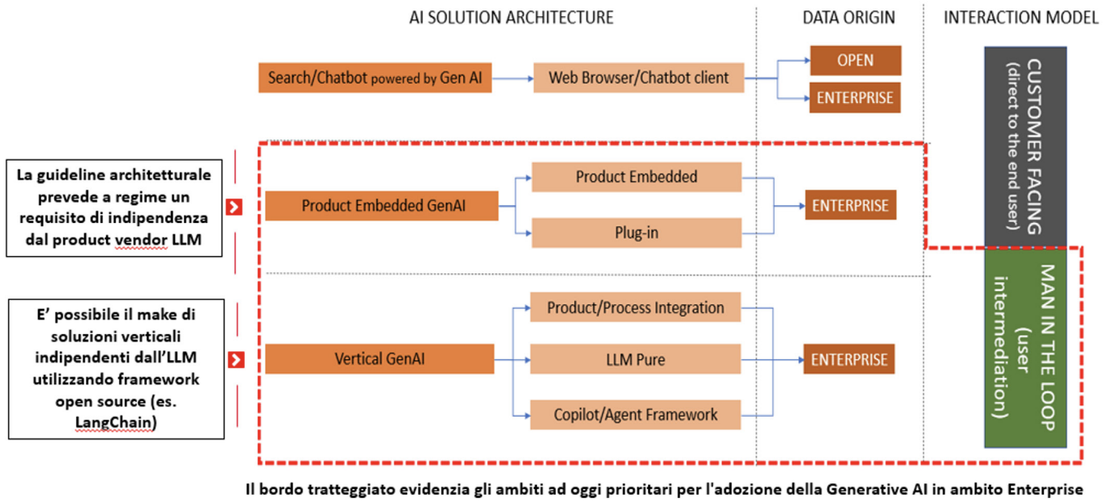


riferimento comune con cui classificarle e confrontarle si è definito lo schema di Fig.2, che indica le tre seguenti "dimensioni caratteristiche" per descrivere le soluzioni generative:

- **AI SOLUTION ARCHITECTURE**: è la dimensione che rappresenta la modalità con la quale gli LLM sono integrati all'interno della soluzione informatica complessiva. In prima analisi si possono riconoscere le tre seguenti modalità:
  - "Vertical Gen AI": una soluzione make, con tuning sui dati privati dell'azienda e realizzata con minimo lock-in utilizzando le APIs native fornite dai vendor LLM;
  - "Product Embedded Gen AI": qualsiasi soluzione third party che integra tecnologie LLM native con l'aggiunta di middleware e/o connettori proprietari;
  - "Search/Chatbot powerd by GenAI": in generale le architetture basate su chat engine oppure motori di ricerca pubblici e/o enterprise integrati in client nativi o in plug-in forniti dai produttori di LLM.
- **DATA ORIGIN**: questa dimensione indica se la soluzione consente l'utilizzo di dati enterprise e/o open;
- **INTERACTION MODEL**: descrive l'interazione diretta (customer facing) o intermediata (man in the loop) degli utenti finali con le tecnologie generative.

In prima battuta, nei contesti enterprise, il modello di interazione di tipo "Man In the Loop" si ritiene sia il più opportuno per gestire e controllare il rischio di allucinazione e tossicità degli LLM, cioè la generazione di risposte e contenuti non conformi a quanto atteso o contrari alle policy aziendali.
Relativamente ai pattern architetturali, dall'analisi dei casi d'uso è emerso che questi ricadono principalmente soluzioni di tipo "Vertical Gen AI" ed in particolare convergano in modo significativo sui due seguenti patten architetturali appartenenti alla categoria RAG ("Retrieval Aumented Generation"):

Figura 2: Schema di riferimento per le soluzioni Generative AI



Il bordo tratteggiato evidenzia gli ambiti ad oggi prioritari per l'adozione della Generative AI in ambito Enterprise

- **AI Summarization Chain**: a fronte di un sistema documentale contenente dati enterprise (immagini oppure testi), la soluzione consente di realizzare una catena di azioni più o meno complessa di "search", "reorganization", "formatting", "prompting" in grado di generare un testo o delle immagini riassuntive di uno o più concetti semantici da trasmettere al cliente finale dopo intermediazione umana (es. brief di servizi, questionari, valutazioni, contenuti di caring support, etc..);
- **AI Data Analytics**: a fronte di un data base enterprise contenente dati relazionali strutturati, oppure fogli di calcolo contenuti in sistemi di archiviazione, la soluzione consente di generare delle analisi dati che producano informazioni aggregate di sintesi, query specifiche dedotte da descrizioni in linguaggio naturale, grafici oppure tendenze da consumare internamente all'azienda, oppure trasmettere al cliente finale dopo intermediazione umana (es. BI reporting, interactive query, etc).

Moltissimi casi d'uso sono stati ricondotti a questi pattern nella cui prototipazione TIM sta attivamente investendo per perfezionarne l'applicazione in molteplici processi aziendali tra cui quelli del Caring, del Marketing Consumer e in generale nell'ottimizzazione all'accesso delle principali sorgenti informative aziendali, preservando gli aspetti di privacy/GDPR e compliance.

Conclusioni

In questo articolo stati illustrati i punti chiave della tecnologia alla base dei Large Language Model e la strategia che si sta adottando in TIM per censire, raggruppare, filtrare e implementare use case di GenAI che siano al contempo di impatto, in termini di ritorno dei benefici sui costi, e di utilità per una molteplicità di funzioni aziendali.

Sebbene i case di interesse promettano ritorni importanti in termini di aumento della produttività dei lavoratori, Customer Satisfaction e Revenue Generation, si pongono al contempo sfide di tipo tecnologico, normativo e anche di rischio reputazionale, da affrontare in modo organico, al fine di governare e mettere a valore le opportunità che questa rivoluzionaria tecnologia offre.■

Generative AI: Search & Summarization documentale

Una delle attività più time consuming nelle aziende, a prescindere dalle dimensioni e dal core business delle stesse, è la ricerca di informazioni chiave all'interno di grandi quantità di documenti, spesso di grandi dimensioni e nei formati più disparati (.pdf, .doc, .xls e/o altri formati proprietari). Immaginiamo, ad esempio, la necessità di recuperare una particolare clausola di un contratto stipulato con un fornitore, o di riassumere in poche righe i punti principali di un allegato tecnico, o la ricerca di un passaggio chiave in una procedura di installazione, soprattutto nei casi in cui queste informazioni sono "affogate" in cartelle o sistemi documentali che contengono centinaia di files di grandi dimensioni. Attività che, di norma, necessita di qualche ora di lavoro di ricerca, raffinamento, estrazione delle informazioni e generazione di un riassunto da poter inviare via e-mail al collega o riportare in un altro documento.

La problematica può essere ulteriormente estesa anche alla ricerca all'interno di log di sistema, transcript di chat o conversazioni del call center o del backoffice, fino ad arrivare all'analisi dei Social Media Trends, attività che coinvolgono elevate quantità di dati testuali.

La tecnologia dei Large Language Models che abilita la Generative AI ha dato risultati sorprendenti sul tema della Search and Summarization, e molti Use Cases raccolti con le strutture aziendali sono riconducibili al pattern architetturale relativo alla ricerca basata su NLP (Natural Language Processing). Utilizzando in modo opportuno le diverse tecnologie è possibile fornire al motore di Generative AI, tramite un'interfaccia di tipo conversazionale, un prompt in linguaggio naturale unitamente ai riferimenti della location dove sono storicizzati i documenti. Il motore risponde con il miglior completamento possibile del prompt stesso, che nel caso in questione corrisponde alla sintesi o elaborazione delle informazioni ricercate in uno o più documenti e, ove richiesto, i riferimenti ai capitoli dei documenti stessi.

L'architettura di riferimento prevede, per questa tipologia di use case, l'istanziazione di un singolo stack applicativo che soddisfi esigenze provenienti da strutture diverse (es. multi-tenancy), possibilmente agnostico rispetto alla tecnologia utilizzata e compatibile con aggiornamenti successivi del modello LLM sottostante (es. GPT-4 vs GPT3.5), così come rappresentata in Fig.A.

L'architettura si compone dei seguenti moduli logici:

- base documentale di partenza, con esposizione di API ReST per la retrieve dei documenti;
- modulo di data preparation: richiama le API di accesso ai documenti, li indicizza e gestisce la preparazione del contest (contenente il documento opportunamente tokenizzato) ed il prompt fornito dall'utente. In caso di uso dell'Embedding (descritto nel seguito) il modulo implementa anche il DB vettoriale della rappresentazione dei documenti;
- interfaccia conversazionale dove l'utente può inserire il prompt e ricevere la risposta;
- applicazione principale che coreografa i diversi passaggi;
- modello LLM di riferimento.

Nel disegno architetturale è necessario approntare alcuni accorgimenti: uno dei problemi principali è la tokenizzazione dei contenuti, in quanto i motori LLM hanno una limitazione intrinseca del contesto, in termini di numero massimo di token da fornire al modello insieme al prompt di domanda. La conformazione del token varia in base alla tecnologia (ad esempio Open AI utilizza una codifica BPE -Byte-Pair encoding- che corrisponde all'incirca ad una sillaba di una parola) e alla frequenza di occorrenza della coppia di caratteri o byte. Di seguito si elencano alcuni dei metodi utilizzati per passare il contesto (documento tokenizzato e prompt) al modello:

- Stuffing: si passa l'intero documento al modello in una singola chiamata. Poco adatto a documenti di grandi dimensioni (per il rischio di superare la dimensione massima del context), ha il vantaggio di essere più veloce e preciso "al primo colpo" nell'elaborazione della risposta.
- MapReduce: implementa una summarization a più fasi. Più adatto a documenti di grandi; dimensioni, prevede la successione di summarization di chunk del documento e la conseguente combinazione dei diversi summary per ottenere la risposta finale;
- Refine: richiede l'uso di due prompts, la domanda principale che genera l'output per il task successivo e il prompt di refine, che serve a raffinare la prima risposta. Il vantaggio rispetto al mapReduce è ottenere un contesto più rilevante, ma necessita di più chiamate conseguenti, in quanto rispetto al precedente, per sua natura non può parallelizzare i task.

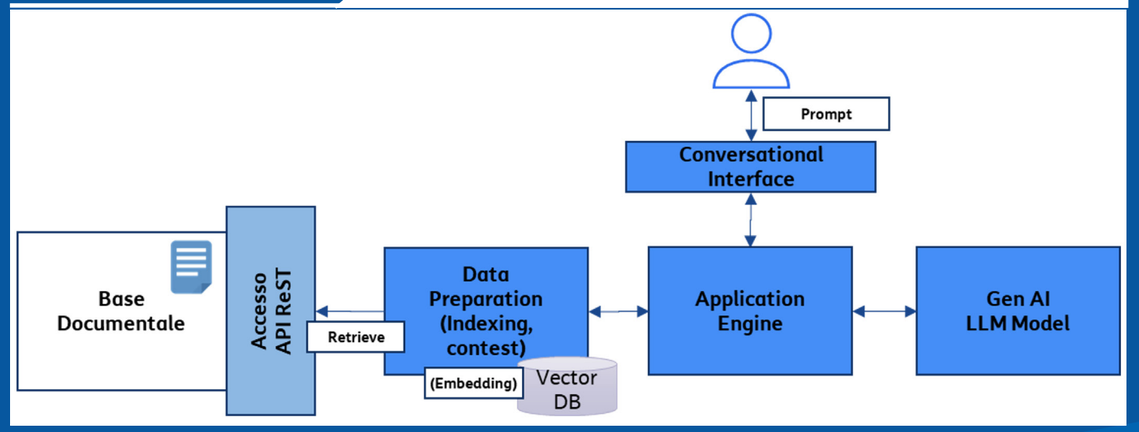
Un altro strumento estremamente potente utilizzabile in questo ambito è l'Embedding, ovvero una rappresentazione vettoriale di un testo. Tale rappresentazione fornisce capacità avanzate di "Semantic Search & similarity". Utilizzata come input al modello generale LLM permette di creare applicazioni di Search & Summarization documentale specifica su basi dati private come ad esempio repository di contratti, documenti aziendali, manuali di grandi dimensioni, mantenendo il contesto dati circoscritto al perimetro aziendale. Il principio di base è la rappresentazione dei documenti stessi, dei metadati associati ad essi e di altri possibili dati ausiliari tramite uno spazio vettoriale, inteso come vettori numeri in virgola mobile. L'embedding può poi essere indicizzato su un DB vettoriale, per permettere una ricerca a bassa latenza, vincolo critico per alcune tipologie di applicazioni su larga scala [4].

Le scelte tecnologiche per implementare le features richieste dipendono, infine, da vari fattori, tra i quali: costi di licenza per l'uso di uno o più prodotti vendor, opportunità di soluzioni più vicine al "chiavi in mano", o la possibilità di combinare tra loro capabilities fornite da strumenti open source, in base al grado di maturità e skills del team di sviluppo IT.

Un framework di sviluppo open source con un buon grado di completezza è, ad esempio, Langchain. Con esso è possibile sviluppare applicazioni basate su LLM in maniera "vendor-agnostic", scegliendo la tecnologia più adatta. Langchain fornisce strumenti per la creazione del contesto, per l'interfacciamento con basi di dati documentali, la possibilità di utilizzare i vari modelli forniti dai principali attori di mercato. La caratteristica principale è la possibilità di creare "chain", cioè catene applicative combinando vari moduli per costruire la soluzione più adatta al proprio requisito.

giovanni.petracca@telecomitalia.it

Figura A: Architettura di riferimento per applicazioni di Search & Summarization



Bibliografia

1. Acc – Accenture. (2023). Generative AI Workshop. Rome.
2. AWS – Altman Solon, Report finanziato da Amazon AWS – Telecommunications Generative AI Study – https://pages.awscloud.com/GLOBAL-other-DL-generative-ai-for-telecom-whitepaper-2023-learn.html
3. Azu – Azure OpenAI Services - Ritiro GPT-35-Turbo 0301 e GPT-4 0314 - https://learn.microsoft.com/it-it/azure/ai-services/openai/concepts/models#gpt-35-turbo-0301-and-gpt-4-0314-retirement
4. EU - EU AI Act: first regulation on artificial intelligence - https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
5. McK – McKinsey - Unleashing developer productivity with generative AI - https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/unleashing-developer-productivity-with-generative-ai
6. MS – Microsot Research – AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation – https://www.microsoft.com/en-us/research/project/autogen/
7. O2 - Virgin Media O2 - How Generative AI is Revolutionising the Telco Industry –https://medium.com/@vmo2techteam/how-generative-ai-is-revolutionising-the-telco-industry-first-steps-56a0d49ea6e4
8. OAI – Andrey Karpathy – OpenAI – State of GPT Speech - https://www.youtube.com/watch?v=bZQun8Y4L2A&ab_channel=MicrosoftDeveloper
9. Squ - https://squirro.com/pricing/
10. Wir – Wired - ChatGPT: George R.R. Martin, Jonathan Franzen e altri scrittori hanno fatto causa contro OpenAI - https://www.wired.it/article/chatgpt-george-r-r-martin-scrittori-americani-causa-open-ai/

Urlografia

- https://github.com/GoogleCloudPlatform/generative-ai/blob/main/language/use-cases/document-summarization/summarization_large_documents_langchain.ipynb?utm_source=pocket_saves
- https://learn.microsoft.com/en-us/semantic-kernel/prompt-engineering/tokens
- https://cloud.google.com/vertex-ai/docs
- https://cloud.google.com/vertex-ai/docs/generative-ai/embeddings/get-text-embeddings
- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks - Patrick Lewis, Ethan Perez , Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela - Facebook AI Research; University College London; New York University
- Large Language Models are Zero-Shot Reasoners - Takeshi Kojima The University of Tokyo Shixiang Shane Gu Google Research, Brain Team Machel Reid Google ResearchYutaka Matsuo The University of Tokyo Yusuke Iwasawa The University of Tokyo
- Attention Is All You Need - Ashish Vaswani∗ Google Brain Noam Shazeer∗ Google Brain Niki Parmar∗ Google Research Jakob Uszkoreit∗ Google Research Llion Jones∗ Google Research Aidan N. Gomez University of Toronto Łukasz Kaiser∗ Google Brain

Autori

Mario Bonnet *mario.bonnet@telecomitalia.it*

Ingegnere delle Telecomunicazioni, in Azienda dal 1998, si è inizialmente occupato di progetti di deployment di piattaforme di rete e servizi VAS, soluzioni di messaggistica, architetture SOA e Web Services per l'esposizione e l'integrazione di servizi/API in ecosistemi e centri servizi (CSP, ISP, MVNO, clienti business).
È stato responsabile, a partire dal 2013, della funzione Service Delivery Platform & Net API di Technology; poi dal 2014 ha assunto la responsabilità delle funzioni Network Function Virtualization, Telco Cloud e Core Network and Automation. Da settembre 2020 ha ricoperto diverse ruoli in ambito IT ed e' attualmente responsabile della funzione IT Corporate & Market Architecture and Security, con il compito di definire il piano tecnologico triennale e l'evoluzione delle architetture integrate IT. ■

Dario Mana *dario.mana@telecomitalia.it*

Dario Mana, ingegnere informatico, entra in TIM nel 2001 e lavora su progetti di innovazione: TimCafè, StarSIP, DTTRun, DynamicTV, WantEat. Nel 2013 entra a far parte dell'iniziativa Joint Open Lab presso il Politecnico di Torino. Dal 2017 al 2019 lavora nel team che ha ideato, sviluppato e messo in campo Angie, l'assistente virtuale, e dal 2020 si occupa di sviluppo di modelli di Natural Language Processing in aree di Innovation, Data Office e Strategy. Dal 2023 si occupa anche del tema della Generative AI e conseguenti valutazioni tecnologiche e della prototipazione di casi d'uso di interesse. ■

Clara Oliva *clarafabiola.oliva@telecomitalia.it*

Laureata con lode in Matematica Applicata all'economia e alla finanza nel 2011 presso la LUISS Guido Carli, ha conseguito un Executive MBA nel 2021 presso la Bocconi School of Management.
Approda in TIM nel 2016 dopo un percorso di crescita in aziende italiane e internazionali.
Dal 2020 è responsabile di un team multidisciplinare che si occupa di Ricerche di mercato, Data Strategy e guida lo sviluppo di soluzioni di Artificial Intelligence per Revenue and Operations. ■

Pier Carlo Paltro *piercarlo.paltro@telecomitalia.it*

Ingegnere elettronico, con Master in Telecomunicazioni, entra in Telecom Italia nel 1996. Lavora inizialmente nella standardizzazione ITU-T MPEG per lo sviluppo tecnologico e brevettuale dei servizi multimediali. Successivamente contribuisce con responsabilità crescente al lancio dei servizi IPTV&Media; al programma Open Innovation nell'ambito della API e IoT community; alle soluzioni Big Data progettando e gestendo l'Enterprise Data Lake di TIM. Si occupa di artificial intelligence dal 2015 guidando la costituzione di un team di data scientist per diffondere a largo spettro il machine learning in TIM, dall'ottimizzazione della rete al caring, ai processi di up/cross selling e in generale di customer value management.
Dal 2020 è Lead Data Architect operando nel gruppo IT come riferimento per i programmi di journey to cloud e di digital transformation del parco applicativo aziendale. Attualmente è IT lead per il programma di Generative AI adoption all'interno di TIM. ■

Le opportunità offerte dall'AI ad un operatore Telco

Cristina Persico, Angelo Solari

La portata della AI è così rivoluzionaria che solo le aziende in grado di adottare una strategia che integri la digitalizzazione, la comprensione dei processi, l'etica e la sicurezza saranno in grado di implementarla con successo e di coglierne l'enorme potenziale. In questo articolo esaminiamo le sfide per un operatore Telco.

Le applicazioni dell'AI nei processi di Operations TIM

Le nuove tecnologie dell'Intelligenza Artificiale, come la Generative AI e i Large Language Model, possono essere impiegate per massimizzare l'efficacia e l'efficienza dei processi operativi nel settore delle telecomunicazioni, consentendo la creazione di un numero illimitato di casi d'uso.
Ad esempio, queste tecnologie possono essere utilizzate per ottimizzare i consumi delle centrali, determinare le competenze necessarie per le operazioni di assurance, prevedere il successo o l'insuccesso di un intervento e fornire consulenza e supporto.

In TIM attualmente sono stati implementati due casi d'uso (VuCAB ed EnergIA) che raccolgono informazioni relative agli apparati, alle centrali e alla componentistica di rete al fine di elaborare indicatori e modelli di analisi predittiva/prescrittiva. Questi modelli sono progettati per supportare i processi di sorveglianza e diagnosi della rete, nonché per prendere decisioni volte a migliorare la sostenibilità e l'efficienza dei costi aziendali. Nel dettaglio:

- VUCAB: questo use case è finalizzato alla manutenzione proattiva degli apparati ONUCab. Comprende indicatori sviluppati per il monitoraggio dei cabinet, che consentono di identificare varie casistiche di guasto o disservizio. Queste casistiche possono includere blocchi o sospensioni dell'energia, allagamenti o condizioni meteorologiche avverse che influenzano il funzionamento dei dispositivi. Il progetto facilita il passaggio dall'identificazione del problema alla pianificazione dell'intervento necessario;
- EnergIA: questo use case è dedicato all'ottimizzazione e all'analisi dei comportamenti anomali legati ai consumi energetici. Comprende indicatori specifici sviluppati per monitorare i consumi energetici delle sedi TIM. Inoltre, offre una dashboard di esplorazione visuale che si concentra sulle variabili più influenti, come i dati meteorologici, ed evidenzia eventuali scostamenti che possono suggerire possibili sprechi di consumo energetico ingiustificati (5).

In questo contesto sono stati individuati ulteriori use case:

- ottimizzazione dell'allocazione dei tecnici per gli interventi di Delivery FTTH: riguarda la realizzazione di un algoritmo che consenta di suggerire la migliore allocazione MOS/MOI dei tecnici per gli appuntamenti di installazione FTTH. Questo use case coinvolge l'acquisizione dei dati relativi alle work request per la gestione degli appuntamenti e fornisce come output la colorazione degli indirizzi in base alla complessità degli interventi;
- riduzione dei KO: finalizzato alla riduzione dei casi in cui un ordine non può essere eseguito per motivi tecnici, noti come "KO di lavorazione". La riduzione di questi KO consentirà di migliorare l'efficienza delle attività dei tecnici sul territorio. In questo caso, si raccolgono dati sugli apparati, dati geografici e informazioni sugli ordini per cercare di prevedere l'esito della lavorazione.

Inoltre, si sta lavorando sull'implementazione di strumenti di Generative AI per sviluppare un chatbot avanzato.

Questo chatbot sarà alimentato con specifiche sugli apparati, linee guida sugli interventi e una Knowledge Base in continua evoluzione.
Il suo obiettivo è fornire supporto evoluto per indirizzare problemi di Delivery e di Assurance.

La Cloud Data Platform

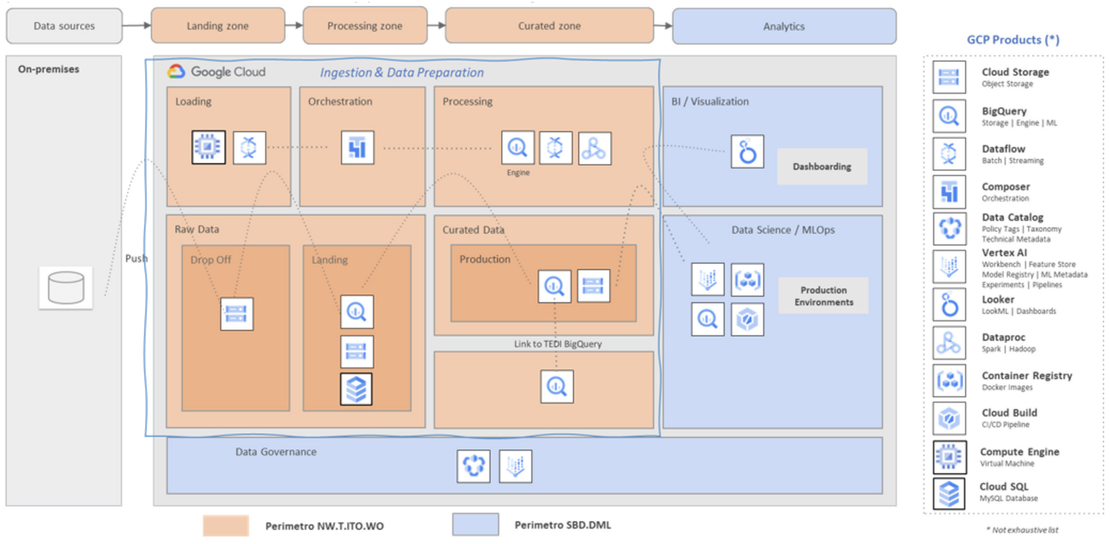
Per la realizzazione degli use cases VUCAB & EnergIA è stata deliverata su Google Cloud Platform una piattaforma la cui architettura è rappresentata di seguito nella Fig.1.
Si tratta di un ambiente dedicato alla Predictive Data Analisys, con possibilità di estensione e sviluppo su base use case, che, grazie all'accordo siglato con Google, consente di avere accesso agli strumenti analitici nativi offerti dal Cloud.

Tale piattaforma, che potrà essere arricchita gradualmente sia in termini di tool disponibili (ora Vertex, Looker, a tendere Palm e, in futuro, Gemini), sia come fonti di alimentazione e finalità d'uso dei dati (scopo predittivo), rappresenta il nucleo della nuova Cloud Data Platform a supporto dei processi di Wholesale & Operations.
Tale piattaforma renderà disponibile un ambiente unificato per Advanced Analytics che consentirà di superare i limiti delle attuali architetture on premise.

Lesson Learned

Da queste esperienze emerge che la disponibilità di strumenti in grado di analizzare enormi quantità di informazioni in modo automatico sta rendendo i dati una risorsa sempre più fondamentale per le aziende e un elemento centrale nei processi decisionali. È quindi fondamentale rafforzare questa consapevolezza e garantire che ogni evento generato dai processi aziendali venga immesso nel Data Lake, in modo da poter essere sfruttato per analisi predittive.

Figura 1: Architettura su GCP per use cases VUCAB & EnergIA



Inoltre, è altrettanto essenziale istituire una catena MLOps che gestisca l'intero ciclo del Machine Learning, che comprende l'ingestion, l'analisi, la trasformazione, la modellazione, l'addestramento, la valutazione, il deploy e la previsione. Questo è cruciale per implementare un processo di Continuous Integration/Continuous Deployment (CI/CD), che migliora la qualità degli artefatti prodotti, riduce gli errori e il tempo in fase di deploy.

Ed ancora. Il Cloud rappresenta un acceleratore nella realizzazione di soluzioni, che possono essere sperimentate attraverso Proof of Concept (POC) con tempi significativamente più rapidi rispetto alle soluzioni tradizionali. Non è più necessario effettuare investimenti ingenti per creare una catena MLOps. L'approccio vincente è lavorare in modo agile per ogni caso d'uso, coinvolgendo un Product Owner nel processo.

L'utilizzo di nuove tecnologie su ambienti Cloud richiede il potenziamento degli skill dei gruppi di lavoro, con formazione mirata sia su aspetti architetturali che di sviluppo. Nasce in TIM la figura del Data Scientist, un professionista specializzato nella raccolta e analisi di grandi quantità di dati, che combina informatica, statistica e matematica per elaborare e modellare i dati, interpretando i risultati per fornire indicazioni strategiche.

L'impiego di strumenti di Generative AI può cambiare il modo di lavorare e le competenze richieste su diversi ambiti (es. tecnici on field). La disponibilità di un assistente conversazionale basato sull'AI che offra suggerimenti in tempo reale per le risposte (cosiddetto co-pilot), può aiutare ad aumentare la produttività, soprattutto nel caso di colleghi meno esperti, e la soddisfazione per i clienti finali.

Conclusioni

L'Intelligenza Artificiale può offrire numerose opportunità alle aziende di telecomunicazioni, tra cui il miglioramento dell'esperienza dei clienti, il co-pilot delle risorse, l'ottimizzazione della gestione delle reti, l'automazione dei processi operativi, la manutenzione predittiva e il supporto all'elaborazione di grandi volumi di dati per estrarre insights utili.

Il valore effettivo che si riuscirà ad estrarre da queste opportunità tecnologiche dipende dal grado di adozione dell'AI nei diversi processi e dalla capacità di individuare e implementare i casi d'uso in grado di generare il maggior valore.

Le aziende che riusciranno per prime ad ottenere risultati concreti, indirizzando anche le sfide legate alla privacy dei dati e alla sicurezza, potranno acquisire un considerevole vantaggio competitivo.■

Bibliografia

1. https://www.ey.com/it_it/beyond-la-nuova-serie-tv-di-ey/intelligenza-artificiale-una-rivoluzione-tra-opportunita-e-rischi
2. https://www.digital4.biz/executive/ai-cos-e-l-intelligenza-artificiale-e-come-puo-aiutare-le-imprese/
3. https://blog.osservatori.net/it_it/storia-intelligenza-artificiale
4. https://exploreai.org/p/generative-ai
5. L'Intelligenza Artificiale per il saving energetico delle centrali. Roberta Giannantonio, Cristina Persico, Federico Tebaldi, Alena Trifirò, https://www.gruppotim.it/it/newsroom/notiziario-tecnico-tim/Anno-2023/n2-2023/Intelligenza_Artificiale_per_il_saving_energetico_delle_centrali.html

Autori

Cristina Persico *cristina.persico@telecomitalia.it*

Laureata con lode in Ingegneria Informatica nel 1996, inizia a lavorare in DATAMAT nella Space & Environment Unit. Nel 1998 entra a far parte di TIM nei sistemi informativi dedicati alla rilevazione delle frodi nel traffico mobile. Dal 2015 coordina un gruppo di lavoro impegnato nei sistemi di CRM e Order Management. Di recente, ha assunto il ruolo di responsabile della funzione Wholesale & Operations Data Platform dell'Area IT Operations & Wholesale Systems di TIM. ■

Angelo Solari *angelo.solari@telecomitalia.it*

Laureato in Ingegneria Elettronica, nel 1988 inizia a lavorare presso Esacontrol, occupandosi della progettazione e sviluppo di sistemi di automazione in ambito automotive. Dopo aver lavorato nel centro ricerca IBM, nel 1995 entra in Sodalia, joint venture tra Telecom Italia e Bell Atlantic, dove fino al 2000 coordina progetti di sviluppo OSS per Bell Atlantic. Dopo l'integrazione di Sodalia in Azienda diventa inizialmente responsabile della SW Factory OSS, in seguito delle funzioni di Ingegneria Sell To Delivery e Usage to Cash, poi di Sviluppo e Operations di tutti i processi a supporto del Billing per Telecom Italia. Dal 2014 coordina le attività di sviluppo ed Operations delle Applicazioni OSS & Wholesale. ■

Large Language Model per il processo documentale TIM

Luca Buriano, Barbara Rinero, Marco Sapienza, Rossana Simeoni

La **Generative AI** è la tecnologia su cui si focalizza il progetto di innovazione TIM denominato “Hybrid Intelligence & Advanced Communications” per individuare nuove opportunità per le attività operative di TIM. Il progetto intende mantenere una forte coniugazione tra gli aspetti “Human” e “not Human” approcciando ad una *Hybrid Human-Artificial Intelligence* [1]. Vediamo come in questo articolo.

Da un punto di vista tecnologico la Generative AI [2] vede come fulcro i **Transformer**, particolari algoritmi di deep learning che con **il concetto di “attenzione”** introducono la capacità di apprendere relazioni tra elementi in input accrescendo la capacità di apprendimento e generativa [3].
Open AI con GPT e ChatGPT è stata dirompente ed ha portato alla democratizzazione di questi nuovi modelli svelandone il salto paradigmatico dalla necessità di modelli specifici per compiti specifici alla disponibilità di un unico modello capace di supportare in più compiti, dalla classificazione, alla summarization, etc [4].
Da un punto di vista applicativo e di business il contesto è sicuramente dominato da grandi Player come **Microsoft** e **Google** [5,6] ma vede anche una miriade di Startup e Centri di eccellenza che propongono soluzioni [7].
A fronte di questa rivoluzione, TIM ha scelto di investire nell’ambito della Generative AI anche in un ambito di innovazione specifico quale **il processo di automazione documentale**.

Il fenomeno ChatGPT: Generative AI e Large Language Model

Con l’avvento di ChatGPT nel novembre 2022 si è diffusa l’idea che possa esistere un’Intelligenza Artificiale generale in grado di automatizzare compiti, fornire idee creative e persino scrivere software, sia per aziende che per consumatori. ChatGPT, acronimo di Chatbot Generative Pre-trained Transformer, si basa su un modello di AI ovvero un **Large Language Model**, nella fattispecie GPT [8, 9], addestrato su grandi quantità di testo al fine di **comprendere** e **generare** testo coerente.

Un ruolo fondamentale nel campo del **Natural Language Processing (NLP)** è legato ai **Transformers** in quanto modelli di apprendimento automatico in grado di “catturare” le relazioni complesse tra parole e concetti.
Questi modelli, introdotti nel 2017 nell’articolo “Attention Is All You Need” [2], sono alla base del Generative Pre-trained Transformer (GPT) e sono composti da due componenti principali: l’encoder, che codifica l’input in una rappresentazione vettoriale numerica comprensiva delle relazioni tra parti del testo, e il decoder, che genera l’output basato su questa rappresentazione per produrre un testo coerente con il significato dell’input.

I pre-trained transformers hanno permesso di superare le difficoltà di codifica delle parti lessicali, sintattiche, grammaticali, etc... grazie alla capacità di apprendere automaticamente le relazioni tra le parti dei dati in input, detti **Token**.
Sarà quindi il Token l’unità fondamentale per: la trasformazione del testo in vettori di

Figura 1: Prompt engineering

numeri, **embeddings**; il **fine tuning** ovvero addestramento su aspetti specifici; il **prompting**, istruzione data per la generazione del testo desiderato; la **completion** ovvero il contenuto generato.

Il modello matematico lavorerà sui token di cui è costituito l'input per apprendere e genererà il risultato fornendo una successione di token su base probabilistica/stocastica.
In questo senso il modello di generative AI e GPT nella fattispecie non sono "consapevoli" della bontà del contenuto generato seppur plausibile e grammaticalmente corretto.
Il **prompt engineering** si configura quindi come una nuova disciplina per assicurare controllo e correttezza tra quanto si desidera ottenere ed il risultato fornito dal modello.
Fake e allucinazioni mascherate dalla plausibilità del testo generato devono quindi essere un punto di forte attenzione ai fini dell'utilizzo professionale e quotidiano di questi sistemi di Intelligenza Artificiale.
Nell'era del prompt, dunque, saper fare le domande diviene una competenza determinante, perché si passa dalla ricerca delle informazioni corrette, alla formazione della corretta ricerca. Nasce, così, la cosiddetta promptologia, cioè la nuova abilità ingegneristica di ideare domande per spingere l'Intelligenza Artificiale a fornire risposte appropriate che richiede anche la comprensione del linguaggio e dell'espressione.
A valle di questa veloce disamina degli elementi tecnico – scientifici alla base della Generative AI, possiamo dire che i Large Language Models (LLM) come GPT possono essere sfruttati come:

1. un mezzo per dialogare in linguaggio naturale con una macchina, e quindi come strumento di interazione uomo-macchina per comunicare in linguaggio naturale;
2. come strumento di una Intelligenza Artificiale generale che, se addestrata costantemente e su grande quantità di dati, può indurci a immaginare la nascita di un'entità senziente o per lo meno in grado di generare informazione plausibile, ma non necessariamente corretta data la natura statistica di questa generazione.

Applicazione e impatto sul processo documentale in ambito Operation

Tra i progetti di Innovazione si è scelto di concentrarsi sull'automazione dei processi documentali a supporto delle Operations di TIM, con l'obiettivo di mettere la Generative AI al servizio di una fruizione più immediata dell'informazione e una semplificazione dei processi di diffusione della conoscenza.

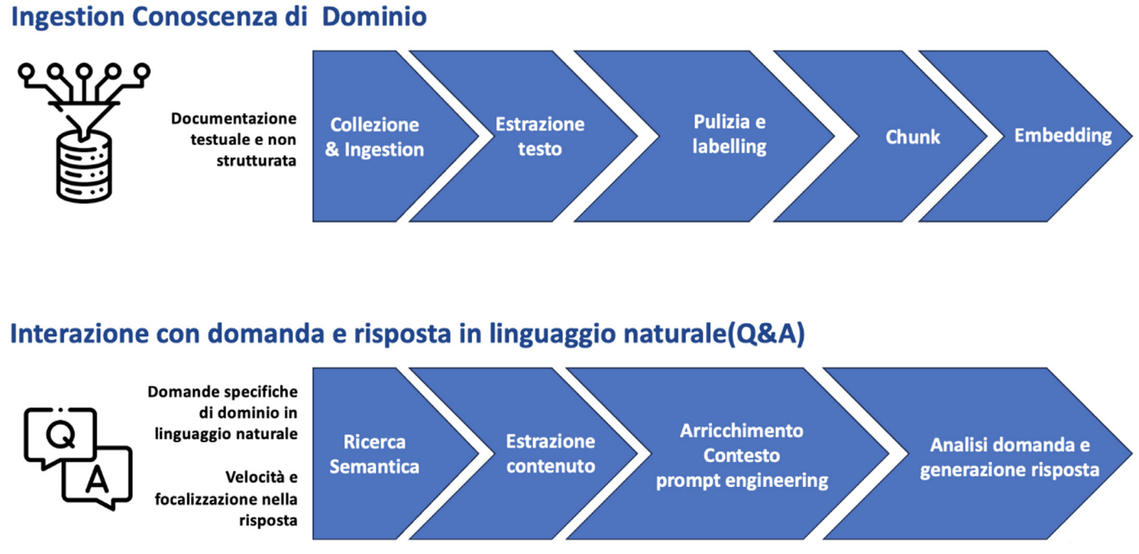
Di seguito vengono descritte le fasi che hanno portato alla costruzione di un **Proof of Concept di Questions & Answers (Q&A)** basato sulla generazione automatica di risposte a domande comuni estraendo informazioni puntuali da contesti sparsi su documenti eterogenei e non strutturati.
La costruzione del Prototipo di Q&A può essere divisa in due macrofasi: **l'ingestion della base di conoscenza e l'interazione con domanda e risposta in linguaggio naturale**.
La fase di **ingestion** ha lo scopo di trasformare la collezione di documenti in opportuni vettori numerici, Embeddings.

Questa prima fase è di fondamentale importanza e ci ha visto affrontare problemi legati all'eterogeneità della documentazione sia in termini di struttura che di tipologia.

Da un punto di vista pragmatico è stata compiuta un'attività di pre-processing sulla documentazione fornita dagli esperti di dominio che è stata di fondamentale importanza per migliorare le performance del Q&A.
Come ogni applicazione data driven, anche nel caso di un Q&A abilitato da Generative AI, è necessario che la sorgente di partenza sia epurata, laddove possibile, di informazioni non necessarie, attraverso operazioni quali esplicitazione della conoscenza, arricchimento tramite labelling, pulizia.
Il tutto per favorire una migliore ricerca del contenuto informativo ed una segmentazione della conoscenza che possa favorire una risposta precisa e puntuale, anche nel caso in cui la domanda fosse posta in modo generico.
Queste operazioni favoriscono la successiva suddivisione in più parti chiamate chunk e vettorizzate tramite l'operazione di embedding, in modo da abilitare i meccanismi statistici di cui discusso nella sezione precedente.

Da un punto di vista architetturale possiamo evidenziare come, soprattutto nel caso di documentazione eterogenea, sia necessario sviluppare o avvalersi di strumenti in grado di dare uniformità alla sorgente informativa e successivamente di tecnologie in grado di gestire il ciclo di vita di una nuova forma di salvataggio delle informazioni, i database vettoriali.
Essi, infatti, a differenza dei DB tradizionali, consentono operazioni di ricerca e analisi avanzate basate sulla similarità tra vettori, rendendo più efficace il recupero di informazioni rilevanti. Infine, il cuore del prototipo: i modelli di Generative AI.

Figura 2: Macro Fasi della soluzione di Q&A basata su generative AI



Ma basta soltanto dare delle informazioni corrette per far rispondere il modello in modo corretto?
La seconda fase di **interazione in linguaggio naturale** ha invece risvolti importanti dal punto vista della modalità di generazione della risposta.
Tale interazione è mediata da un modello di Generative AI che permette di trasformare la domanda in un vettore numerico, embedding, abilitando così una ricerca semantica nello spazio vettoriale della base di conoscenza. Il risultato di questa ricerca potrà così essere ritrasformato in testo per fornire una risposta in linguaggio naturale.

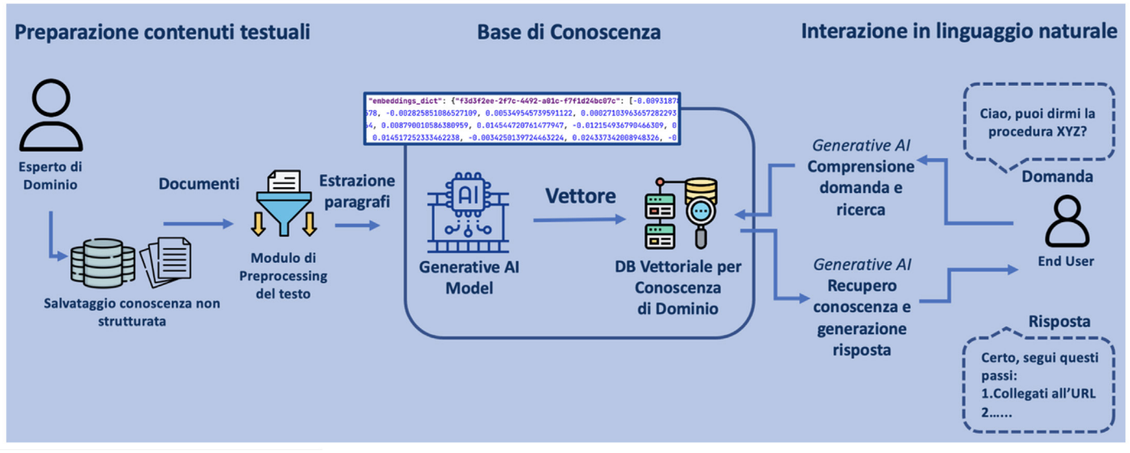
Ma come essere sicuri che la risposta sia adeguata all'interlocutore?
A questa domanda non vi è una risposta univoca, in quanto molto dipende dallo Use Case. Possiamo dire in generale che un'operazione di arricchimento della domanda posta attraverso una contestualizzazione porta il modello ad essere più allineato a quanto atteso dall'interlocutore ovvero dall'utilizzatore finale del Q&A; questa attività di arricchimento è propria del **prompt engineering**.

Le due macrofasi descritte mettono in luce nuovi compiti e ruoli che ricadono nel concetto dello "human in the loop", ovvero della necessità di un controllo ed una validazione continua della soluzione basata su Generative AI da parte degli esperti di dominio della conoscenza specifica.
E' altresì necessario semplificare molto la diffusione e il reperimento dell'informazione dalla fonte al suo utilizzo di valore nei processi operativi come, ad esempio, quello dei tecnici presso casa Cliente o altro.

Conclusioni

Il lavoro svolto fino ad ora ci ha permesso di fare esperienza sulla Generative AI ed analizzare competenze e ruoli professionali che impatteranno nei prossimi anni.

Figura 3: L'architettura logico-funzionale della soluzione



Le tecnologie sono state validate attraverso lo sviluppo del prototipo di Q&A, con l'obiettivo di dimostrare l'efficacia della Generative AI, opportunamente istruita e supervisionata dall'uomo, nell'automazione dei sistemi documentali; ovvero nella semplificazione e velocizzazione dell'accesso all'informazione, nell'organizzazione della conoscenza, nella revisione ed ottimizzazione dei processi di creazione della documentazione tecnica e commerciale.

Le prime evidenze ottenute sono incoraggianti e permettono di tranquillizzarci circa il timore di perdere il controllo dei sistemi basati sulla Generative AI e di non riuscire ad indirizzarli secondo i nostri desideri.
Sarà però solo la messa a terra di una sperimentazione in campo a permetterci di misurare gli impatti di una soluzione tecnologica basata sulla Generative AI sullo Use Case descritto e su scenari futuri.■

Generative AI per arte e design nel Metaverso

Una nuova generazione di artisti, designer ed architetti sta utilizzando la Generative AI come strumento di espressione creativa; inoltre, la diffusione di software come Stable Diffusion, Midjourney e DALL-E, basati su modelli quali Diffusion Models e Transformers, ha reso accessibili al grande pubblico queste tecnologie.

In TIM stiamo esplorando le tecniche di generazione tramite AI di contenuti multimediali per il metaverso; strumenti chiave sono:

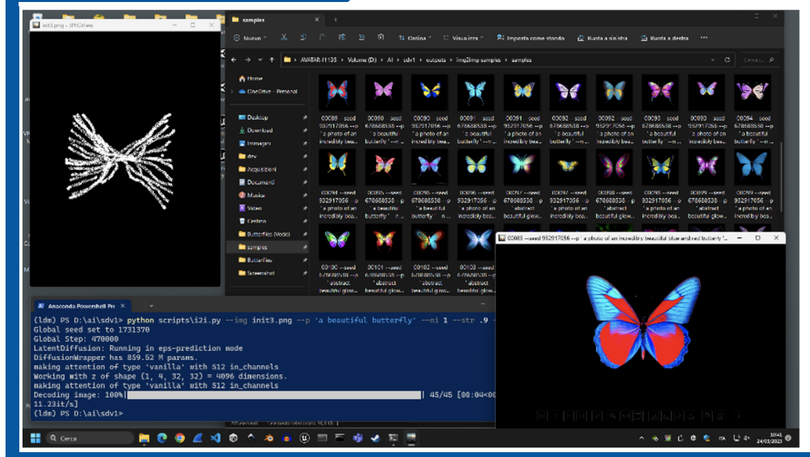
- prompt engineering: creazione di una descrizione testuale funzionale ad ottenere al meglio il contenuto desiderato;
- creazione di un pattern/contenuto iniziale che guiderà il modello di AI nella generazione dei risultati;
- fine tuning: riaddestramento parziale dei modelli di generative AI allo scopo di personalizzarli su contesti specifici;
- selezione: scelta, tra i molti risultati ottenibili, di quelli che soddisfano al meglio i criteri funzionali ed estetici di interesse.

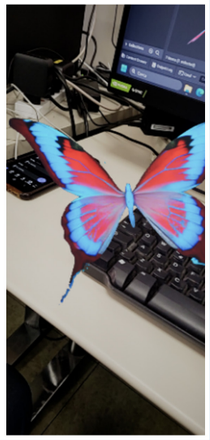
In TIM è stata creata un'installazione artistica interattiva in Augmented Reality, chiamata "Butterflies", come Proof of Concept del processo creativo e tecnologico mirante alla fusione tra Generative AI ed esperienze nel Metaverso.

La realizzazione si è articolata in:

- generazione di forme e disegni delle ali di farfalle virtuali, tramite prompt engineering, impostazione di un pattern visivo iniziale e selezione guidata da criteri estetici. Come software di Genertive AI è stato utilizzato Stable Diffusion;
- trasformazione dell'immagine delle farfalle generate in oggetti 3D, tramite algoritmi di visione artificiale ed applicazione del software di modellazione Blender;
- animazione del battito delle ali e simulazione del volo delle farfalle, tramite Blender e l'XR engine Unreal Engine 5 (UE5);
- sviluppo in UE5 di un'app in Augmented Reality che permette di creare farfalle virtuali nel Metaverso, osservarne forme, disegno e volo, ed interagire con esse.

Figura A: Installazione artistica frutto di fusione tra Generative AI e Metaverso





Bibliografia

1. Hybrid Human-Artificial Intelligence (computer.org)
2. “Attention Is All You Need” - Vaswani et al., 2017
3. What is Generative AI? | World Economic Forum (weforum.org)
4. openai.com
5. https://learn.microsoft.com/it-it/azure/ai-services/openai/overview
6. https://cloud.google.com/vertex-ai?hl=it
7. Generative AI: A Creative New World | Sequoia Capital US/Europe
8. “Language Models are Few-Shot Learners” - Tom B. Brown et al., 2020
9. https://www.marktechpost.com/2023/03/21/a-history-of-generative-ai-from-gan-to-gpt-4/

Autori

Luca Buriano *luca.buriano@telecomitalia.it*

Laureato in Matematica all’Università di Torino, lavora in TIM a temi quali Metaverso, Artificial Intelligence, Information Visualization e Human-Machine Interaction, alla ricerca della sintesi tra algoritmi, tecnologia e arte. ■

Barbara Rinero *barbara.rinero@telecomitalia.it*

Dopo essersi laureata in Economia e Commercio presso l’Università di Torino ed aver partecipato al Corso avanzato in Management, Economia e Diritto dei Servizi a Rete (MEDiR) presso il MIP Politecnico di Milano, entra in TIM nel 2001 occupandosi di metodologie di accounting regolatorio. Dal 2004 si occupa del disegno di servizi innovativi in ambito fisso, mobile e TV per clientela consumer, business ed interna dell’azienda. Tra gli altri, ha curato lo User Experience Design di TIM Wallet, TIMpersonal, TIM Developer, di soluzioni prototipali di realtà aumentata, virtuale, mista e di soluzioni prototipali di comunicazione evoluta nel Metaverso con avatar, ologrammi, video 360°. Attualmente è responsabile, all’interno di Service Innovation, delle attività di Hybrid Intelligence Service Design sul tema dell’Intelligenza Artificiale generativa. ■

Marco Sapienza *marco1.sapienza@telecomitalia.it*

Laureato in Ingegneria Informatica, nel 2006 ha intrapreso il suo percorso presso TIM all’interno delle Operations. Il suo percorso professionale lo ha visto spostarsi successivamente in Service Innovation, dove ha concentrato i suoi sforzi su progetti innovativi e standard internazionali come OMA DM e One M2M. In particolare, Marco Sapienza ha lavorato intensamente su temi legati all’Internet of Things (IoT) ed al Data Intelligence.
Oggi lavora all’interno del progetto “Hybrid Intelligence and Advanced Communication”, dove ricopre il ruolo di task leader responsabile della parte tecnologica del progetto, su temi quali Generative AI, Natural Language Processing, Graph Neural Network, impegnandosi costantemente nel cercare soluzioni innovative e modelli avanzati di Intelligenza Artificiale e Machine Learning che possono apportare un miglioramento significativo ai processi aziendali di TIM. ■

Rossana Simeoni *rossana.simeoni@telecomitalia.it*

Rossana Simeoni received the master degree in computer science and she has been applying her competences in Telco R&D departments since 1992. Her interests include Interaction Design, Intelligent Interactive Systems, Conversational Agent, Natural Language Processing. She is innovation project manager at TIM and adjunct professor at the University of Torino.
Rossana Simeoni ha conseguito la laurea magistrale in Informatica ed applica le sue competenze nei dipartimenti di R&D ed Innovation in ambito TELCO dal 1992. I suoi interessi includono Interaction Design, Sistemi interattivi intelligenti, Conversational Agent, Natural Language Processing. È Innovation Project Manager presso TIM e professore a contratto presso l’Università di Torino. ■

AI & Machine Learning per la rete TIM

Aurelio Giammusso, Roberta Giannantonio, Michele Ludovico, Andrea Marafante

Le potenzialità dell'Intelligenza Artificiale e del machine learning sono utilizzate, con grande successo, dagli operatori di telecomunicazioni principalmente per ottimizzare la gestione della propria rete con l'obiettivo di fornire una sempre maggiore qualità del servizio ai propri clienti e razionalizzare gli investimenti.

Mai come in questo periodo il termine Intelligenza Artificiale è sulla bocca di tutti e spesso accompagnato da falsi miti e infondate paure sul futuro. È indubbio che l'avvento dell'AI generativa e in particolar modo la disponibilità di motori generativi quali, tra gli altri, ChatGPT [a] e Midjourney [b] ha dato a tutti la possibilità di toccare con mano una tecnologia che può sembrare magica, misteriosa e generare paure ed incertezze. Le potenzialità delle soluzioni di AI sono però utilizzate da molte aziende da diversi anni per estrarre valore ed informazioni oggettive dai dati, aiutando in tal modo a prendere decisioni manageriali sulla base delle evidenze oggettive, migliorare i processi interni e fornire servizi di qualità sempre maggiore verso i clienti.

In TIM sono presenti le competenze multidisciplinari legate alle tematiche di AI che hanno reso possibile la realizzazione e l'utilizzo in campo di soluzioni, principalmente basate su approcci di Machine Learning, per l'ottimizzazione della rete TIM.

Un approccio multidisciplinare

Le competenze richieste per ideare e realizzare un progetto *data driven* di impatto, con reali ricadute sul core business aziendale, sono variegate, intrafunzionali e vanno ben oltre la pura componente algoritmica, la cui qualità rimane, per altro, fondamentale per la riuscita dell'attività.

Nella fase iniziale del lavoro, la visualizzazione dei dati è lo strumento più efficace per definire dove direzionare le attività in base ai dati a disposizione e alle prime evidenze visuali: in tutte le fasi del progetto le competenze di **data visualization** e **user experience** rivestono un ruolo cruciale per definire gli obiettivi e condividere i risultati. Una volta identificato il potenziale caso d'uso si passa alla raccolta e all'analisi dei dati. In questa fase sono fondamentali le piattaforme software che mettono a disposizione dei **data analyst** non solo i dati stessi, ma anche gli strumenti per l'analisi statistica e le attività di *data quality*, che tipicamente occupano una buona parte delle attività. Le fasi di analisi dei dati si svolgono tipicamente in parallelo con i primi approcci algoritmici, in cui i **data scientist** sperimentano i primi modelli anche per capire come indirizzare al meglio le attività. Le successive fasi del progetto procedono, in modo iterativo, dalla realizzazione dei modelli di machine learning, alla condivisione dei risultati con gli esperti di processo delle linee operative (esperti di dominio), che devono fornire i feedback necessari per il corretto addestramento del modello, fino ad arrivare ad un primo trial per valutare in campo le prestazioni e l'utilità della soluzione ideata. Le tecniche di AI e ML utilizzate sono le più svariate, da semplici modelli ad albero, alle più complesse reti neurali profonde (**Deep Learning**) o all'apprendimento dinamico del **Reinforcement Learning**, fino ad arrivare alle tecniche generative (**Gen AI** e **LLM**). Le competenze sono sempre in continuo aggiornamento anche grazie alla collaborazione con importanti atenei italiani, quali ad esempio il Politecnico di Torino, e la partecipazione ad associazioni di settore, quale ad esempio l'associazione italiana di Intelligenza Artificiale (AIxIA) di cui TIM è socio. Spesso però sono testate le più evolute soluzioni algoritmiche e poi viene selezionata la soluzione con il miglior compromesso tra le prestazioni e i requisiti infrastrutturali necessari: si preferisce ad esempio un algoritmo "ad albero" rispetto ad una rete neurale se le differenze prestazionali non giustificano la necessità di un hardware specifico necessario per far girare la soluzione di deep learning. Un grande supporto per rendere disponibili velocemente in campo le soluzioni ideate è fornito dalle piattaforme cloud IT per il ML grazie al paradigma dell'**MLOps** (vedi articolo "Le opportunità offerte dall'AI ad un operatore Telco"- Notiziario Tecnico TIM 3-2023).

In contesti specifici, quali ad esempio la configurazione dinamica della rete, i requisiti di bassa

latenza e alta affidabilità necessitano di un approccio **EdgeML** "*train in the cloud, infer at the edge*", che sfrutta le potenzialità dell'Edge Computing per rendere disponibile la logica algoritmica in modo decentralizzato ai bordi della rete, lì dove è necessaria.
I progetti sono inoltre realizzati in modo olistico e responsabile (**Responsible AI**) non limitandosi solo agli aspetti tecnici, ma abbracciando ambiti più ampi, in completa sinergia con le funzioni aziendali competenti, quali gli aspetti regolatori, il rispetto delle linee guida etiche dell'azienda e l'analisi degli impatti che le soluzioni di AI possono avere sulla società.

I dati della rete TIM

La rete TIM, con le sue oltre 10.000 centrali di rete fissa, 135.000 apparati stradali e 74.000 nodi radiomobili sparsi per tutto lo stivale, produce una enorme quantità di dati, preziose "tracce digitali" del suo funzionamento che, se correttamente analizzate, possono fornire utili indicazioni sullo stato attuale della rete e su possibili ottimizzazioni anche in ottica predittiva. Per poter trarre utili informazioni i dati stessi devono essere completi, digitali e affidabili: in altre parole la *data quality* è una prerogativa fondamentale per costruire algoritmi di successo. I dati infatti sono le fondamenta su cui si basano gli algoritmi di Machine Learning che riescono a fornire nuove informazioni solo se sono addestrati su dati che descrivono correttamente e in modo completo il contesto di interesse.
Le potenzialità predittive degli algoritmi di Machine Learning sono sfruttate a pieno quando si riesce ad applicare con successo il *supervised machine learning*, ovvero si ha una base dati da cui gli algoritmi possono correttamente apprendere il contesto di riferimento.
È quindi importante, per una transizione verso una *AI native company*, diffondere la data culture a tutti i livelli dell'organizzazione, per progettare i processi con il criterio fondante di generare dati affidabili, digitali e ben strutturati. Le prime attività di TIM in tale contesto si sono focalizzate su casi d'uso a supporto della supervisione del funzionamento delle reti di accesso fissa e mobile (*assurance*), della diagnosi di guasti della rete fissa e dell'ottimizzazione della rete radiomobile (*planning*), come sarà approfondito nei prossimi paragrafi.

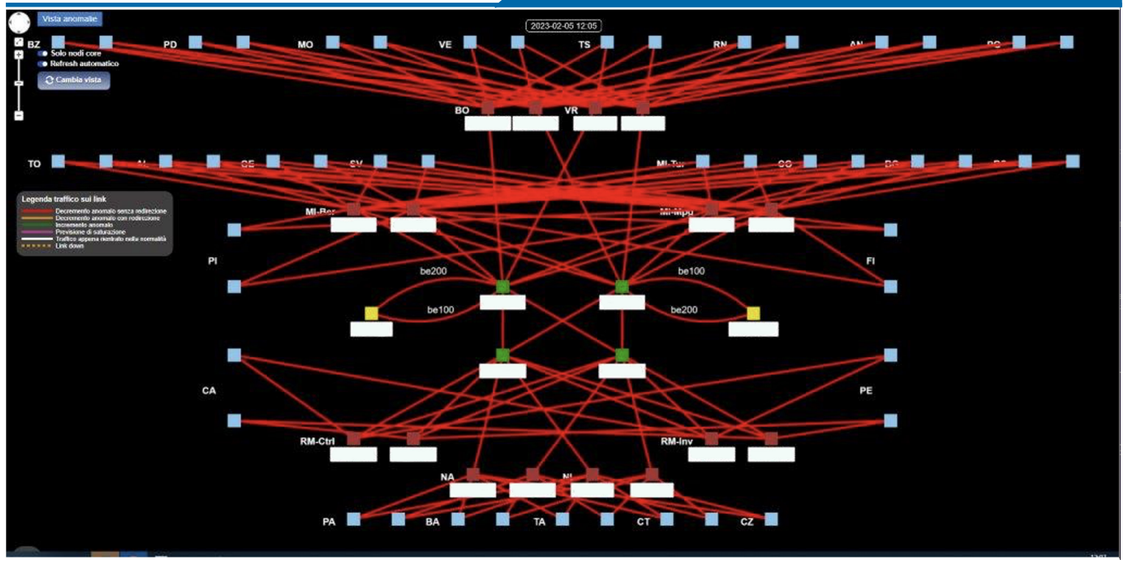
Un altro grande settore in cui estrarre valori dai dati è quello della manutenzione (*maintenance*) degli apparati di rete.
In parallelo sono in corso altre attività quali ad esempio l'ottimizzazione dei processi di installazione e configurazione dei servizi sottoscritti dagli utenti (*delivery*). Con particolare riferimento all'installazione presso la casa del cliente del servizio di FTTH (Fiber To The Home), si stanno sviluppando modelli predittivi per pianificare al meglio gli interventi.
Anche nel contesto di ottimizzazione dei consumi energetici (*energy*) sono in corso attività, come già descritto in [1], sulla visualizzazione dei dati ed algoritmi di *anomaly detection*.

Machine Learning per la gestione degli allarmi di rete

I dispositivi della rete TIM inviano numerose segnalazioni sul loro funzionamento, ma solo le più critiche, che possono evidenziare l'insorgenza di un disservizio, devono essere analizzate da personale specializzato per intervenire in tempo e ripristinare la normale operatività. In questo contesto è attivo da quattro anni il sistema **PANAMA** che, grazie ad algoritmi di Machine Learning che apprendono dai dati storici, è in grado di discriminare quando un allarme deve essere gestito con urgenza da un tecnico sul territorio e quando invece si tratta di allarmi "deboli", che facilmente si risolveranno da soli. Un caso tipico si verifica durante temporali improvvisi con forti scariche elettriche, che creano instabilità nella rete di alimentazione, con temporanee interruzioni del funzionamento degli apparati che al termine del fenomeno riprendono la regolare attività. Ovviamente questi fenomeni possono avve-

Figura 1: Principali ambiti di applicazione di AI&ML per la rete TIM

Figura 2: ALADDIN visualizza anomalia di rete prima di un #timdown

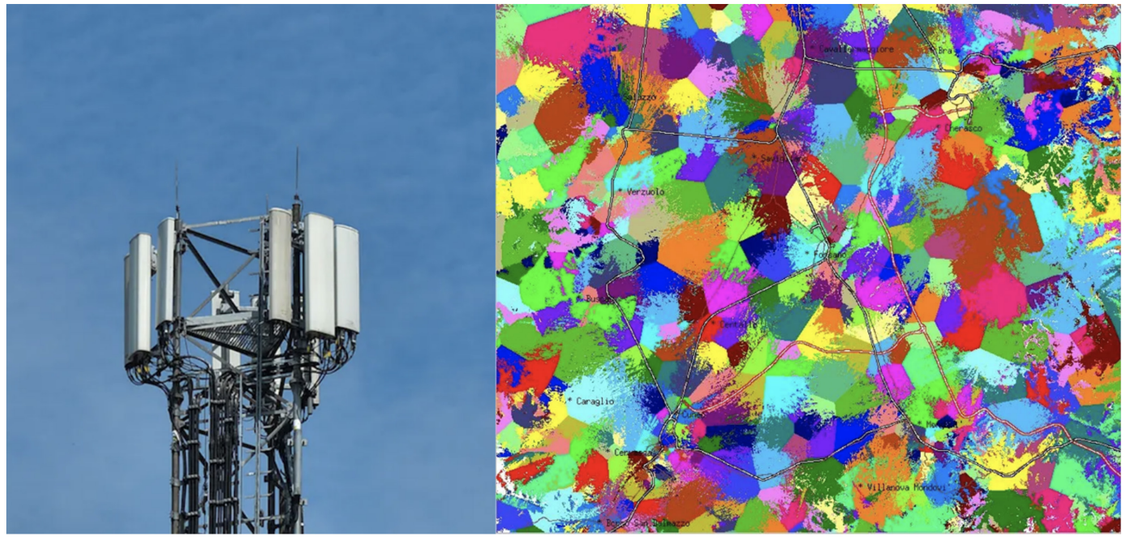


nire a centinaia di chilometri di distanza dai centri di supervisione e quindi senza che gli operatori ne abbiano evidenza diretta. I meccanismi di ML permettono di filtrare questi eventi, alleggerendo il carico degli operatori remoti ed evitando interventi non necessari in campo.
Per avere una idea dei numeri, nel 2022 il sistema ha evitato oltre 880.000 interventi di tecnici sulla rete di accesso fissa e mobile su un bacino di 5,7 milioni di allarmi analizzati. In questo modo, come è facile intuire, si abbassano i costi di gestione della rete e si migliora la qualità del servizio, intervenendo prontamente nella risoluzione dei problemi più importanti. Inoltre, l'algoritmo di PANAMA viene addestrato automaticamente ogni notte, per cui è stato ad esempio in grado di imparare a gestire gli allarmi della rete 5G senza nessuna specifica istruzione: questo dimostra ulteriormente la potenza e la versatilità di un approccio *data driven*.
Grazie al Machine Learning, inoltre, TIM ha realizzato il sistema **Aladdin** per l'identificazione delle anomalie di traffico sulla rete IP, basato su un *ensable* di algoritmi di Machine Learning tra cui un autoencoder convoluzionale.
Gli algoritmi implementati segnalano proattivamente quando il traffico si differenzia dalla normalità in modo significativo e permettono di allertare gli operatori di potenziali criticità, che potrebbero portare, se non affrontate in tempo, ad un grave disservizio su tutta la rete nazionale.

Ottimizzazione rete radiomobile

La gestione e l'ottimizzazione della rete radiomobile rappresenta un ambito di particolare interesse per l'utilizzo di algoritmi di AI, data la complessità e la specificità del dominio (propagazione, interferenza, incertezza della posizione e mobilità degli utenti, come descritto ad esempio in [2]). La realizzazione di applicazioni AI, ad esempio orientate alla realizzazione di use cases Self Organizing Network (SON), richiede quindi un approccio sistemico, in grado di abilitare un "accesso strutturato" a tale complessità [3].

Figura 3: Ottimizzazione delle aree di copertura cellulare mediante algoritmo ERIS



A questo scopo TIM, facendo leva sulle competenze di eccellenza nell'ambito della simulazione radio [4], ha sviluppato un framework software in grado di realizzare un vero e proprio *digital twin* della rete di accesso radio, inserendo in un unico modello sia simulazioni radiomobili ad elevata accuratezza, sia le migliaia di indicatori, anche geolocalizzati, e di parametri che possono essere ricavati dagli elementi di rete.
In questo contesto, è stato sperimentato in esercizio **ERIS** [5], un sistema che, basandosi su algoritmi di *Reinforcement Learning (RL)*, suggerisce la configurazione ottimale di tilt delle antenne, al fine di garantire la migliore copertura possibile senza dover ricorrere all'installazione di nuovi siti.
Questo approccio, molto potente in situazioni dove già si evidenza un potenziale miglioramento delle prestazioni della rete radiomobile, potrà essere utilizzato anche preventivamente andando a riconfigurare le antenne in base alle predizioni di carico della rete stessa. Grazie, infatti, alle potenzialità del deep learning e ad una base dati storica da cui poter apprendere, si sono sviluppati numerosi modelli in grado di predire, con un buon grado di accuratezza, i principali indicatori significativi per ogni cella, quali la quantità di utenti connessi e il traffico.
Queste predizioni possono essere poi usate in modo preventivo per riconfigurare la rete, sia per far fronte ad attesi carichi maggiori, ad esempio attraverso algoritmi di MLB (Mobility Load Balancing), sia per ottimizzarne l'utilizzo con interventi di energy saving come descritto in [6]. Inoltre, l'analisi statistica delle serie storiche degli indicatori disponibili fornisce utili informazioni sullo stato della rete e consente l'emissione di *early warning* di situazioni potenzialmente migliorabili, prima che si verifichino degradi sulla qualità del servizio percepita dagli utenti.

Data set analoghi, comprendenti anche indicatori georeferenziati derivati da misure MDT (Minimization of Drive Tests), sono utilizzati anche per attività innovative svolte in collaborazione con CNIT, sempre finalizzate alla realizzazione di algoritmi di ottimizzazione radio [7] ed all'estensione dei modelli predittivi anche ad altri indicatori, come la latenza, particolarmente rilevanti per la "Quality of Experience" dei servizi 5G [8].

Conclusioni

Nel contesto della rete TIM sono già presenti soluzioni mature basate su Machine Learning per l'ottimizzazione della gestione della rete di telecomunicazioni, sia fissa che mobile.

Oltre a quelli citati gli ambiti di applicazione possibile spaziano dalla pianificazione dell'agenda dei tecnici sul territorio, alla gestione degli investimenti, ad un supporto sempre più evoluto per le attività di gestione della rete.

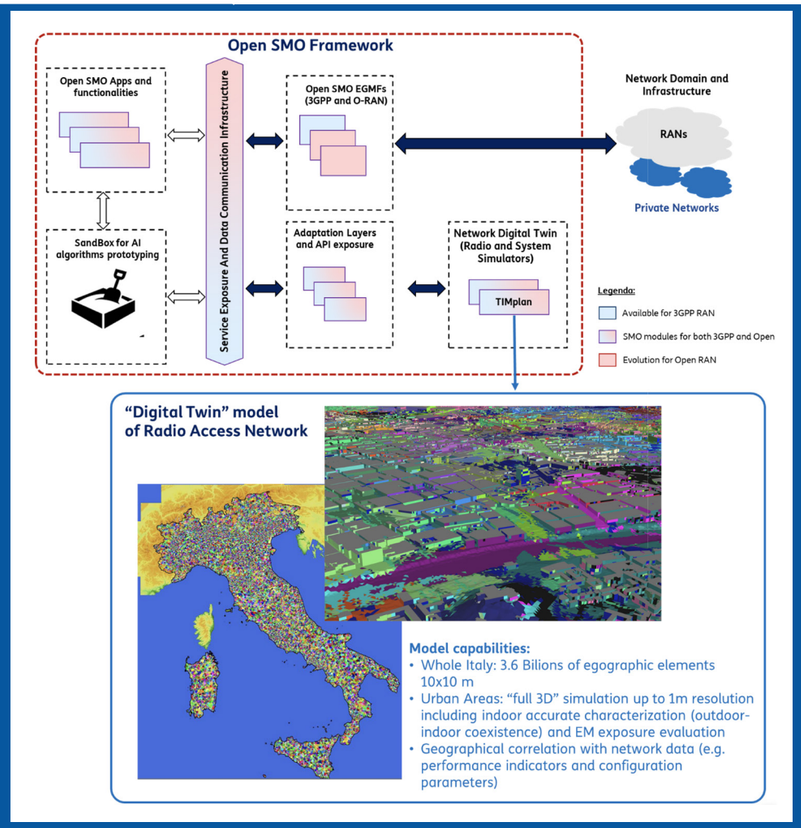
Un approccio multidisciplinare, il continuo aggiornamento delle competenze, la disponibilità di dati di qualità, di piattaforme dati e di strumenti informatici sono gli ingredienti fondamentali per realizzare soluzioni *data driven* di successo.■

“Digital Twin” della rete di accesso radio

TIM ha costruito un approccio che prevede la realizzazione di un vero e proprio modello “digital twin” (https://en.wikipedia.org/wiki/Digital_twin) della rete di accesso radio, una replica virtuale della rete stessa che permette di monitorare, simulare e governare il suo comportamento in real time, parallelamente al normale funzionamento della rete in campo. Ne è un esempio la piattaforma *TIMplan*, utilizzata per la progettazione radio dagli specialisti TIM, ma resa disponibile, mediante API e librerie di sviluppo, anche come base per l’addestramento di algoritmi come ERIS [4].

Il digital twin di rete sfrutta una piattaforma CI/CD di progettazione e successiva integrazione in rete attraverso l’approccio MLOps per l’automazione e il monitoraggio di tutte le fasi del ciclo di vita degli algoritmi di AI/ML (sviluppo, test, rilascio, deployment).

Figura A: Architettura dell’ambiente Open SMO con approccio Digital Twin



In quest’ottica rientra la realizzazione, all’interno del Framework Open SMO di gestione e orchestrazione della rete d’accesso radio di TIM, di una Sandbox offerta da TIMatom (la piattaforma cloud di sviluppo e integrazione sviluppata internamente) per consentire la prototipazione di algoritmi di ottimizzazione, sia internamente a TIM che verso terze parti, come ad es. le università. Si tratta di un ambiente JupiterLab (https://jupiter.org) di sviluppo interattivo che consente l’accesso supervisionato ai dati di rete mobile, opportunamente classificati sulla base della terminologia e delle definizioni degli standard 3GPP di riferimento per le differenti tecnologie radio. In particolare, sono resi disponibili, ad es. su base cluster, diverse fonti dati di rete d’accesso radio, sia interne che esterne alla piattaforma, che possono essere utilizzate per l’elaborazione e l’analisi algoritmica, tra cui:

- Dati e/o topologia di celle/siti con aggiornamento quotidiano;
- Copertura radioelettrica simulata, mediante il tool *TIMplan*;
- Configurazioni di cella/nodo giornaliere utilizzate per il Configuration Management (CM) per ciascun fornitore RAN;
- Contatori e KPI di Performance Management (PM) su base cella e cluster di celle e rop (15 min) per ciascun fornitore RAN;
- Misurazioni delle prestazioni e2e basate sulla raccolta dati da app installate su terminali mobili;
- Tracing MDT (Minimization of Drive Test) per raccogliere i dati di rete di accesso radio direttamente dagli UE sulla rete (geolocalizzati, quando disponibili) per diversi contatori/KPI e resi disponibili dai fornitori RAN.

Si noti che la soluzione combina, nell’approccio a digital twin, dati derivati dalla rete reale e dati simulati basati su modelli accurati della rete, in base all’architettura di massima dell’ambiente Open SMO rappresentata in Fig.A.

Lo sviluppo algoritmico in Sandbox all’interno di un contesto di digital twin produce un artefatto utilizzabile per passare dalla prototipazione, all’ingegnerizzazione fino alla messa in esercizio delle applicazioni di AI. Gli output della Sandbox sono moduli e/o relative librerie che possono essere condivise a livello programmatico (ad es. tramite API). La chiusura della catena di ottimizzazione in closed loop avviene tramite la gestione del ciclo di vita dell’artefatto tramite la catena CI/CD, al fine di consentirne il rilascio in esercizio. Sono quindi previste funzionalità di monitoring per la raccolta dei feedback dalla rete che saranno utilizzati dalla catena MLOps per un tuning continuo degli algoritmi.

Le caratteristiche di apertura e flessibilità di tale approccio rappresentano un abilitatore, potenzialmente a livello nazionale, per un nuovo modo di realizzare ed ottimizzare le reti 4G/5G e futuro 6G, mettendo a fattor comune le competenze interne, le collaborazioni con le università e i partner industriali favorendo lo sviluppo di un ecosistema attraverso opportuni progetti di ricerca, ad es. quelli relativi al programma Restart (https://www.fondazione-restart.it/it/home-italiano/) sulle Reti Programmabili (Progetto Super - https://www.fondazione-restart.it/it/progetti/s2-super/) e sulle Industrial Networks (Progetto IN - https://www.fondazione-restart.it/it/progetti/s9-in/). L’approccio descritto è pienamente coerente con il lavoro che, anche con il contributo di TIM, è portato avanti negli enti di standardizzazione tecnica affinché l’AI possa essere nativamente contemplata nelle architetture e nelle funzionalità delle nuove release di rete 5G e nell’evoluzione verso il 6G.

danilo.dolfini@telecomitalia.it
francesco.epifani@telecomitalia.it
giuseppe.minerva@telecomitalia.it
simone.piacco@telecomitalia.it

ML e InfoVis a supporto della manutenzione proattiva dei cabinet stradali

Con oltre 9ML di linee, la rete FTTC (Fiber To The Cabinet) costituisce un segmento estremamente rilevante nel dominio dell'accesso fisso di TIM ed è destinata a rimanere tale nei prossimi anni. Infatti, se da un lato c'è una forte spinta orientata alla migrazione verso FTTH (Fiber To Te Home), dall'altro – specialmente in aree rurali – molte delle vecchie linee in rame vengono progressivamente sostituite con collegamenti FTTC
Uno dei nodi fondamentale della rete FTTC, è chiamato Optical Network Unit (di seguito "ONU"). Questi apparati sono contenuti in opportuni **cabinet** co-locati con gli armadi stradali e costituiscono il punto di interconnessione tra le linee in rame della rete secondaria (provenienti dalle abitazioni degli utenti) e il collegamento in fibra primario verso la centrale più vicina.
Essendo dislocate in strada, le ONU presentano problematiche particolari rispetto agli altri apparati di rete tipicamente posizionati in centrale. Esse sono particolarmente sensibili alle condizioni meteo avverse e a problematiche relative all'alimentazione, a causa della composizione della stazione d'energia ad essi dedicata.

Il progetto **ONUCab** (dall'unione di ONU e Cabinet) utilizza tecniche di machine learning e advanced information visualization per supportare una manutenzione intelligente delle ONU e dei cabinet stradali, mirata a ridurre i malfunzionamenti e, al contempo, ottimizzare l'operatività riducendo quindi i costi di gestione.
Le principali attività del progetto al momento hanno indirizzato 2 problematiche estremamente rilevanti: il fenomeno degli spegnimenti e il proliferare delle porte "dichiarate guaste".
Lo spegnimento di una ONU causa chiaramente un disservizio immediato su tutti i clienti (circa 100 in media) ad essa attestati e in un anno si registrano circa 900.000 spegnimenti sul totale dei 135.000 ONU di TIM. Gli spegnimenti possono essere divisi in 2 macrocategorie: quelli dovuti a una interruzione nella fornitura di energia da parte del gestore e quelli dovuti a problematiche dell'apparato o della sua stazione di energia, come ad esempio un blocco alimentazione dovuto a surriscaldamento. Poter distinguere tra queste due categorie è fondamentale perché ci permette di differenziare le azioni che ne conseguono. Infatti, se nel primo caso non è utile alcun intervento in loco ma è importante dare evidenza agli utenti dell'origine del disservizio, nel secondo occorre analizzare le cause alla base dello spegnimento ed intervenire in modo mirato così da evitare il ripetersi del problema. Sfortunatamente non c'è modo di effettuare questa distinzione in maniera certa, ma con opportune tecniche di aggregazione spaziale e temporale si riesce a categorizzare gli spegnimenti ed indirizzare di conseguenza le azioni. I cluster individuati sono 3: i "*grappoli*", 4 o più apparati vicini che si spengono simultaneamente e che, quindi, quasi certamente non necessitano di intervento poichè legati a mancanza di corrente su una cabina di media tensione del gestore; i "*singoli*", spegnimenti isolati molto probabilmente dovuti a problematiche dell'apparato che necessitano di intervento; i "*multipli*" ovvero gruppi di solo 2/3 spegnimenti simultanei per i quali quindi la causa è incerta.

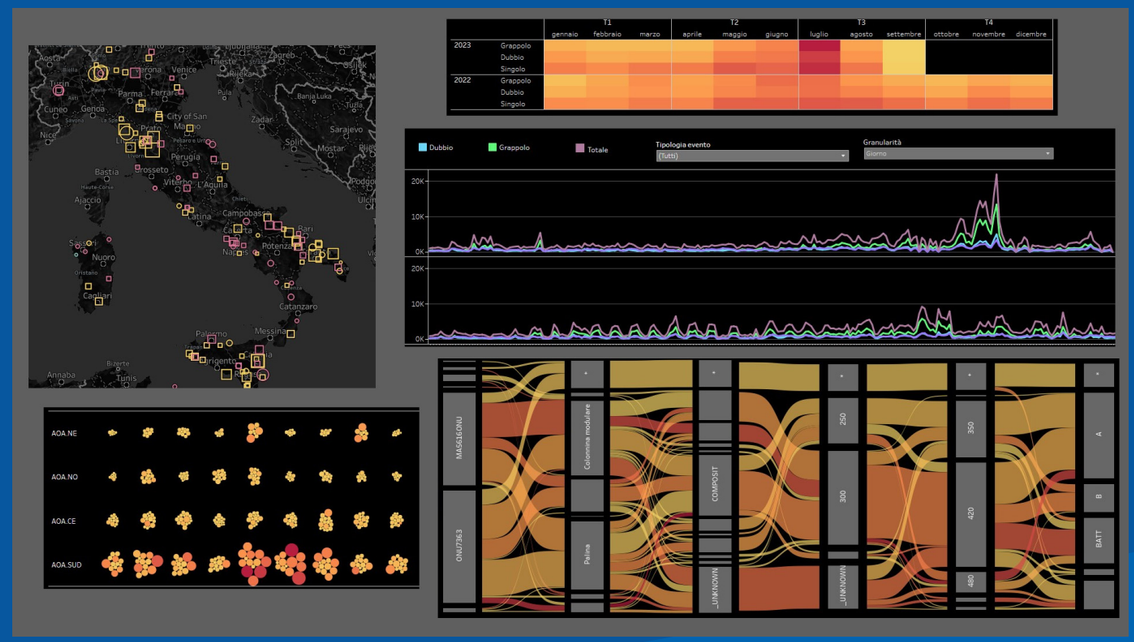
I principali output sono stati la creazione di strumenti di analisi visuale per esplorare i dati in modo intuitivo ed interattivo che hanno consentito di evidenziare la portata del fenomeno degli spegnimenti a "grappolo" e la definizione delle logiche di aggregazione per l'identificazione dei grappoli stessi le quali, implementate sul sistema PANAMA per la prioritizzazione degli allarmi descritto nell'articolo principale, hanno consentito di risparmiare circa 10.000 interventi a vuoto dei tecnici on-field all'anno.

Le porte degli ONU sulle quali si attestano le linee dei clienti, come qualunque componente meccanico, possono essere soggette a guasti. In tal caso la linea deve essere spostata su un'altra porta mentre quella vecchia viene marcata come "guasta" (e quindi non più commercializzabile) negli inventory di TIM. Analizzando i dati storici si nota però una significativa discrepanza tra il numero delle porte marcate "guaste" e i reali tassi di guastabilità, in particolar modo su certe tipologie di schede. Sulla base di questi tassi, del tipo di scheda, delle caratteristiche del territorio e della distribuzione dei guasti nel tempo e sull'apparato è stato quindi studiato un algoritmo che calcola la probabilità che una porta oggetto di Cambio-Porta sia in realtà funzionante. Questo algoritmo è stato integrato nel processo di Cambio-Porta in modo tale che le porte con una elevata probabilità di essere funzionanti non vengano marcate come "guaste" e rimangano disponibili per la commercializzazione. Con un tasso di successo di circa il 70%, questa funzionalità ha consentito ad oggi di recuperare più di 300.000 porte per un risparmio hardware corrispondente a quasi 8ML di euro annui.

giovanni.caire@telecomitalia.it
mario.mirabelli@telecomitalia.it
simone.pescetelli@telecomitalia.it

Figura A: Cabinet ONU e stazione di energia dedicata

Figura B: Strumenti di analisi visuale per l'analisi dei "grappoli" di spegnimenti



Bibliografia

1. L'Intelligenza Artificiale per il saving energetico delle centrali https://www.gruppotim.it/it/newsroom/notiziario-tecnico-tim/Anno-2023/n2-2023/Intelligenza_Artificiale_per_il_saving_energetico_delle_centrali.html
2. DIGIRAN: IL VALORE DELL'AUTOMAZIONE NELL'ACCESSO RADIO https://www.gruppotim.it/content/dam/telecomitalia/it/archivio/documenti/Innovazione/MnisitoNotiziario/2018/1-2018/capitolo6/capitolo%2006.pdf
3. Open Service Management & Orchestration: un nuovo paradigma per la gestione automatizzata delle reti mobili https://www.gruppotim.it/it/newsroom/notiziario-tecnico-tim/Anno-2023/n1-2023/Open-Service-Management-Orchestration.html
4. MANAGING COMPLEXITY: AUGMENTED INTELLIGENCE FOR 5G RADIO ACCESS DESIGN AND OPTIMIZATION https://www.gruppotim.it/content/dam/telecomitalia/it/archivio/documenti/Innovazione/MnisitoNotiziario/2019/2-2019/capitolo5/cap05_augmented_intelligence_5G.pdf
5. Algoritmo ERIS per l'ottimizzazione della copertura e della capacità di rete con un approccio basato su Reinforcement Learning https://www.gruppotim.it/it/newsroom/notiziario-tecnico-tim/Anno-2023/n1-2023/Open-Service-Management-Orchestration/ERIS-algoritmo-ERIS-per-ottimizzazione-della-copertura-e-della-capacita-di-rete.html
6. L'Intelligenza Artificiale per l'ottimizzazione energetica della rete di accesso radiomobile https://www.gruppotim.it/it/newsroom/notiziario-tecnico-tim/Anno-2023/n2-2023/Intelligenza_Artificiale_per_il_saving_energetico_delle_centrali/Intelligenza_Artificiale_per_ottimizzazione_energetica_della_rete_accesso_radiomobile.html
7. Cellular Network Capacity and Coverage Enhancement with MDT Data and Deep Reinforcement Learning. «COMPUTER COMMUNICATIONS», 2022, 195
8. Data-driven Predictive Latency for 5G: A Theoretical and Experimental Analysis Using Network Measurements, presented at PIMRC'23

Urlografia

- [a] https://chat.openai.com/
- [b] https://www.midjourney.com/

Acronimi

AI	Artificial Intelligence
AIxIA	Associazione Italiana di Intelligenza Artificiale
API	Application Programming Interface
CCO	Coverage and Capacity Optimization
CI/CD	Continuous Integration/Continuous Deployment
CM	Configuration Management
CNIT	Consorzio Nazionale Interuniversitario per le Telecomunicazioni
EdgeML	Machine Learning at the Edge
ERIS	Enhanced Reinforcement learning for Innovating Self organizing networks
FTTH	Fiber To The Home
Gen AI	Generative Artificial Intelligence
IP	Internet Protocol
IT	Information Technology
KPI	Key Performance Indicators
LLM	Large Language Models
MDT	Minimization of Drive Test
ML	Machine Learning
MLB	Mobility Load Balancing
MLOps	Machine Learning Operations
PANAMA	Predictive Algorithms for Network Alarms MAnagement
PM	Performance Management
QCI	QoS Class Identifier
RAN	Radio Access Network
RESTART	RESearch and innovation on future Telecommunications systems and network, to make Italy more smart
RL	Reinforcement Learning
SMO	Service Management and Orchestration
SON	Self Organizing Network

Autori

Aurelio Giammusso *aureliomariaa.giammusso@telecomitalia.it*

Ingegnere elettronico, dopo una breve esperienza nel campo dell'elaborazione delle immagini nel settore medicale, entra in Azienda nel 1988. Ha sempre operato nel settore tecnico sia in ambito in esercizio della rete, sia in ambito progettazione e sviluppo. Nel 2015 è stato parte attiva del progetto di creazione del Front End unico di supervisione della rete e da allora coordina le iniziative di sviluppo dei sistemi in ottica di automatismo ed efficienza dei processi lavorativi con ricorso a tecniche di AI e ML. ■

Roberta Giannantonio *roberta.giannantonio@telecomitalia.it*

Ingegnere delle telecomunicazioni, entra in Azienda nel 2004 per occuparsi dei progetti di innovazione con tecnologie wireless. Dal 2016 si occupa di progetti di Intelligenza Artificiale a supporto dell'operatività di TIM in contesti come l'assurance, la pianificazione di rete e l'ottimizzazione dei consumi energetici e da fine 2022 è responsabile della funzione Data Network Learning di TIM. ■

Michele Ludovico *michele.ludovico@telecomitalia.it*

Ingegnere elettronico, ha iniziato ad operare nel gruppo TIM progettando sistemi a micro-onde per comunicazioni via satellite. Dal 2001 si occupa di strumenti e metodologie di progettazione ed ottimizzazione dell'accesso radio, che TIM sviluppa "in house" a supporto dell'evoluzione della rete mobile. Dal 2014 è responsabile della funzione di TIM che assicura l'ingegneria e lo sviluppo delle soluzioni di automazione per la rete di accesso radio, secondo il paradigma "Self Organizing Network". Ha svolto, inoltre, attività di formazione e consulenza in Italia ed all'estero ed è co-inventore di diversi brevetti nel campo della progettazione wireless e della gestione delle risorse radio. ■

Andrea Marafante *andrea.marafante@telecomitalia.it*

Ingegnere elettronico, entra in Azienda nel 1996 iniziando da progetti di ingegneria inerenti l'innovazione della rete di accesso per servizi xDSL e Larga Banda. Numerose esperienze nella ingegnerizzazione dei processi operativi in ambito Open Access con diverse responsabilità in ambito Field Management, Delivery, Maintenance ed attualmente Assurance. Ha seguito, come riferimento progettuale, le iniziative di trasformazione legate alla revisione della catena di Delivery per l'Equivalence ed al nuovo WFM dei tecnici on field. Da febbraio 2023 è responsabile in ambito Assurance&Maintenance della funzione Process Digital Engineering. ■

Verso una Greener AI

Gabriele Elia

I sistemi di Intelligenza Artificiale (AI) sono estremamente "energivori", cioè consumano grandi quantità di energia sia nella fase in cui sono sviluppati sia nel loro utilizzo, perché usano grandi quantità di server e sistemi.
Questo li rende differenti dalla maggioranza dei sistemi IT classici.

Ci sono molti esempi che si possono fare a questo riguardo anche con notizie curiose e paradossali: per esempio, uno studio di qualche tempo fa dell'Università della California[1], Riverside, e dell'Università del Texas, Arlington, ripreso anche dalla stampa generalista, avvisava che ChatGPT utilizza l'equivalente di una bottiglietta di acqua di mezzo litro[2] per il raffreddamento dei server ogni 20-50 domande a cui risponde.

Secondo Gartner[3], p.es. "l'Intelligenza Artificiale consumerà più energia della forza lavoro umana" entro il 2025, a meno che non vengano compiuti passi significativi in termini di efficienza."
Un altro studio recente stima che Google, se passasse tutti i suoi sistemi di ricerca su AI, consumerebbe e più energia elettrica dell'Irlanda[4]. Google attualmente consuma oltre 18 TWh di elettricità all'anno, quasi 10 volte l'energia necessaria a un operatore di rete come TIM.
Uno studio di SemiAnalysis[5] rivela che ChatGPT si basa su un numero impressionante di quasi 29.000 GPU NVIDIA per fornire risposte e ha un costo operativo giornaliero superiore a 694.000 dollari.
Ed ancora Microsoft[6] sta assumendo esperti in costruzione di centrali nucleari!! Si ipotizza voglia costruirne una privata per alimentare alcuni dei propri data center.

Il picco dell'AI

Secondo uno studio della International Energy Agency[7] pubblicato nel 2022 sulla rivista Nature Sustainability, si prevede che il consumo di energia per la formazione e l'implementazione dell'IA aumenterà da 10 a 50 volte entro il 2030.

Note

(1) https://www.businessinsider.com/chatgpt-generative-ai-water-use-environmental-impact-study-2023-4. Questo perché ChatGPT è un modello linguistico di grandi dimensioni e richiede molta potenza di calcolo per essere eseguito. Questa potenza di calcolo genera calore, che deve essere raffreddato utilizzando acqua. Sebbene possa sembrare una piccola quantità di acqua per interazione, l'effetto cumulativo di milioni di persone che utilizzano ChatGPT per varie query può essere significativo. Inoltre, è probabile che l'impronta idrica di ChatGPT aumenti man mano che il modello diventa più grande e più sofisticato

(2) A titolo di confronto: la produzione di un kg di carta consuma in media in Italia 26 litri di acqua. Produrre un foglio A4 richiede quindi circa 0,1 litri di acqua! Fonte: https://www.assocarta.it/it/pubblicazioni.html e https://www.assocarta.it/it/documenti/category/6-pubblicazioni.html?download=404:rapporto-ambientale-2022 pag. 18

(3) https://www.gartner.com/en/articles/keep-ai-from-doing-more-climate-harm-than-good lo studio confronta i consumi dei sistemi AI con i "consumi" di 2000 Kcal giornalieri per i 7 miliardi di essere umani al mondo

(4) https://thenextweb.com/news/googles-ai-could-consume-as-much-electricity-as-ireland
L'articolo è stato pubblicato da Alex de Vries presso la VU Amsterdam School of Business and Economics.
Nel 2021, il consumo totale di elettricità di Google è stato di 18,3 TWh, di cui l'Intelligenza Artificiale rappresenta il 10%-15%. Se Google Search fosse sostituito da Google Bard si utilizzerebbero 29,3 terawattora all'anno, equivalenti oltre 29 TWh all'anno cioè al consumo di elettricità dell'Irlanda

(5) https://www.semianalysis.com/p/the-inference-cost-of-search-disruption

(6) https://www.theverge.com/2023/9/26/23889956/microsoft-next-generation-nuclear-energy-smr-job-hiring

(7) https://www.researchgate.net/publication/352384074_Artificial_intelligence_on_economic_evaluation_of_energy_efficiency_and_renewable_energy_technologies

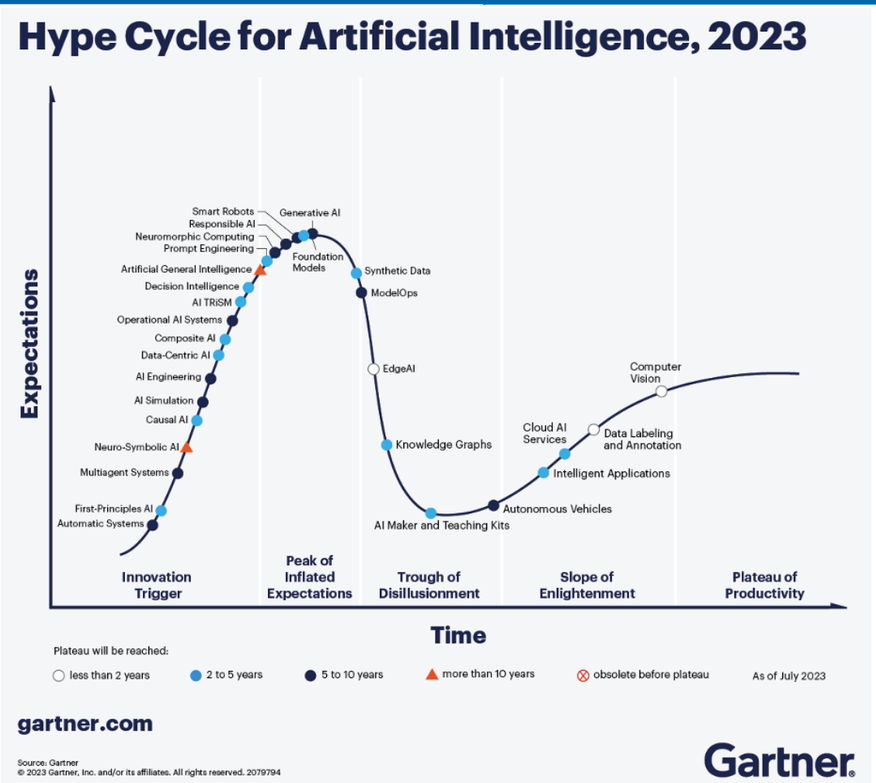
Ciò è dovuto a una serie di fattori, tra cui:

- il crescente utilizzo dell'AI in un'ampia gamma di applicazioni;
- lo sviluppo di modelli di Intelligenza Artificiale più complessi e potenti;
- la crescente quantità di dati che devono essere elaborati per addestrare e implementare modelli di Intelligenza Artificiale.

Lo studio ha inoltre rilevato che il consumo energetico dell'AI è concentrato in un numero limitato di paesi, tra cui Stati Uniti, Cina ed Europa. Questi paesi ospitano i più grandi data center cloud e aziende di Intelligenza Artificiale.

Siamo infatti al picco dell'Hype relativamente AI Generativa[8] estremamente energivoro. Una delle caratteristiche dei nuovi sistemi AI è che mentre in passato, diversi scenari applicativi richiedevano lo sviluppo di modelli diversi, gli attuali modelli, di grandi dimensioni, assorbendo enormi quantità di conoscenza, possono adattarsi a molteplici scenari di business, abbassando significativamente la soglia per lo sviluppo e l'applicazione dell'IA ed accorciando il ciclo dalla tecnologia all'applicazione".

Figura 1: Hype Cycle Garner AI, 2023 (fonte: Gartner)

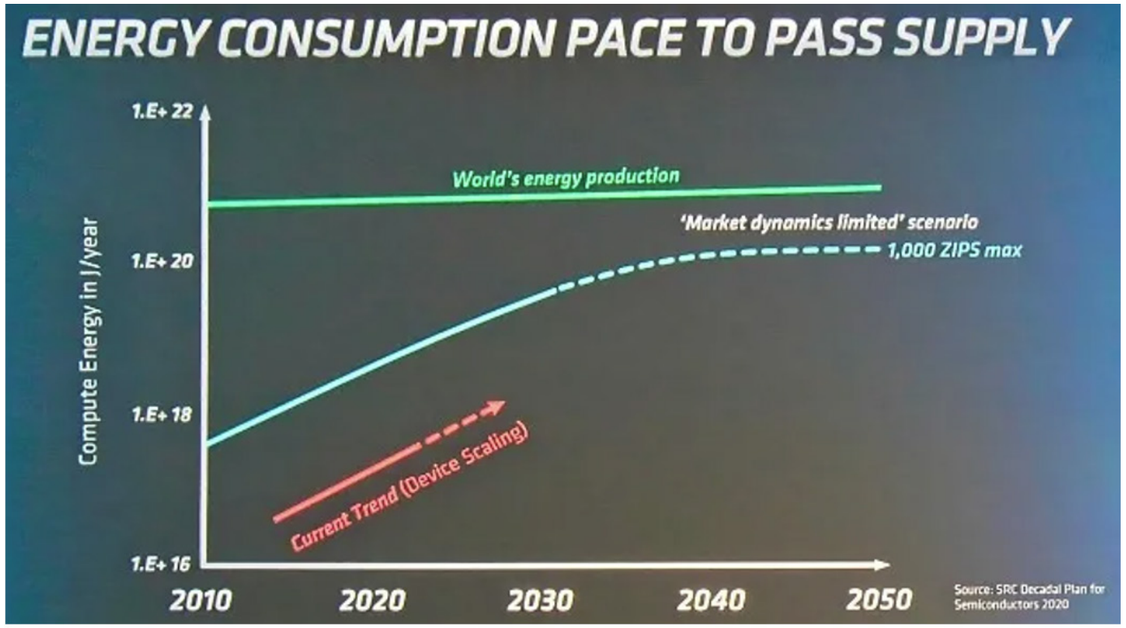


Note

(8) https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle

Ad esempio, uno studio ha scoperto che l'addestramento di un grande modello linguistico chiamato GPT-3 ha richiesto 626.000 tonnellate di emissioni di CO2, che equivalgono alle emissioni annuali di 500 auto americane[9].
La Fig.2 indica il paradosso di questa crescita, insostenibile.
Come misura dello sviluppo impetuoso dell'AI, possiamo guardare agli investimenti degli ultimi anni: ebbene gli investimenti di Venture Capital nelle società di Intelligenza Artificiale hanno subito un'accelerazione negli ultimi dieci anni. Secondo i dati Crunchbase[11], tra gennaio 2013 e il terzo trimestre del 2023 sono stati investiti più di 300 miliardi di dollari in finanziamenti di venture capital in oltre 16.000 aziende del settore. A partire dal terzo trimestre del 2023, circa 1 su 4 dollari di rischio negli Stati Uniti quest'anno è andato a una startup che incorpora l'Intelligenza Artificiale nella sua attività.
A partire dalla fine del 2022 con l'apertura al grande pubblico di ChatGPT, l'attenzione ai sistemi di Intelligenza Artificiale (AI) si è particolarmente rivolta ai sistemi di cosiddetta Generative AI.
Questi sistemi in grado di "generare" risposte verosimili a domande poste

Figura 2: Energy sistemi di AI se la crescita fosse non controllata – (fonte : AMD[10])



Note

(9) hbr.org/2023/07/how-to-make-generative-ai-greener
(10) AI Power Consumption Exploding (semiengineering.com)
(11) The Rise of Generative AI - Tech Scaleup Silicon Valley (mindthebridge.com)

in linguaggio naturale, tradurre testi, generare immagini, musica, video o altri dati "artificiali" con un minimo intervento umano, tipicamente partendo da input fatti di descrizioni testuali in linguaggio naturale. Con innumerevoli campi di applicazione, le tecnologie di Generative AI usano modelli di "Machine Learning" per trovare correlazioni statistiche tra dati di addestramento (training data) e gli output che sono in grado di generare.
Questo perché molte applicazioni di Intelligenza Artificiale generativa sono costruite su modelli di base chiamati "foundation models". Si tratta di algoritmi di apprendimento automatico che sono stati pre-addestrati su enormi set di dati e che sono adattabili a dati "adiacenti".
In linea di massima, la fase di addestramento di questi modelli è molto onerosa se pensiamo alle risorse di calcolo di cui ha bisogno, ma non è estremamente complessa algoritmicamente. Anche la fase di "generazione" è molto esigente come risorse di calcolo.
Un modello generativo usato come chatbot o motore di ricerca è per esempio molto più impegnativo come risorse informatiche rispetto ad una query in un motore di ricerca tra documenti preesistenti come Google.

Secondo un rapporto del 2022 dell'Agenzia internazionale per l'energia, l'AI è attualmente responsabile di circa l'1% del consumo globale di elettricità[12]. Tuttavia, il rapporto prevede anche che il consumo di elettricità dell'IA potrebbe aumentare fino a 50-100 volte entro il 2030.

Cosa succederebbe se Google passasse ad AI?

Un'azienda specializzata, SemiAnalysys[13], ha provato a stimare cosa succederebbe al conto economico di Google se il modello ChatGPT venisse intromesso nelle attività di ricerca esistenti di Google.
Ne risulta che l'impatto sarebbe devastante. Ci sarebbe una riduzione di 36 miliardi di dollari nel reddito operativo, dovuti ai costi di inferenza di sistemi basati su Large Language Model (LLM) come appunto Google Bard o OpenAi ChatGPT.
Secondo lo studio, l'implementazione dell'attuale ChatGPT in ogni ricerca effettuata da Google richiederebbe circa mezzo milione di server con un totale di oltre 4 milioni di GPU A100. Il costo totale di questi server e reti supera i 100 miliardi di dollari solo in Capex.

Strategie per ridurre l'impatto ambientale dell'AI

Sono dunque necessarie strategie e azioni per ridurre l'impatto ambientale dei prossimi sistemi AI.

Le azioni che si stanno intraprendendo sono su tre linee di lavoro:
- utilizzare hardware e software più efficienti dal punto di vista energetico, (p.es. utilizzare hardware per ridurre il consumo energetico dei modelli di Intelligenza Artificiale);
- utilizzare in maniera più mirata i modelli AI disponibili;
- addestrare modelli di IA generativa su set di dati più piccoli e utilizza algoritmi di training più efficienti.

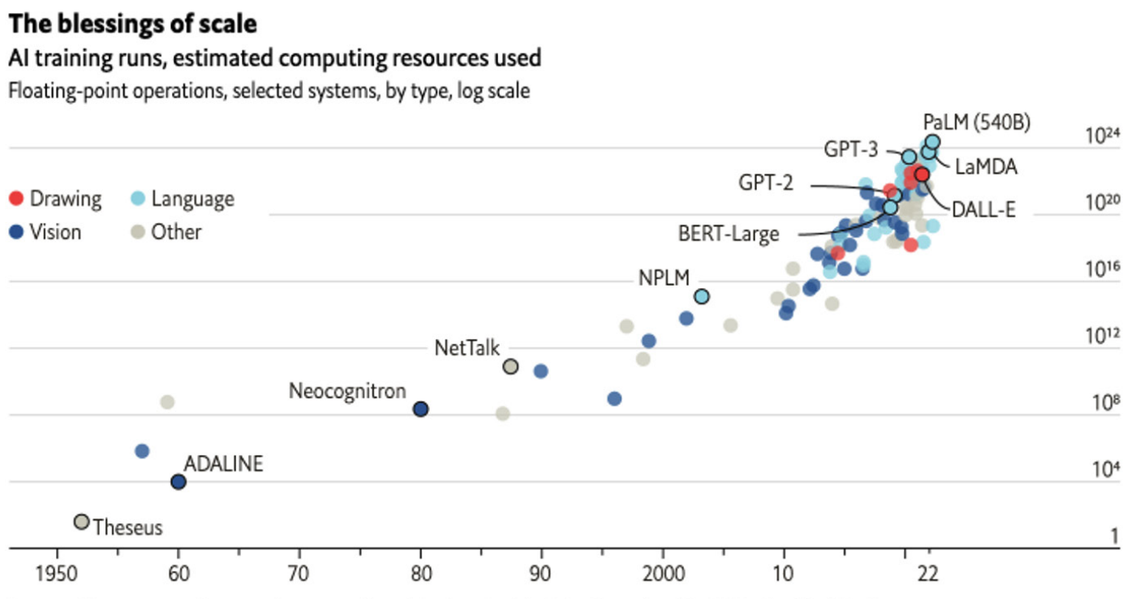
Nel resto dell'articolo sintetizzeremo alcune di queste opportunità, in particolare l'evoluzione di alcuni chip specializzati e lo sviluppo di nuovi modelli AI meno energivori.

Nuovi CPU e Semiconduttori energeticamente efficienti per AI

Come illustra la Fig.3[14], il numero di parametri (di conseguenza la larghezza e la profondità) delle reti neurali e quindi la dimensione del modello sta aumentando. Per creare modelli di deep learning migliori e potenziare le applicazioni di Intelligenza Artificiale generativa, le organizzazioni necessitano di maggiore potenza di calcolo e larghezza di banda di memoria.

Anche la Fig.4 mostra come i chip generici (come le CPU) non possono supportare modelli di deep learning alta-

Figura 3: Crescita esponenziale della complessità dei modelli di training AI



Note

(12) Rapporto VIA "Intelligenza Artificiale per l'efficienza energetica e le energie rinnovabili - 6 applicazioni attuali", 2021 https://www.researchgate.net/publication/352384074_Artificial_intelligence_on_economic_evaluation_of_energy_efficiency_and_renewable_energy_technologies

(13) Peeling The Onion's Layers - Large Language Models Search Architecture And Cost (semianalysis.com)

Note

(14) https://research.aimultiple.com/ai-chip-makers/

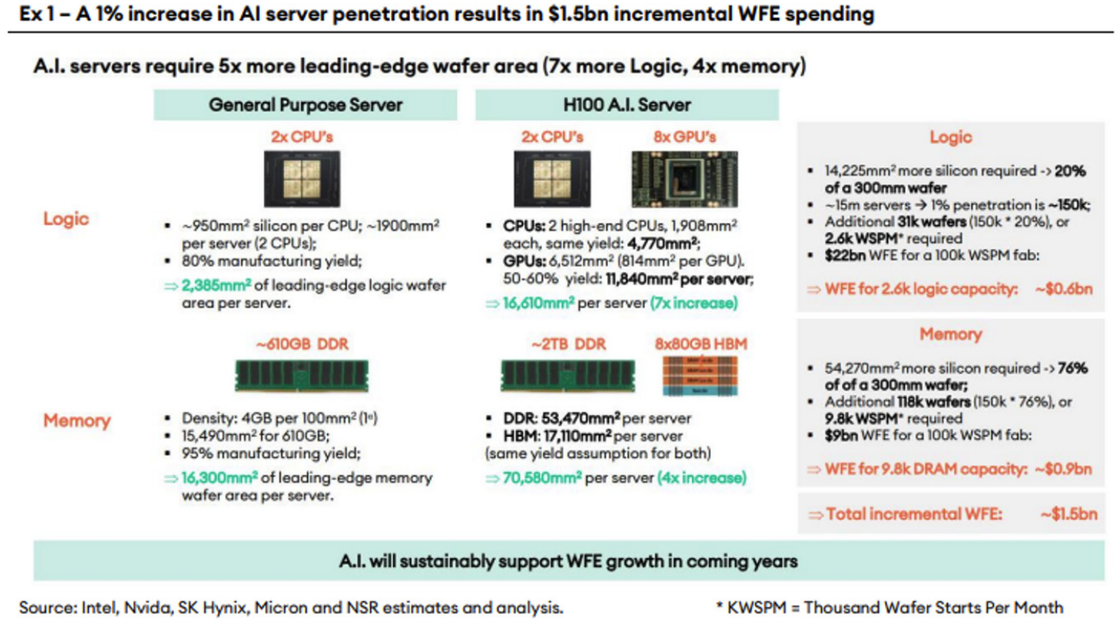
mente parallelizzati. Pertanto, i chip AI che consentono capacità di elaborazione parallela sono sempre più necessari. Questo processo migliora anche l'efficienza energetica.
Nuove CPU e microchip possono contribuire a rendere l'AI generativa più efficiente dal punto di vista energetico.
L'hardware più recente è in genere più efficiente dell'hardware precedente, il che significa che può eseguire le stesse attività consumando meno energia.
Oltre alle nuove CPU e microchip, ci sono una serie di altre innovazioni hardware che potrebbero contribuire a rendere l'AI generativa più efficiente dal punto di vista energetico. Ad esempio, i ricercatori stanno sviluppando nuovi tipi di memoria che sono più efficienti dal punto di vista energetico rispetto alla DRAM tradizionale.
Di seguito elenchiamo alcuni di questi progressi.

Nvidia H100 e A100: i principali chip per AI

Nvidia è un leader mondiale nella produzione di chip per l'Intelligenza Artificiale (AI), le sue "GPU" sono considerate il benchmark in molti contesti.
I suoi chip sono utilizzati in una vasta gamma di applicazioni, tra cui la visione artificiale, il riconoscimento del linguaggio naturale e l'apprendimento automatico.
I principali chip Nvidia per AI sono l'H100 e l'A100.

Un tema di sicurezza nazionale

Figura 4: Impatto della diffusione dell'AI nella industria dei Chip secondo NSR[15]



Note

(15) https://www.newstreetresearch.com/research/openai-drama-our-read-and-implications/attachment

I chip Nvidia sono così potenti che la loro esportazione dagli USA è soggetta a restrizioni. In particolare, il 23 ottobre 2023, il Dipartimento del Commercio degli Stati Uniti ha imposto nuove restrizioni all'esportazione di chip e tecnologie avanzate verso la Cina.
Queste restrizioni includono l'H100 e l'A100 di Nvidia, che sono i chip AI più avanzati dell'azienda. Le nuove restrizioni impongono a Nvidia di richiedere una licenza al Dipartimento del Commercio per esportare i chip H100 e A100 in Cina. Il Dipartimento del Commercio esaminerà ogni richiesta di licenza caso per caso e può negare la licenza se ritiene che l'esportazione del chip possa rappresentare una minaccia per la sicurezza nazionale degli Stati Uniti.

Intel

Intel è il più importante player nel mercato dei chip AI. L'azienda offre una varietà di chip AI, sia per l'uso nei data center che nei dispositivi mobili.
I principali chip Intel per AI sono i seguenti: Gaudi2 e Ponte Vecchio.
Ponte Vecchio ha una potenza di calcolo FP32 di 52 teraFLOPS, mentre l'H100 ha una potenza di calcolo FP32 di 60 teraFLOPS. Ciò significa che l'H100 ha una potenza di calcolo FP32 del 16% superiore a quella del Ponte Vecchio.
Il seguente grafico confronta il consumo energetico del Ponte Vecchio con quello di alcuni altri chip AI di fascia alta:

Chip	Consumo energetico (watt)
Ponte Vecchio	500
H100	450
A100	350

Come si può vedere dalla tabella, il Ponte Vecchio ha un consumo energetico superiore a quello di altri chip AI di fascia alta. Ciò è dovuto alla sua maggiore potenza di calcolo.

Qualcomm

Qualcomm, un altro gigante dei semiconduttori, dispone di chip proprietari per AI per data center. I primi chip AI di Qualcomm, datati 2022, sono denominati Cloud AI 100.
Qualcomm ha annunciato che avrebbe lanciato un secondo chip AI per data center, chiamato Cloud AI 200 ed un altro per il 2024, chiamato Cloud AI 300, con prestazioni maggiori.
Qualcomm ha dichiarato che il Cloud AI 200 ha un consumo energetico di 300 watt per chip. Questo consumo energetico è inferiore rispetto a quello di altri chip AI per data center, come le GPU di Nvidia.

I nuovi chip Microsoft per AI

E' interessante notate che i principali attori cloud sviluppino chip "in house" oltre che utilizzare chip dei principali produttori di semiconduttori come ARM, Nvidia e Intel e altri.

Microsoft ha presentato nel novembre 2023 i suoi nuovi chip per l'Intelligenza Artificiale, denominati Maia 100. Questi chip sono progettati per migliorare le prestazioni e l'efficienza energetica dei carichi di lavoro AI, come il machine learning e il deep learning.
Microsoft ha affermato che i chip Maia saranno utilizzati nei suoi data center per migliorare le prestazioni di una serie di servizi cloud, tra cui Azure Machine Learning, Azure Cognitive Services e Azure Bot Service.

Come molti altri, questi chip utilizzano per efficienza energetica un sistema di raffreddamento liquido, che aiuta a mantenere il chip a una temperatura operativa ottimale.

Google
Google ha sviluppato una serie di chip proprietari per cloud e AI, tra cui: Tensor Processing Unit (TPU); Edge TPU e Bard; anch'essi utilizzano un sistema di raffreddamento liquido.
In particolare, Google ha annunciato il lancio di una nuova generazione di TPU, denominata TPUv5. Le TPUv5 offrono prestazioni ed efficienza energetica significativamente superiori rispetto alle generazioni precedenti. Google prevede di utilizzare le TPUv5 nei suoi data center per alimentare applicazioni AI sempre più complesse.

Amazon
Amazon ha due chip proprietari per l'Intelligenza Artificiale nel cloud: Trainium ed Inferentia. Amazon ha iniziato a utilizzare questi chip nei suoi servizi cloud nel 2022.

I chip Trainium vengono utilizzati per alimentare applicazioni basate sull'Intelligenza Artificiale, come il riconoscimento vocale e la visione artificiale. I chip Inferentia vengono utilizzati per alimentare applicazioni basate sull'Intelligenza Artificiale, come la traduzione automatica e il riconoscimento delle immagini.
Amazon prevede di rendere questi chip disponibili anche ai clienti terzi in futuro.
Secondo Amazon, il chip Trainium offre un'efficienza energetica fino a 10 volte superiore rispetto a una GPU tradizionale per l'addestramento dei modelli di AI.
Questo significa che il chip Trainium può eseguire lo stesso lavoro con un consumo energetico inferiore, il che può portare a risparmi significativi sui costi operativi.

Il chip Inferentia offre un'efficienza energetica fino a 4 volte superiore rispetto a una GPU tradizionale per l'inferenza dei modelli di AI.
Questo significa che il chip Inferentia può eseguire lo stesso lavoro con un consumo energetico inferiore, il che può portare a miglioramenti delle prestazioni e dell'economicità delle applicazioni AI.
Ad esempio, Amazon afferma che il chip Trainium può ridurre il consumo energetico del 70% per l'addestramento di un modello di Intelligenza Artificiale per la visione artificiale.
Questo può comportare un risparmio fino a 1 milione di dollari all'anno per un'azienda che utilizza il chip per addestrare i propri modelli AI.

Meta
A dicembre 2023, Meta non dispone ancora di chip proprietari per AI. Tuttavia, l'azienda ha annunciato di essere al lavoro sullo sviluppo del suo primo chip proprietario, chiamato MTIA (Meta Training and Inference Accelerator), che dovrebbe essere disponibile nel 2025.
In attesa dell'uscita del MTIA, Meta utilizza chip di terze parti, come le GPU di Nvidia e le CPU di Intel.
L'azienda ha anche sviluppato un supercomputer da 16.000 GPU, chiamato Research SuperCluster (RSC), che viene utilizzato per addestrare modelli di Intelligenza Artificiale di grandi dimensioni.

AMD
AMD, anch'esso un gigante dei chip, ha annunciato il 7 dicembre 2023 i nuovi chip Instinct MI300X per l'Intelligenza Artificiale.
Questi chip offrono prestazioni ed efficienza energetica notevolmente migliorate rispetto ai modelli precedenti.

La caratteristica più importante dei nuovi chip MI300X è la presenza di 192 gigabyte di memoria nella HBM3[16].
Questa memoria ultraveloce è in grado di gestire grosse moli di dati in trasferimento, rendendo i chip MI300X particolarmente adatti per i modelli AI più grandi.

I nuovi chip MI300X sono in grado di eseguire l'inferenza AI a una velocità di 300 teraflops, il che li rende circa il 30% più veloci rispetto ai modelli precedenti. Inoltre, i chip MI300X consumano circa il 20% di energia in meno rispetto ai modelli precedenti, rendendoli più efficienti dal punto di vista energetico.

AMD ha già annunciato che i chip MI300X saranno utilizzati da Meta e Microsoft per i loro prodotti e servizi basati sull'Intelligenza Artificiale.

Cenni a nuove architetture di chip per AI orientate al risparmio energetico

D-Matrix Corsair: la scommessa del "in memory computing"

D-Matrix è una startup inglese attiva nel campo dei chip "analogici" per AI.

Sebbene le GPU siano incredibilmente potenti per i giochi o il mining di criptovalute, le loro prestazioni non sono ottimali per l'esecuzione di un'Intelligenza Artificiale generativa.
L'azienda dichiara[17]:
- miglioramento del costo totale di proprietà (TCO) di 13-27 volte rispetto alle GPU quando si eseguono modelli LLaMA2-13B con contesto 4K;
- efficienza energetica 20 volte migliore;
- latenza 20 volte inferiore;
- larghezza di banda della memoria 40 volte superiore.

Mythic
Un'altra startup, Mythic[18] sfrutta la bassa latenza e il basso consumo energetico del calcolo analogico. Mythic afferma di aver creato una soluzione unica e rivoluzionaria che promette di affrontare i limiti del digitale fornendo allo stesso tempo specifiche migliorate rispetto alle migliori soluzioni digitali della categoria: un motore di calcolo analogico (ACE).

Expedera
Expedera[19] è un'altra azienda di semiconduttori con sede in California. L'azienda ha collaborato con importanti aziende tecnologiche, tra cui Google, Microsoft e Amazon ed ha recentemente aperto un centro di R&D in UK.

Note
(16) https://en.wikipedia.org/wiki/High_Bandwidth_Memory
(17) https://www.d-matrix.ai/wp-content/uploads/2023/09/d-Matrix-WhitePaper-Approved-w-cover.pdf
(18) https://mythic.ai/product/
(19) https://www.expedera.com/

Nuovi approcci software per rendere energeticamente efficienti i sistemi AI

Nel complesso, il nuovo hardware ha il potenziale per svolgere un ruolo significativo nel rendere l'AI generativa più efficiente dal punto di vista energetico. Tuttavia, è importante notare che l'hardware non è l'unico fattore che influisce sul consumo energetico. Anche il software utilizzato per addestrare e distribuire modelli di Intelligenza Artificiale generativa svolge un ruolo.
Ecco alcuni miglioramenti del software[20] che possono aiutare a ridurre l'impatto energetico dell'addestramento e dell'utilizzo dell'IA:

- l'utilizzo di architetture di rete neurali convoluzionali più efficienti, come le reti neurali convoluzionali sparse;
- l'utilizzo di tecniche di compressione dei parametri, come la quantizzazione;
- l'utilizzo di algoritmi di apprendimento più efficienti, come l'apprendimento con rinforzo;
- l'utilizzo di tecniche di ottimizzazione della memoria, come la condivisione dei dati;
- l'utilizzo di hardware dedicato, come acceleratori AI, per l'esecuzione dei modelli.

Esistono poi diversi studi su modelli LLM pensati per essere efficienti dal punto di vista energetico.

Questi studi si concentrano su una serie di tecniche, tra cui:

- l'utilizzo di architetture di rete neurale più efficienti dal punto di vista energetico. Ad esempio, alcuni studi hanno proposto l'utilizzo di architetture di rete neurale che utilizzano meno pesi e connessioni;
- l'utilizzo di tecniche di compressione dei dati. Queste tecniche possono essere utilizzate per ridurre la dimensione dei modelli LLM, il che può ridurre il consumo energetico necessario per l'addestramento e l'esecuzione dei modelli;
- l'utilizzo di tecniche di ottimizzazione del codice. Queste tecniche possono essere utilizzate per migliorare l'efficienza del codice utilizzato per implementare i modelli LLM.

Riportiamo a titolo di esempio studi su modelli LLM efficienti dal punto di vista energetico[21]:

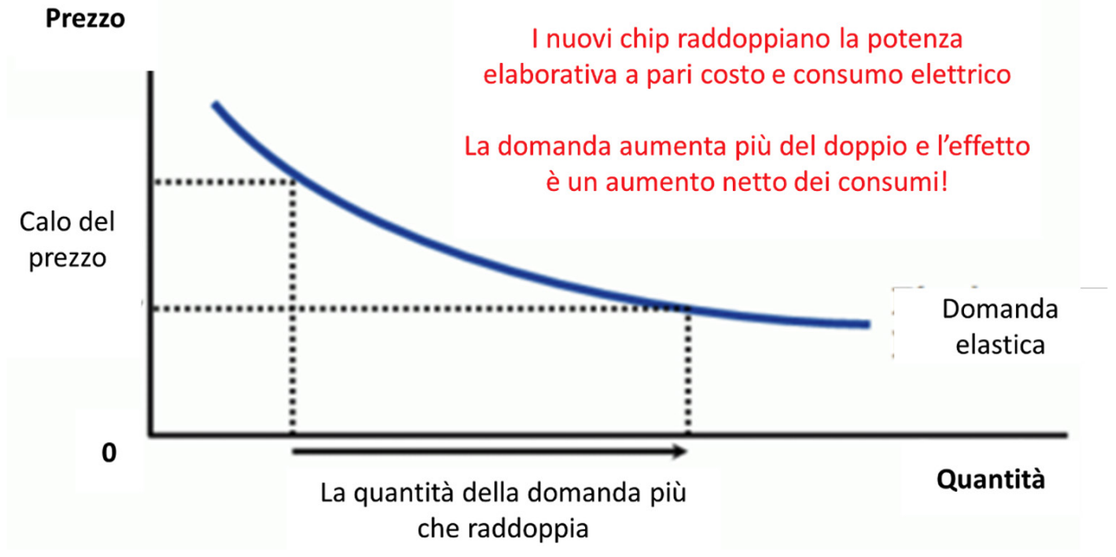
- uno studio di Google AI ha proposto un nuovo tipo di architettura di rete neurale chiamata "Sparse Transformer". I Transformer sono un tipo di rete neurale che viene comunemente utilizzato nei modelli LLM. I Sparse Transformer utilizzano meno pesi e connessioni rispetto ai Transformer tradizionali, il che li rende più efficienti dal punto di vista energetico;
- un altro studio di Google AI ha proposto una tecnica di compressione dei dati chiamata "Quantization". La quantizzazione consente di ridurre la dimensione dei modelli LLM senza sacrificare troppo le prestazioni;
- un terzo studio di Nvidia ha proposto una tecnica di ottimizzazione del codice chiamata "Tensor Core Optimization". La Tensor Core Optimization consente di migliorare l'efficienza del codice utilizzato per implementare i modelli LLM su GPU Nvidia.

Questi studi stanno dimostrando che è possibile sviluppare modelli LLM efficienti dal punto di vista energetico senza sacrificare troppo le prestazioni. Questo è importante per ridurre l'impatto ambientale dell'Intelligenza Artificiale.

Conclusioni

L'effetto Jevons[22] si basa sull'idea che un aumento dell'efficienza nell'uso di una risorsa può portare a un aumento assoluto dell'uso di quella risorsa anziché a una diminuzione.
Ad esempio, se si sviluppa una tecnologia più efficiente dal punto di vista energetico per un certo processo industriale, potrebbe diventare più conveniente utilizzare quella tecnologia, spingendo le imprese ad espandere la produzione o ad utilizzare tale processo più frequentemente.
Questo aumento dell'efficienza potrebbe portare a un aumento netto del consumo totale di energia, anziché ad una diminuzione, proprio perché l'efficienza rende più conveniente utilizzare quella risorsa.

Figura 5: Curva domanda secondo l'effetto Jevons[23]



Note

(20) http://versodatascience.com/pruning-neural-networks-1bb3ab5791f9
https://link.springer.com/chapter/10.1007/978-3-030-01249-6_18
https://link.springer.com/article/10.1007/s10766-018-00624-9

(21) Google AI: Sparse Transformers for Efficient Natural Language Processing https://arxiv.org/abs/2201.07285
Google AI: Quantized Transformers for Efficient Language Modeling https://arxiv.org/abs/2203.07803
Nvidia: Tensor Core Optimization for Efficient Large Language Model Inference https://arxiv.org/abs/2203.08188

Note

(22) https://en.wikipedia.org/wiki/Jevons_paradox

(23) Grazie a Nicola Magnani per la discussione sul paradosso di Jevons

Per evitare questa trappola occorre adottare specifiche politiche che incentivino un uso più responsabile delle risorse.
L'AI sta vivendo con lo sviluppo dei sistemi generativi basati su LLM un'ulteriore accelerazione.
Tuttavia, lo sviluppo e l'utilizzo di soluzioni AI, unito al gran numero di progetti in partenza, ha un impatto energetico importante. Le tecniche per ridurre questo impatto richiedono chip sempre più specializzati, i cui costi di R&D e di produzione sono enormi, ed un utilizzo più ottimale dei modelli attraverso soluzioni software.

Entrambe le strade saranno necessariamente perseguite nei prossimi anni.■

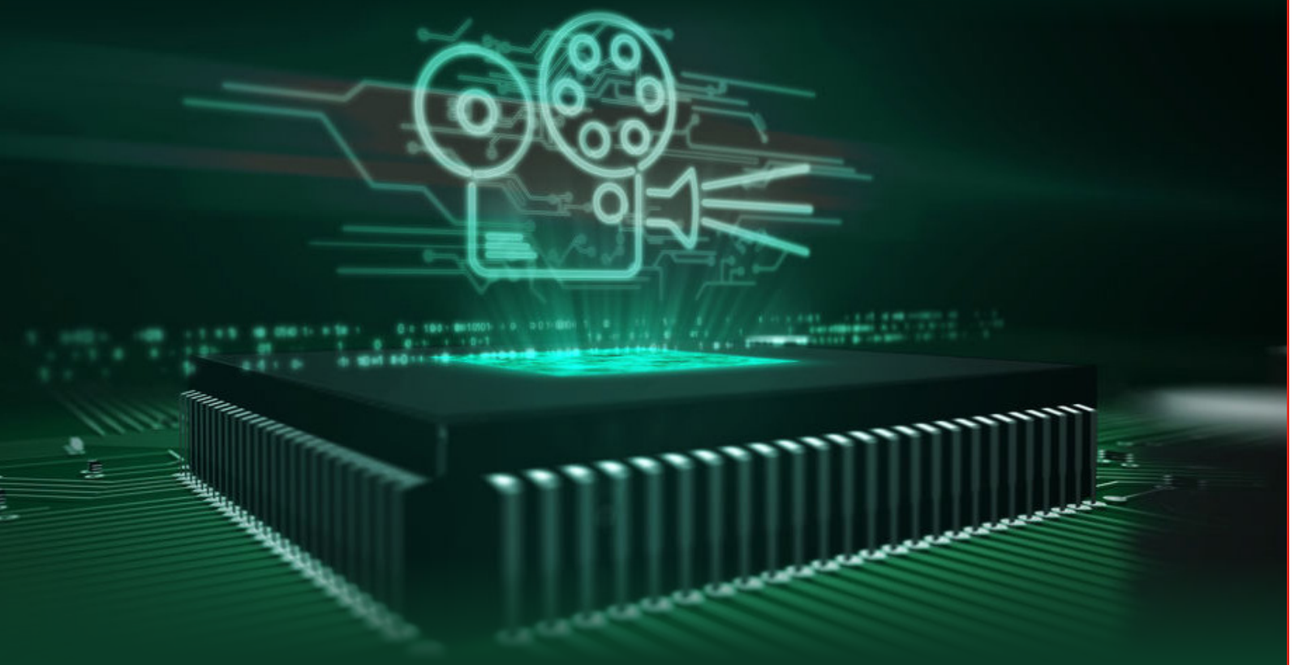
Autori

Gabriele Elia *gabriele.elia@telecomitalia.it*

Ingegnere elettronico e Dottore di Ricerca al Politecnico di Torino, in Azienda dal 1994. Lavora presso la divisione Technology Innovation di TIM, guida il gruppo Technological Scouting, Trend Analysis & Future Center dove si portano in evidenza i trend di medio termine di reti e softwarizzazione; vita digitale; digitalizzazione delle industrie; bigdata, robotica e AI; trend provenienti dal mondo scientifico e di processi di innovazione.
Si è sempre occupato di innovazione nei settori tecnologici sui temi servizi IP, media, applicazioni del broadband fisso e mobile, sia più recentemente di iniziative di Open Innovation, startup acceleration e costruzione di collaborazioni innovative di ricerca, formazione e imprenditorialità con il tessuto universitario. Ha iniziato il suo lavoro negli anni '90 nel primo gruppo di progetto sui temi Internet in Telecom Italia, che sviluppò le fasi iniziali di Interbusiness, TOL - Telecom On Line e poi TIN.IT, occupandosi dell'architettura della rete di accesso e del centro servizi. Autore di vari brevetti, è Ingegnere elettronico e Dottore di Ricerca al Politecnico di Torino, è stato assunto in CSELT, il Centro Studi e Ricerche di Telecom Italia a Torino nel novembre 1994. ■

Intelligenza Artificiale per la Compressione Video

Diego Gibellino



L'Intelligenza Artificiale è entrata a pieno titolo tra l'insieme di tecnologie impiegate stabilmente dai Media. Tecniche di Machine Learning sono in grado di comprendere, classificare, manipolare, ottimizzare formati e modalità di distribuzione e raccomandare contenuti sulla base delle abitudini e preferenze di fruizione degli utenti. L'Intelligenza Artificiale generativa può creare immagini, video, audio e testi, supportando i processi di creazione artistica, di produzione e post-produzione. La Comunità Scientifica negli ultimi anni ha sviluppato nuove tecniche basate sulle reti neurali profonde che possono essere applicate alla compressione dei video. Gruppi di lavoro internazionali, come ISO/IEC JTC1 SC29 WG5 (MPEG)/ITU-T SG16 Joint Video Experts Team (JVET) e Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI), hanno avviato attività esplorative che dovrebbero portare alla definizione di una nuova generazione di compressori video (codec) basati su IA nel corso di questo decennio, con benefici in termini di risparmio di banda e flessibilità di utilizzo rispetto alle soluzioni oggi disponibili. Queste nuove tecnologie potrebbero rivoluzionare servizi come streaming di video e gaming, videoconferenza, realtà virtuale e aumentata, e public safety.

Intelligenza Artificiale e Media

Gli algoritmi di *Machine Learning* in ambito Media sono stati inizialmente impiegati per la gestione automatica di operazioni di indicizzazione e ricerca su cataloghi video e, sulla base delle preferenze di fruizione degli utenti, per realizzare motori di raccomandazione sempre più sofisticati che facilitano la ricerca e l'accesso a nuovi contenuti. Queste soluzioni si basano su tecnologie in grado di estrarre le caratteristiche principali dei contenuti, chiamate *feature*, e di sfruttare queste informazioni per operazioni di annotazione automatica, arricchimento e classificazione.

Altre tecniche permettono, a partire da un contenuto, di ricavare segmenti corrispondenti ad ogni scena o con particolari requisiti (presenza di oggetti, attori, personalità o con una particolare semantica associata). I metadati prodotti possono essere utilizzati per offrire una migliore navigazione del contenuto originale, per generare in modo automatico highlights da eventi sportivi appena conclusi, o per facilitare operazioni di ricerca ed estrazione di filmati da archivi digitali per la produzione di nuovi contenuti, per la realizzazione di servizi per i notiziari o per la creazione di asset a corredo del contenuto stesso (per esempio le immagini da presentare nell'interfaccia di navigazione dei servizi di streaming).

Le prime applicazioni di tecniche di *Machine Learning* ai processi di preparazione e distribuzione dei contenuti in streaming sono legate principalmente a soluzioni di Content Adaptive Encoding (CAE). Si tratta di soluzioni in grado di lavorare su uno specifico contenuto o su parti di esso (ad esempio ogni scena) analizzandone caratteristiche e complessità.
Processando un numero molto elevato di contenuti è possibile addestrare un algoritmo per individuare la migliore configurazione possibile in termini di parametri di transcodifica (numero di livelli ABR, risoluzione, framerate, bitrate, ...), ottimizzando il rapporto tra qualità percepita e banda impiegata per ogni singolo contenuto. Queste tecniche sono diventate sempre più efficaci e sono oggi disponibili in soluzioni commerciali di encoding e packaging proposte da vari fornitori, sia in modalità on premises che in cloud, fornendo fino al 30% di risparmio in termini di banda per qualità equivalente per scenari offline (on-demand). Nel 2021 Harmonic e TIM hanno collaborato all'estensione dello standard ISO/IEC 23001 Parte 10 "Carriage of timed metadata metrics of media in ISO-base file format" [1], proprio per permettere il trasporto delle informazioni generate per scenari di transcoding intelligente distribuito.

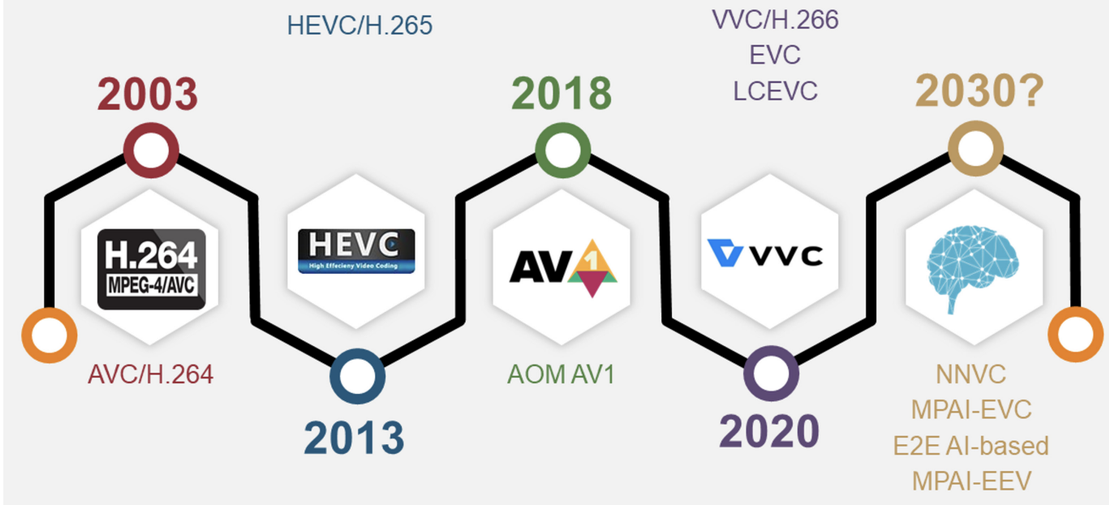
Un ulteriore, recente sviluppo è l'utilizzo di tecniche di Intelligenza Artificiale generativa, che attraverso modelli LLM (*Large Language Models*) possono creare testo, immagini e video artificiali originali basati sulle informazioni testuali di contesto (prompt) fornite da personale umano e su quantità estremamente elevate di dati utilizzati per l'apprendimento.
Queste soluzioni, in alcuni casi già commercialmente disponibili, stanno dimostrando grandi potenzialità per uno sfruttamento nel settore della produzione ed editing dei contenuti, ma il loro uso sta anche alimentando polemiche per le implicazioni etiche e di tutela delle proprietà intellettuali ed artistiche.

Stato dell'arte ed evoluzione dei compressori video

I compressori video permettono di produrre rappresentazioni estremamente compatte di un contenuto visuale naturale o sintetico rimuovendo le informazioni meno percepibili dal sistema visivo umano e applicando tecniche che sfruttano al massimo la ridondanza di informazione all'interno del contenuto stesso.
Si tratta di una tecnologia abilitante che ha permesso la creazione e lo sviluppo dei mercati televisivi digitali broadcast e IPTV negli anni Novanta e, più recentemente, l'esplosione dei servizi di streaming OTT sulle reti broadband. In oltre 30 anni di evoluzione, l'architettura delle soluzioni di compressione video più diffuse è rimasta sostanzialmente la stessa, basata su un modello di codifica ibrido che utilizza sia la compressione spaziale che quella temporale. Su questi principi di base sono state tuttavia definite soluzioni e standard sempre più efficienti che, circa ogni 10 anni, hanno dimezzato la banda richiesta per comprimere un contenuto ad una determinata qualità rispetto ai compressori della generazione precedente (Fig.1).

Nel 2020 il gruppo di lavoro congiunto ISO/IEC JTC1 SC29 WG5 (MPEG)/ITU-T SG16 Joint Video Experts Team (JVET), ha pubblicato lo standard internazionale noto come Versatile Video Coding (VVC o H.266)[2]. VVC è il codec video più efficiente disponibile ad oggi. Perfeziona ed estende ulteriormente l'insieme di tecnologie sviluppate per i codec di generazione precedenti quali AVC/H.264 e HEVC/H.265. Nello stesso anno, il gruppo di lavoro ISO/IEC JTC1 SC29 WG4 MPEG Video, ha pubblicato due ulteriori standard denominati Essential Video Coding (EVC) e Low Complexity Enhancement Video Coding (LCEVC) che possono offrire vantaggi specifici per alcuni scenari di servizio e di mercato.
A queste soluzioni si affianca AV1, il codec definito dal consorzio industriale Alliance for Open Media nel 2018, che, pur offrendo performance inferiori a VVC, viene reso disponibile in modalità royalty-free agli integratori e può contare su un buon livello di supporto tra i prodotti consumer come TV, Set-top box e laptop.

Figura 1: Evoluzione delle tecnologie e standard di compressione video



Il gruppo JVET, a partire dal completamento della prima versione dello standard VVC, ha avviato una serie di attività esplorative con l'obiettivo di valutare nuove tecnologie in grado di migliorare ulteriormente l'efficienza di compressione. Le attività si sviluppano su due filoni principali: il primo analizza nuove tecniche basate sulla manipolazione dei segnali, in linea con l'approccio seguito finora per l'evoluzione dei codec; il secondo considera per la prima volta l'impiego di tecniche di Intelligenza Artificiale e le reti neurali profonde applicati alla compressione video. Le due attività prendono il nome rispettivamente di "*Enhanced compression beyond VVC capability*" [3] e di "*Neural Network-based Video Coding*" (NNVC) [4].

Nell'ambito dei lavori su NNVC vengono studiate tecniche che adottano le reti neurali profonde per migliorare le prestazioni di alcuni componenti dell'architettura di compressione ibrida tradizionale basata su VVC (in-loop filter, predizione inter e intra frame), ma anche approcci innovativi che rimpiazzano completamente o in parte l'attuale architettura esistente con architetture end-to-end AI, basate su autoencoder. Si tratta, come detto, di un'attività esplorativa e non ancora di un vero e proprio nuovo standard, ma è significativo come sia già disponibile un modello software di riferimento utilizzato per la verifica delle prestazioni rispetto a VVC e per l'esecuzione degli *exploration experiment* per valutare le diverse proposte dei vari membri.
Recentemente JVET ha pubblicato un'estensione dello standard Versatile Supplemental Enhancement Information (VSEI) che permette di segnalare, attraverso appositi metadata nel bitstream del compressore VVC, la possibilità per un decoder conforme di utilizzare una particolare rete neurale come filtro da applicare a valle del processo di decodifica per migliorare la qualità dell'immagine risultante o generare nuovi frame tramite interpolazione. Questa soluzione permette la massima flessibilità di utilizzo lato client, mantenendo la compatibilità a livello di bitstream.

Un interessante ulteriore sviluppo è l'avvio delle attività MPEG relative a Video Coding for Machines (VCM), un nuovo standard video che ha lo scopo di definire un formato compresso facilmente utilizzabile da processi di analisi software visuale. In questo caso l'obiettivo è di rendere possibili scenari di monitoring e analisi automatiche quali object detection e object tracking e occasionali fruizioni da parte di esseri umani, per supervisione e conferma dei risultati. Lo standard prevede la possibilità di riutilizzare come base la codifica VVC o la codifica di feature salienti.

Creato nel 2020, Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) è un ente no-profit per la definizione di standard basati sull'impiego dell'Intelligenza Artificiale per la codifica dei dati, inclusi contenuti video, con

una chiara gestione degli IP attraverso un apposito licence framework.

Il gruppo sviluppa, tra gli altri, due progetti relativi alla compressione video: "AI-based End-to-End Video Coding" (MPAI-EEV) [5] e "AI-Enhanced Video Coding" (MPAI-EVC) [6]. Nel primo vengono esplorate tecniche basate sull'impiego di autoencoder, mentre nel secondo, a partire dal profilo base del codec EVC, vengono analizzate soluzioni basate su reti neurali profonde per sostituire i componenti esistenti di predizione intra (cioè sfruttamento delle ridondanze spaziali), di in-loop filtering e per integrare funzionalità di super resolution.

Le reti neurali profonde applicate ai segnali video

Le reti neurali profonde (Deep Neural Network - DNN), ed in particolare le reti convoluzionali (CNN), hanno dimostrato di essere, almeno per ora, le tipologie di reti che offrono le migliori performance in scenari di classificazione e riconoscimento visuale. Si basano su una serie di strati in grado di ricavare progressivamente le informazioni salienti da una o più immagini in ingresso. Il processo prevede l'estrazione di caratteristiche e pattern visuali con livelli di astrazione via via maggiori procedendo attraverso gli strati successivi della rete, in modo molto simile a quanto avviene nel sistema percettivo umano. Le operazioni effettuate sono una serie progressive di convoluzioni matriciali e vettoriali e relativi sottocampionamenti. I risultati in uscita dalla rete vengono infine generati da uno o più strati completamente connessi.
Il processo di apprendimento (*training*) delle reti CNN utilizza un ampio numero di contenuti con caratteristiche diverse (*dataset*) per ottimizzare i parametri di ciascuno strato rispetto ad una funzione di costo definita.
Una CNN opportunamente addestrata è in grado di generalizzare, adattandosi efficacemente a nuovi contenuti (processo di inferenza) e fornendo risultati migliori rispetto ad algoritmi basati su una serie di logiche predefinite. La capacità di estrarre le caratteristiche salienti di un contenuto visuale (feature), permette rappresentazioni estremamente compatte dei contenuti visuali.

Su questo principio, la comunità scientifica ha impiegato le CNN anche per le operazioni vere e proprie di compressione video.

Estensione delle architetture ibride di compressione con DNN

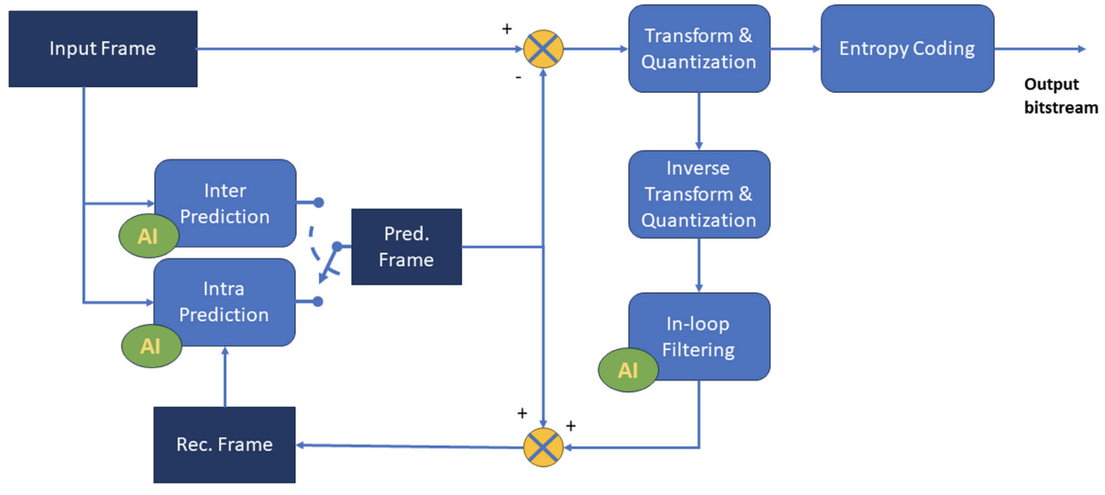
Questo approccio mantiene l'architettura tradizionale ibrida a blocchi dei compressori video sostituendo o aggiungendo componenti basati su reti neurali per specifiche funzionalità. Il diagramma sotto riportato illustra il processo di codecodifica impiegato dai compressori video moderni, evidenziando gli stadi per i quali si stanno valutando alternative basate su AI.
Ciascun frame viene suddiviso in blocchi di diversa dimensione per le componenti di luminanza e crominanza. Ad ogni blocco viene sottratto il risultato di una predizione che può utilizzare informazioni presenti nel frame stesso o in quelli temporalmente adiacenti al frame corrente.
Il residuo viene elaborato con operazioni di trasformazione e quantizzazione.
Infine, un processo di codifica entropica si occupa di comprimere in modo efficiente i coefficienti rimanenti diversi da zero generando il vero e proprio bitstream. Nel processo di codifica i coefficienti vengono inoltre riportati nel dominio originale attraverso un processo di trasformata inversa e dequantizzazione.
Il risultato, ai quali vengono applicati ulteriori filtri, viene utilizzato insieme al frame predetto per generare il frame di riferimento per nuove predizioni al ciclo successivo.
Come riportato nel diagramma (Fig.2), in ambito JVET e MPAI-EVC, si sta lavorando per valutare alternative per i seguenti componenti:

- *In-loop filtering* per migliorare la qualità del frame ricostruito eliminando artefatti derivanti dal processo di suddivisione in blocchi dell'immagine;
- *Predizione Intra Frame* per generare predizioni relative al blocco in oggetto utilizzando le informazioni presenti nel frame;
- *Predizione Inter Frame* per generare predizioni relative al blocco in oggetto utilizzando le informazioni presenti nei frame temporalmente adiacenti.

Ulteriori temi di studio sono l'impiego di:

- *Super-resolution*, in cui per ogni frame, è possibile scegliere se codificarne una versione sottocampionata ed applicare nel processo di ricostruzione del frame una rete neurale in grado di effettuare l'upsampling del frame alla dimensione originaria;
- *Post filtering*, in cui vengono attivate reti neurali specifiche a valle del processo di decodifica da parte del client

Figura 2: Codificatore Ibrido MPEG e componenti per i quali sono in fase di studio tecniche basate su AI



per migliorare la qualità video risultante senza richiedere modifiche al codec.

Il modello di riferimento software sviluppato da JVET, denominato NNVC e attualmente disponibile in versione 5.0, si basa sulla combinazione dei seguenti strumenti: un numero di reti neurali, ognuna per le diverse dimensioni di blocchi supportate, dedicate alla predizione intra frame costituite da strati interamente connessi (non basati su CNN) [7]. Una rete convoluzionale da utilizzare nello stadio di in-loop filtering in parallelo al deblocking filter, ottimizzata per mantenere una bassa complessità (NNLF low OP) e una rete denominata Unified NNLF HP, sempre per in-loop filtering, con maggiore complessità e prestazioni migliori. Sono inoltre integrate funzionalità di NN-based super resolution e post filtering (seppure disattivate nella configurazione di default).
È in fase di integrazione anche una proposta di rete neurale per la generazione di frame di riferimento che migliorino il processo di predizione tra frame denominata *Deep Reference Frame* (DRF) generation che utilizza algoritmi di optical flow [8].

Ad oggi il modello NNVC nella configurazione standard (con una serie di strumenti disabilitati) offre prestazioni in grado di migliorare di circa il 20% le prestazioni di VVC, a qualità confrontabile, con un incremento di complessità tuttavia significativo, in particolare sul decoder.

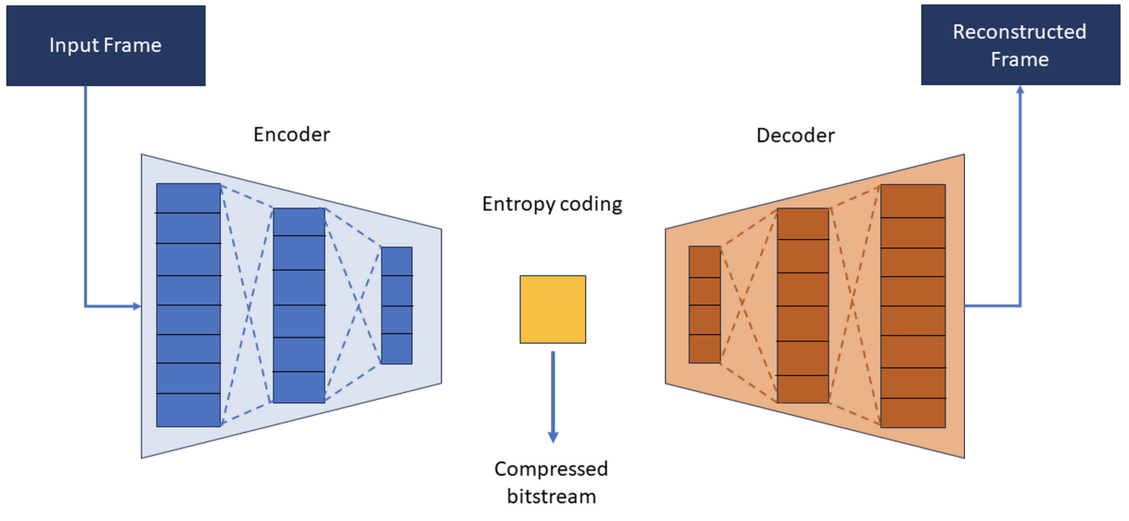
Il gruppo MPAI-EVC sta lavorando ad un *evidence project* con l'obiettivo di migliorare, attraverso un'opportuna combinazione di tecniche basate su reti neurali, le performance del profilo base di Essential Video Coding (EVC). Tecniche di super resolution vengono integrate in uno stadio di post processing a valle della decodifica, è stata migliorata la predizione intra facendo uso di una CNN e, inoltre, sono in fase di analisi soluzioni di in-loop filtering basate su Multi-Frame In-Loop Filter (MIF-NET) [9].
Attualmente il miglioramento rispetto ad EVC è pari al 25%.
Le differenti proposte e tecniche vengono valutate in base all'incremento di prestazioni in termini di efficienza di codifica rispetto alle versioni di reference software VVC nel caso di JVET e EVC per MPAI-EVC.
Per la valutazione dell'efficienza di codifica viene utilizzato il valore BD-rate, basato su PSNR (*Peak Signal-to-Noise Ratio*), tradizionalmente impiegato per lo sviluppo dei codec video in ambito MPEG e ITU-T. Vengono inoltre riportati anche il valore MS-SSIM (Multi-Scale Structural Similarity Index), una metrica in grado di catturare maggiori informazioni sulla struttura vera e propria dell'immagine e il VMAF (Video Multi-Method Assessment Fusion), metrica più correlata alle prove soggettive.
È importante notare come questi valori vengano analizzati in parallelo a valori derivanti da campagne di test soggettive, ad indici che valutano la complessità computazionale delle reti utilizzate (kMAC: numero di operazioni di multiply-accumulate per 1000 campioni durante lo stadio di inferenza o FLOPS: numero di operazioni in floating point richiesto per un singolo passaggio) e la dimensione in memoria utilizzata per i parametri, insieme alle percentuali di incremento per i tempi di codifica e decodifica.

Compressori end-to-end DNN

Un approccio alternativo a quello descritto nel paragrafo precedente consiste nella creazione di framework end-to-end basati su reti neurali profonde, in grado di ottimizzare globalmente, attraverso il processo di apprendimento, le tecniche di predizione intra, gestione del moto e codifica dei residui.
In questo modello si superano i limiti imposti dallo schema a blocchi ibrido tradizionale, riducendo quindi sensibilmente gli artefatti derivanti dalla scomposizione in blocchi delle immagini. L'obiettivo è utilizzare un numero minimo di reti neurali, in modo da semplificare sensibilmente l'architettura, in grado di apprendere autonomamente il modo migliore per gestire il trade-off tra bitrate e qualità visuale percepita.
I modelli attualmente analizzati utilizzano autoencoder, particolari reti neurali in grado di apprendere, in modalità non supervisionata, la codifica più efficiente per una determinata tipologia di contenuto in uno spazio latente diverso da quello basato sui pixel e di decodifica per ripristinare il dato originale, secondo lo schema di alto livello riportato nella Fig.3.
Gli autoencoder vengono impiegati per la predizione intra, sfruttando soluzioni ad alte prestazioni sviluppate per la compressione delle immagini, per la codifica dei residui e per la stima del moto e predizione inter frame. In quest'ultimo caso, tipicamente si combinano algoritmi di optical flow in grado di modellare il movimento a livello di pixel e non solo come semplici traslazioni o trasformazioni geometriche tra blocchi.

Il gruppo MPAI-EEV ha realizzato un modello di compressore end-to-end che integra reti autoencoder per la codifica dei residui e del moto. Nello specifico, sulla base delle informazioni più recenti pubblicate da MPAI, l'architettura utilizza Deep

Figura 3: Schema di architettura E2E autoencoder



Video Compression (DVC) [10] per la codifica intra e una rete *multi-scale optical flow pyramid* per la stima del moto [11]. Reference frame e vettori di moto sono quindi inseriti in una rete per la compensazione del moto (MC-Net) per generare il frame predetto. Residui e vettori di moto vengono codificati come immagini in modo da permettere un'ottimizzazione globale con una unica funziona di costo.
Il gruppo JVET sta analizzando le performance di alcuni nuovi tool end to end applicati alla codifica intra frame [12] e predizione inter frame [13], associati comunque all'architettura basata su compressore VVC.
Le proposte non sono al momento considerate abbastanza mature per la definizione di *exploration experiment* specifici.

Vantaggi e limiti

L'impiego di reti neurali profonde per la compressione di contenuti video sta rivoluzionando un settore che per oltre trent'anni ha lavorato con l'obiettivo di perfezionare ed estendere l'architettura ibrida a blocchi derivata dalle prime soluzioni e standard adottati dall'industria. Tale processo è tuttora in corso, con attività MPEG/ITU-T SG16 che mostrano ulteriori margini di miglioramento possibili rispetto a VVC sfruttando sostanzialmente la stessa architettura, a fronte di un moderato incremento nella complessità dei nuovi strumenti di codifica.
L'approccio tradizionale si basa su raffinate euristiche in grado di ottimizzare localmente per ogni modulo il trade off costo in bit rispetto alla qualità visuale fornita.

Le logiche di funzionamento sono definite sulla base di trasformazioni lineari e modelli statistici predefiniti, con guadagni di performance possibili tipicamente attraverso l'aggiunta di nuovi tool o ampliando ad esempio le modalità e aree di ricerca per la generazione dei predittori.

Attraverso l'utilizzo di reti neurali profonde è possibile superare questi limiti. I compressori basati su reti neurali profonde possono in prospettiva offrire migliori prestazioni, sfruttando feature ed informazioni di contesto ed adattandosi in modo efficace alle diverse tipologie di video in ingresso grazie al processo di apprendimento e generalizzazione effettuato su un numero elevato di contenuti.
Attraverso la comprensione della struttura del contenuto e di come viene interpretato dal sistema visivo umano è possibile gestire ciascun frame e ciascuna area all'interno di un frame in modo diverso in base all'impatto relativo sulla qualità percepita globale e combinare in modo ottimale logiche di modifica dinamica dei valori di framerate e di risoluzione utilizzate nel processo di codifica.

I livelli di prestazione e flessibilità offerti da questi compressori di nuova generazione hanno però, almeno per ora, impatti rilevanti sulla complessità nelle implementazioni lato encoder, ma soprattutto lato decoder.
I tempi di decodifica legati alle operazioni sequenziali associate a ciascun strato e i requisiti di memoria per i vari parametri necessari alle reti neurali possono esplodere per soluzioni non ottimizzate, rendendo irrealistiche implementazioni su dispositivi utente a costo contenuto.
La compressione video basata su reti neurali è una tecnologia ancora in fase embrionale, nonostante sia oggetto di un numero crescente di studi. Negli ultimi anni c'è stata una forte accelerazione sullo sviluppo e ottimizzazione di queste soluzioni, un processo a cui stanno partecipando tutti i maggiori attori globali nel mercato ICT e Media e che ha recentemente registrato notevoli progressi.
Il percorso per selezionare, ottimizzare ed integrare in modo efficace le migliori proposte basate su reti neurali non è ad oggi completamente definito, per questo motivo si stanno valutando due approcci diversi basati rispettivamente sulla sostituzione o aggiunta di elementi alla tradizionale architettura di codifica ibrida e su nuove architetture end to end, che potranno avere tempistiche di realizzazione e prestazioni diverse.

ISO/IEC JTC1 WG5 MPEG/ITU-T SG16 JVET e MPAI stanno esplorando in modo indipendente entrambi gli approcci, con l'obiettivo di produrre standard utilizzabili dall'industria nel corso di questo decennio.■

Bibliografia

1. Carriage of timed metadata metrics of media in ISO base media file format - Amendment 1: Support for content-guided transcoding and spatial relationship of immersive media, Standard ISO/IEC 23001-10:2020/AMD.1:2021, ISO/IEC JTC 1, Sep. 2021
2. Versatile Video Coding, Standard ISO/IEC 23090-3 2nd Ed., ISO/IEC JTC 1, Sep. 2022
3. JVET-AE2025, “Algorithm description of Enhanced Compression Model 10 (ECM 10)”
4. JVET-AE2019, “Description of algorithms and software in neural network-based video coding (NNVC) version 4”
5. AI-based End-to-End Video Coding (MPAI-EEVC), https://mpai.community/standards/mpai-eev/
6. AI-Enhanced Video Coding (MPAI-EVC), https://mpai.community/standards/mpai-evc/
7. JVET-AD0212-v5, “AHG11: neural network-based intra prediction with reduced complexity”, input document to JVET
8. JVET-AD0162, EE1-2.1-related: DRF Model without QP Input, input document to JVET
9. T. Li, M. Xu, C. Zhu, R. Yang, Z. Wang and Z. Guan, “A Deep Learning Approach for Multi-Frame In-Loop Filter of HEVC,” in IEEE Transactions on Image Processing, vol. 28, no. 11, pp. 5663-5678, Nov. 2019, doi: 10.1109/TIP.2019.2921877
10. “DVC: An End-to-End Deep Video Compression Framework,” Guo Lu et al., CVPR 2019
11. Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4161–4170, 2017
12. JVET-AA0063-v2, “AHG 11: A hybrid codec using E2E image coding combined with VVC video coding”, input document to JVET
13. JVET-Z0077, “AHG11: Extension of DOVC to Regular 2D Videos”, input document to JVET

Acronimi

ABR	Adaptive Bitrate Streaming
AI	Artificial Intelligence
CAE	Content Adaptive Encoding
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DRF	Deep Reference Frame
DVC	Deep Video Compression
E2E	End-to-End
EVC	Essential Video Coding
IA	Intelligenza Artificiale
JVET	Joint Video Experts Team
LCEVC	Low Complexity Enhancement Video Coding
LLM	Large Language Models
MIF-NET	Multi-Frame In-Loop Filter
MPAI	Moving Picture, Audio and Data Coding by Artificial Intelligence
MPAI-EEV	AI-based End-to-End Video Coding
MPAI-EVC	AI-Enhanced Video Coding
MS-SSIM	Multi-Scale Structural Similarity Index
NNVC	Neural Network-based Video Coding
PSNR	Peak Signal-to-Noise Ratio
VCM	Video Coding for Machines
VMAF	Video Multi-Method Assessment Fusion
VSEI	Versatile Supplemental Enhancement Information
VVC	Versatile Video Coding

Autori

Diego Gibellino *diego.gibellino@telecomitalia.it*

Responsabile del laboratorio di Tecnologie Video & TV di TIM.
Partecipa attivamente a diversi Enti di Standardizzazione e Forum Industriali Nazionali e Internazionali.
È Presidente della Commissione Tecnica Nazionale UNI/CT 512 “UNINFO SC29” per JPEG e MPEG, è capo della delegazione italiana in SC29 e advisor in MPEG.
È membro dello Steering Board DVB. ■

Glossario

Artificial Intelligence (AI):

L'Intelligenza Artificiale è un campo dell'informatica che si occupa di creare sistemi e algoritmi in grado di eseguire compiti che richiedono intelligenza umana, come il ragionamento, il problem solving, il riconoscimento di pattern e il linguaggio naturale. L'AI può includere una vasta gamma di tecniche e approcci, tra cui il machine learning, il deep learning e l'elaborazione del linguaggio naturale.

Deep Learning:

Nel Deep Learning, gli algoritmi sono basati su reti neurali artificiali profonde (Deep Neural Networks o DNN). Queste reti sono composte da strati di neuroni artificiali interconnessi e sono in grado di apprendere rappresentazioni complesse dei dati direttamente dai dati stessi. Il termine "profondo" si riferisce al fatto che queste reti possono essere molto complesse, con molti strati nascosti. Nel Deep Learning, le reti neurali sono in grado di apprendere automaticamente le rappresentazioni dei dati, eliminando in gran parte la necessità di un'estesa ingegneria delle caratteristiche. Questo rende il processo di addestramento più automatizzato e meno dipendente dalla conoscenza esperta. Il Deep Learning ha dimostrato di avere prestazioni eccezionali su grandi dataset, e spesso richiede grandi quantità di dati per ottenere risultati significativi. L'aumento delle dimensioni dei dati è una delle ragioni per cui il Deep Learning è diventato così potente negli ultimi anni. Le reti neurali profonde possono richiedere molte risorse computazionali, tra cui GPU (unità di elaborazione grafica) potenti, per l'addestramento. Sono più complesse da addestrare rispetto a molti algoritmi di Machine Learning tradizionali.

Generative AI:

La Generative AI è una sottocategoria dell'Intelligenza Artificiale che si concentra sulla creazione di modelli in grado di generare dati o contenuti originali. Questi modelli possono essere addestrati per generare testo, immagini, suoni e altro ancora. Quindi Genera nuovi dati invece di discriminare i Dati. Il Generative AI è spesso basato su reti neurali artificiali, come le reti generative avversarie (GAN) e le reti neurali ricorrenti (RNN).

Large Language Model:

Un Large Language Model (Modello di Linguaggio di Grandi Dimensioni) è un tipo di Generative AI che è stato addestrato su una vasta quantità di testo scritto per comprendere il linguaggio naturale e generare testo in modo coerente e contestualmente appropriato. Questi modelli sono spesso formati su miliardi di parole o più e possono essere utilizzati per una varietà di applicazioni, come l'elaborazione del linguaggio naturale, la generazione di testi creativi e molto altro.

Machine Learning (Apprendimento Automatico):

E' un sottoinsieme dell'Artificial Intelligence dove le decisioni vengono prese a partire da algoritmi costruiti a partire dai dati e non da regole derivanti dalla teoria. Nel Machine Learning non esiste una programmazione esplicita. Possono essere di Classificazione o Continui. Nel Machine Learning, gli algoritmi sono spesso basati su modelli statistici e possono utilizzare una varietà di tecniche, tra cui alberi decisionali, regressione lineare, support vector machines (SVM), k-means clustering e molto altro. Questi modelli sono spesso progettati manualmente per adattarsi ai dati e ai compiti specifici. Nel Machine Learning, spesso è necessario estrarre manualmente le caratteristiche rilevanti dai dati. Questo processo è noto come ingegneria delle caratteristiche e richiede una conoscenza esperta del dominio. Il Machine Learning può essere efficace anche con dataset di dimensioni relativamente ridotte o medie. Molti algoritmi di Machine Learning hanno una complessità computazionale inferiore rispetto al Deep Learning, il che li rende più adatti a situazioni in cui le risorse computazionali sono limitate.

Predictive Analytics (Analisi Predittiva):

L'Analisi Predittiva è un'applicazione dell'Intelligenza Artificiale che utilizza algoritmi statistici e machine learning per identificare pattern nei dati e fare previsioni o anticipazioni sul futuro. Questa tecnica è utilizzata in vari settori, tra cui il marketing, le finanze, la salute e molti altri, per migliorare la presa di decisioni basate sui dati.


TIM