

L'intelligenza artificiale generativa multimodale in ambito umanistico. Alcune considerazioni introduttive

Generative Multimodal Artificial Intelligence in the Humanities. An Introduction

Gino Roncaglia

Università degli Studi Roma Tre
gino.roncaglia@uniroma3.it

| abstract

L'articolo propone un'introduzione al tema delle intelligenze artificiali generative multimodali, descrivendone natura e caratteristiche, e si sofferma quindi su due aspetti che potrebbero costituire contesti particolarmente interessanti per il loro uso, in particolare nel campo delle digital humanities: il tema dell'accessibilità, con la generazione automatica di descrizioni alternative fortemente personalizzate basate su codici comunicativi diversi rispetto al contenuto primario, e il tema della crono- e georeferenziazione, con la possibilità di includere la dimensione geografica e quella cronologica all'interno delle caratteristiche di un modello multimodale.

The paper offers an introduction to the topic of multimodal generative artificial intelligence, describing the nature and characteristics of multimodal AI systems. It then focuses on two aspects of their use that could be of special interest, especially in the field of digital humanities: (a) accessibility, involving the automatic generation of highly personalized alternative descriptions based on communicative codes different from those of the primary content, and (b) chrono- and georeferencing, enabling the incorporation of both geographical and chronological dimensions within a multimodal model.

DOI 10.36158/97888929589208

Premessa

I sistemi di intelligenza artificiale generativa rappresentano indubbiamente la novità più significativa degli ultimi anni nell'ecosistema digitale, con effetti potenziali enormi e ancora in parte imprevedibili anche sulla società nel suo complesso. Per capire la portata rivoluzionaria di questo sviluppo anche per il mondo delle digital humanities e delle culture digitali, è bene ricordarne due aspetti fondamentali. Per un verso, i meccanismi di funzionamento di tali sistemi si allontanano notevolmente dai paradigmi prevalentemente logico-simbolici che erano stati propri della cosiddetta "Good, Old-Fashioned Artificial Intelligence" (GOFAI): al posto del tentativo di formalizzare i meccanismi di base del ragionamento logico deduttivo e del linguaggio e di programmare deterministicamente un computer in modo da fargli acquisire le relative capacità, si costruiscono e si addestrano reti neurali che funzionano su

base statistico-probabilistica; reti che modificano i propri parametri interni in maniera largamente oscura per gli stessi programmatori che ne hanno ideato l'architettura, e che vengono selezionate e via via perfezionate prevalentemente sulla base dei risultati empirici ottenuti, più che sulla base di una specifica concezione del funzionamento dell'intelligenza o del linguaggio.

Va notato che, pur se nel contesto di un paradigma assai diverso, la centralità dell'elemento linguistico viene comunque confermata: i Large Language Models costituiscono la forma più nota e diffusa di IA generativa, e anche i sistemi che generano contenuti visivi o sonori, o codice informatico, funzionano spesso attraverso "prompt" testuali. Ma il linguaggio è preso in considerazione attraverso processi associativi statistico-probabilistici di livello assai basso, relativi a "token" che sono spesso singole parole o parti di parole, se non addirittura gruppi di pochissimi caratteri. Questo porta a considerare sintassi e semantica – e, in un senso più generale, la stessa competenza linguistica – come proprietà in qualche misura "emergenti" in sistemi che lavorano a un livello altamente granulare e il cui comportamento di basso livello non è prevedibile deterministicamente.

L'interrogativo sulla portata che questo radicale cambiamento di prospettiva può avere anche sulla comprensione dei meccanismi alla base del nostro uso del linguaggio è credo legittimo: potremmo dire – anche se si tratta solo di un'analogia – che ci troviamo davanti a qualcosa di simile al cambiamento di paradigma vissuto dalla fisica nel passaggio dall'"universo della precisione" della fisica classica al mondo sconcertante e spesso imprevedibile della meccanica quantistica, in cui modelli statistico-probabilistici sostituiscono le leggi deterministiche.

Ho descritto e discusso altrove (Roncaglia, 2023) il contesto e alcune conseguenze di questi sviluppi, in particolare per quanto riguarda il rapporto fra il tradizionale modello architettonico-sistematico di organizzazione delle conoscenze e la produzione di contenuti in forma probabilistico-predittiva. Vorrei invece prendere qui sinteticamente in considerazione alcune prospettive che credo potrebbero avere un particolare rilievo, anche teorico, nel prossimo futuro, legate allo sviluppo dei sistemi multimodali e alle loro applicazioni al campo dell'accessibilità e alla costruzione di modelli per dati crono- e georeferenziati in campi quali la linguistica storica o la geografia sociale.

Le IA generative multimodali

La maggior parte dei sistemi di intelligenza artificiale generativa attualmente più diffusi sembrano poter essere caratterizzati, almeno in prima istanza, come monomodali: i Large Language Models – l'esempio più noto è rappresentato dalle prime versioni di ChatGPT – sono nati per produrre testi, mentre sistemi come Midjourney o Dall-e sono nati per produrre immagini. Si tratta dunque di sistemi orientati ciascuno a uno specifico codice comunicativo: quello linguistico nel caso dei LLM, quello visivo nel caso della generazione di immagini.

Fra i due casi vi è però una importante differenza. Nel primo, l'orizzonte è puramente linguistico: il corpus di addestramento è composto da testi, il prompt è testuale, l'output è testuale. Un sistema di questo tipo è dunque effettivamente monomodale. Nel secondo caso, invece, abbiamo a che fare con sistemi text-to-image: anche se il sistema produce un'immagine, lo fa sulla base di un prompt testuale che il sistema deve saper interpretare e associare alle immagini, e la dimensione linguistica è quindi necessariamente presente. Si tratta allora di sistemi multimodali? Non del tutto, giacché

l'output prodotto non è di norma in grado di “fondere” pienamente testo e immagini. Così, ad esempio, se chiediamo a MidJourney di produrre un manifesto pubblicitario o un fumetto, il sistema ci restituisce per ora un'immagine che include qualcosa che si avvicina all'apparenza grafica di un testo, ma non è un testo sintatticamente e semanticamente coerente (figura 1). In altri termini, sistemi text-to-image di questo tipo producono, appunto, solo immagini, e da questo punto di vista non sono dunque pienamente multimodali.

Questo esempio basta a farci capire che il concetto di “multimodalità” è più sfuggente di quanto non potrebbe sembrare, e prevede gradazioni. E la situazione si complicherebbe se volessimo descrivere e analizzare i molti meccanismi diversi attraverso i quali questi sistemi possono essere addestrati per produrre i loro contenuti: si può partire da immagini accompagnate da descrizioni, ma anche dall'integrazione fra un modello addestrato solo su immagini e un LLM addestrato solo su testi, che possono essere “sovrapposti” in uno spazio vettoriale comune in cui, ad esempio, l'immagine di un cane corrisponde a una matrice numerica non troppo lontana da quella che corrisponde alla parola “cane”. Non entrerà qui nei dettagli di funzionamento di questi e altri meccani-



Figura 1. Immagine generata da Midjourney; prompt: “A comic strip with a girl and her dog, talking, with the text of the conversation in a bubble”. Il sistema sa di dover generare il tipico testo di un fumetto, ma – essendo orientato unicamente alla produzione di immagini – non è in grado di produrre testo sintatticamente e semanticamente coerente.

smi possibili¹, limitandomi a osservare come sembri possibile individuare uno spettro abbastanza variegato di possibili significati del termine “multimodale”, al cui estremo più avanzato sono sistemi nativamente multimodali sia nelle tipologie di contenuti di addestramento, sia nella costruzione di uno spazio vettoriale concepito fin dall’inizio – e non solo per sovrapposizione – come capace di integrare contenuti basati su codici comunicativi diversi, sia infine nella produzione di output in cui codici comunicativi diversi sono pienamente ed efficacemente integrati (figura 2).

In questi sistemi, al posto dei Large Language Models abbiamo ad esempio – facendo riferimento al caso più diffuso, quello di sistemi che integrano testo e contenuti visivi – i cosiddetti Large Vision Language Models (VLMs). In un sistema di questo tipo i “token” possono essere sia visivi sia testuali (e combinazioni delle due cose), e il relativo embedding – la loro rappresentazione attraverso vettori numerici multidimensionali – avviene fin dall’inizio in uno spazio vettoriale comune. Ovviamente la multimodalità può estendersi anche ad altre tipologie di contenuti, oltre a testo e immagini: indicativo al riguardo è ad esempio un articolo piuttosto impressionante prodotto da un gruppo di ricercatori di Facebook Research che presenta ImageBind, «an approach to learn a joint embedding across six different modalities – images, text, audio, depth, thermal, and IMU data» (Girdhar et al., 2023). L’aspetto interessante di questo modello è che l’allineamento di modalità diverse fra loro collegate (ad esempio l’immagine di un gabbiano, la parola “gabbiano”, il suono del verso di un gabbiano, l’immagine di un gabbiano visto da una telecamera termica...) avviene in maniera in parte emergente: «It does not need datasets where all modalities co-occur with each other. Instead, we leverage the binding property of images, and we show that just aligning each modality’s embedding to image embeddings leads to an emergent alignment across all of the modalities» (Girdhar et al., 2023). Inoltre, il sistema mostra la capacità non solo di integrare ma anche di collegare semanticamente fra loro informazioni multicodicali: così, ad esempio, un input che unisca la parola “gabbiano” al suono delle onde del mare può produrre l’immagine di un gabbiano su una spiaggia.

La discussione sulle proprietà emergenti in sistemi di questo tipo è affascinante e meriterebbe attenzione specifica: troppo spesso quando si parla di proprietà emergenti il discorso sembra trasformarsi nella trama di un film di fantascienza e limitarsi al tema della coscienza, come se la coscienza fosse il solo tipo di (possibile) proprietà emergente. E dato che è decisamente implausibile attribuire coscienza ai sistemi attuali di IA generativa, inclusi quelli multimodali, si tende spesso a negare che questi sistemi abbiano davvero proprietà emergenti, o almeno proprietà emergenti interessanti. Mi sembra che questo approccio sia fuorviante, e porti a semplificare e sottovalutare un campo assai più complesso e sfaccettato². Non è questa la sede per approfondire questo tema, di rilevante portata filosofica, ma spero che le osservazioni proposte in questa sede possano aiutare a guardare all’ambito delle proprietà emergenti da una prospettiva almeno in parte diversa.

1. Per una introduzione relativamente accessibile a uno fra i molti esempi di sistemi text-to-image, Google Imagen, si veda (O’Connor, 2022). Per un approfondimento sul funzionamento dello stesso sistema si veda (Saharia et al., 2022).

2. Un discorso analogo si potrebbe fare rispetto a concetti come quelli di “intelligenza” o di “agency”, a loro volta spesso inclusi fra le possibili proprietà emergenti “interessanti”: il problema è che, proprio come nel caso della coscienza, per un verso si tratta di concetti assai difficili da definire e analizzabili in molti modi e da molti punti di vista diversi, per altro verso si tratta di (possibili) proprietà emergenti di livello particolarmente alto, che tendono a nascondere l’ampiezza dello spettro di proprietà emergenti che può essere interessante e utile prendere in considerazione.



Figura 2. Chat GPT 4 V è in grado di produrre, sulla base dello stesso prompt, un'immagine con testo (anche se – in assenza di indicazioni testuali specifiche – di qualità per ora inferiore rispetto all'output tipico del modello linguistico).

IA generative multimodali e accessibilità

In termini molto generali, per “accessibilità” si intende la capacità di un ambiente, di un dispositivo, di un servizio, di un contenuto o una risorsa, di risultare utilizzabile nel modo più facile, più efficace e più completo possibile da un insieme il più possibile ampio di utenti, indipendentemente dalle loro caratteristiche e capacità individuali e collettive, inclusa la provenienza etnica, le condizioni socioeconomiche, gli stili cognitivi, le eventuali disabilità o diverse abilità, la tipologia degli strumenti di accesso utilizzati³. L'accessibilità in ambito informatico (inclusa l'accessibilità di contenuti, risorse e servizi di rete) rappresenta dunque solo un sottoinsieme di un campo più ampio, e va tenuto presente che strumenti informatici possono essere utilizzati anche per garantire l'accesso ad ambienti, servizi o risorse fisiche: così, ad esempio, una app destinata a utenti non vedenti, che utilizzi la fotocamera di uno smartphone per raccogliere informazioni visive sull'ambiente circostante e trasformarle in informazione sonora, può essere utilizzata in

3. Ho cercato qui di riassumere in forma sintetica i principali tratti comuni ad alcune fra le molte e diverse definizioni del concetto di accessibilità proposte in letteratura; per una rassegna abbastanza recente che include alcuni fra i più importanti lavori di riferimento in materia, in particolare nel campo dell'interazione uomo-computer, si veda (Mack et al., 2021).

una varietà di situazioni della vita quotidiana, inclusa l'interazione con ambienti fisici (attraversare una strada, visitare un museo...).

Per quanto sintetiche, le osservazioni proposte nella prima sezione dell'articolo dovrebbero essere sufficienti a capire le possibilità potenzialmente rivoluzionarie dell'IA multimodale in questo ambito: una componente particolarmente significativa (anche se non certo l'unica) del lavoro svolto nel campo dell'accessibilità riguarda infatti proprio la compensazione di difficoltà relative alla percezione o interpretazione di un codice comunicativo, attraverso il ricorso a informazioni basate su altri codici comunicativi. Così, per fare solo alcuni fra i molti esempi possibili, una descrizione sonora potrà aiutare nella fruizione di contenuti visivi da parte di persone non vedenti, una trascrizione testuale potrà aiutare nella fruizione di contenuti sonori da parte di persone non udenti, una traduzione, una riformulazione più semplice o una descrizione visiva potranno aiutare a comprendere informazioni testuali complesse da parte di persone con competenze linguistiche insufficienti rispetto alla lingua di partenza, e così via. La disponibilità di strumenti che consentano il passaggio di un contenuto informativo da una modalità espressiva all'altra, e lo facciano in modo automatico, veloce e qualitativamente efficace, rappresenta dunque un'innovazione di enorme portata.

Va aggiunto che l'accessibilità ha molto spesso un carattere contestuale, che gli strumenti compensativi tradizionali non sempre riescono a gestire adeguatamente. Pensiamo ad esempio alla descrizione testuale di un reperto museale. Questa descrizione ha un primo compito di livello abbastanza generale: presentare e rendere "interpretabile" il reperto da parte di un visitatore non specialista. In alcuni casi (ma difficilmente in tutti) un museo può anche offrire una o più descrizioni ulteriori: ad esempio, una descrizione in Braille destinata a persone non vedenti, e/o una descrizione "facile" per bambine e bambini, e/o una descrizione più approfondita, ad esempio attraverso una guida o un'audioguida. Ciascuno di questi livelli descrittivi ha le proprie caratteristiche e si rivolge – per riprendere un termine largamente indagato in ambito semiotico – a un proprio "lettore ideale". Ma i visitatori reali offrono uno spettro di situazioni e necessità di fruizione assai più ampio. Possono esserci visitatori di varie età, con provenienze e interessi diversi, con curiosità specifiche, con diversi livelli di conoscenze e competenze pregresse. Chiaramente, nessun museo al mondo potrebbe personalizzare i propri strumenti informativi a un livello di differenziazione e di dettaglio tale da rispondere a queste molteplici situazioni e necessità. Un sistema di intelligenza artificiale, invece, potrebbe trasformarsi in una sorta di "guida" personalizzabile, capace di interagire con il visitatore per determinare il codice comunicativo, la tipologia e il livello di descrizione di volta in volta più opportuni e di generare poi il proprio output sulla base di un contesto specifico.

Per farci un'idea delle possibilità in quest'ambito, ho fornito a ChatGPT 4V (la V indica la presenza del modello visivo) un file con l'immagine del quadro "La notte stellata" di Vincent van Gogh, senza specificare nel prompt la natura dell'immagine ma chiedendogli unicamente di fornirne una descrizione. ChatGPT ha immediatamente riconosciuto il quadro e ha restituito la seguente descrizione:

L'immagine che hai caricato è una riproduzione del famoso dipinto "La notte stellata" di Vincent van Gogh. Questo capolavoro è stato creato nel 1889 e rappresenta una vista dalla finestra dell'ospedale psichiatrico di Saint-Rémy-de-Provence, dove van Gogh si trovava ricoverato. Il dipinto è noto per il suo cielo turbolento e onirico, ricco di vortici di stelle scintillanti e una luna crescente luminosa. Le pennellate sono spesse e dinamiche, creando un movimento che

sembra animare il cielo notturno. Il cielo è dominato da tonalità di blu e giallo, con tocchi di bianco per le stelle e la luna.

Nella parte inferiore del quadro, il paesaggio è più calmo e ordinato. Ci sono colline ondulate, una chiesa con un campanile alto e sottile che si staglia contro il cielo, e case basse con finestre che sembrano piccoli quadri di luce nella notte. Un cipresso alto e sinuoso, tipico del paesaggio meridionale francese, si erge in primo piano, fungendo da ponte tra la terra e il cielo.

I colori della terra sono più muti rispetto al cielo, con varie tonalità di verde, blu e marrone che danno un senso di quiete in contrasto con il tumulto sopra. La composizione del dipinto guida l'occhio a muoversi attraverso la scena, dal cipresso al cielo dinamico e poi verso il paese addormentato.⁴

Chiedendo al sistema di fornire una descrizione del quadro per bambine o bambini di otto-nove anni di età, il risultato è invece il seguente:

Questo quadro sembra una festa nel cielo ed è stato dipinto molto tempo fa da un pittore che si chiamava Vincent van Gogh, che amava usare molti colori e fare le pennellate tutte curve e mosse.

Il quadro si chiama "La notte stellata" e se lo guardi, puoi vedere un sacco di stelline che girano e si rincorrono nel cielo. Il cielo sembra dipinto con onde di colore blu e giallo, e se ti concentri, puoi quasi sentire la musica delle stelle.

Sotto questo cielo che sembra muoversi, c'è un villaggio che dorme sotto un grande piumone scuro. Le casette hanno le lucine spente perché è notte e tutti stanno sognando. C'è anche una chiesetta con un campanile che sembra toccare il cielo, quasi come se volesse raggiungere la luna.

E poi c'è un albero grandissimo, che assomiglia a una gigantesca piuma verde, e sembra come se stesse facendo il solletico alle stelle. Van Gogh ha dipinto questo albero per collegare la terra con il cielo, facendoci capire che tutto è connesso.

Quando guardi questo quadro, puoi immaginare di essere parte di questo mondo magico che van Gogh ha creato. E anche se siamo in un museo, se chiudi gli occhi, puoi sognare di essere lì, in mezzo a quella festa nel cielo.⁵

Ecco invece la risposta alla richiesta di descrivere il quadro dal punto di vista più strettamente pittorico:

Dal punto di vista pittorico, "La notte stellata" di Vincent van Gogh è un capolavoro della tecnica espressionista, dove l'uso del colore e delle pennellate è tanto emotivo quanto innovativo. Van Gogh utilizza la tecnica a impasto, applicando il colore in strati densi e spessi, che aggiunge una ricca tridimensionalità alla tela.

Le pennellate sono vigorose e visibili, creando un ritmo pulsante su tutta la superficie. L'artista esegue mosse circolari per rappresentare il cielo, creando vortici che catturano l'occhio e lo guidano attraverso la composizione. Queste pennellate non solo definiscono la forma e il movimento del cielo, ma anche la luminosità delle stelle e della luna.

Van Gogh sceglie una palette di colori che è allo stesso tempo realistica e fantastica. Usa tonalità di blu e ciano per il cielo notturno, intervallate da tocchi di giallo, bianco e arancio per

4. La risposta è stata data da ChatGPT 4 V in data 28 dicembre 2023. Prompt: "Forniscimi una descrizione dettagliata di questa immagine".

5. La risposta è stata data da ChatGPT 4 in data 28 dicembre 2023. Prompt: "Forniscimi ora una descrizione dell'immagine, adatta a bambine e bambini di 8-9 anni".



rappresentare la luminosità delle stelle e della luna. Il contrasto tra il caldo giallo delle stelle e il freddo blu del cielo crea un vibrante effetto di luce che sembra emanare dall'interno del quadro.

Il paesaggio sotto è più terreno e rassicurante, con i blu e i verdi a creare un senso di calma stabilità, in netto contrasto con il tumulto del cielo. Il cipresso, che si staglia in primo piano, è reso con pennellate verticali e sinuose, enfatizzando la sua natura di collegamento tra terra e cielo. In questo dipinto, van Gogh non si preoccupa della precisione fotografica, ma piuttosto esprime la sua esperienza emotiva del paesaggio, usando la pittura come mezzo per trasmettere la sua tumultuosa ricerca interiore. La tecnica pittorica diventa un'estensione della sua psiche, facendo di "La notte stellata" un'opera profondamente personale e universale allo stesso tempo.⁶

Quelli forniti sono ovviamente solo esempi, e – attraverso la mediazione del prompt o di un contesto adeguato (per "contesto" in questo caso si intende un insieme di informazioni fornite al sistema dopo il suo addestramento, con la funzione di indirizzare l'output nel senso più adeguato rispetto agli obiettivi specifici che si vogliono raggiungere) – il sistema può migliorare e personalizzare ulteriormente la descrizione in base alle necessità.

Il quadro di van Gogh è molto noto, probabilmente fa parte del corpus di addestramento sulla cui base è stato costruito il VLM di ChatGPT 4V, e il sistema lo sa identificare; ma, ricordiamo, il corpus di addestramento non è usato come un database: gli output prodotti da ChatGPT non sono "copiati" dalle informazioni su cui è stato addestrato ma prodotti generativamente, un token alla volta, su base statistico-probabilistica. Lo possiamo verificare constatando che i risultati sono qualitativamente ottimi – e per certi versi ancor più impressionanti – se usiamo immagini meno note o del tutto sconosciute. Fabio Ciotti ha discusso questo tema in un thread assai interessante su Facebook, mostrando come il sistema descriva – cogliendone correttamente anche le sfumature emotive – un quadro dipinto da sua figlia e sicuramente non presente in rete⁷.

La possibilità di generare descrizioni multimodali dettagliate e personalizzabili rappresenta una vera e propria rivoluzione nel campo dell'accessibilità: alcuni casi sono già abbastanza noti⁸, molti altri vengono esplorati proprio in questi mesi. Fra gli effetti di questa rivoluzione, vi sarà probabilmente anche il ripensamento di alcune scelte che parevano consolidate. Ad esempio, nel campo della web accessibility e dell'editoria digitale un elemento essenziale è rappresentato dalle descrizioni testuali alternative fornite attraverso l'attributo "ALT" del marcatore "IMG". Tuttavia, questo attributo ha (almeno) due funzioni diverse: da un lato, le descrizioni testuali alternative aiutano nell'interpretazione delle immagini da parte di persone con disabilità visive; dall'altro, però, sono anche uno strumento utile ai motori di ricerca, sia per migliorare l'indicizzazione del testo sia per permettere di ricercare e ritrovare più efficacemente le immagini che vi compaiono. In altri termini, il testo fornito attraverso l'attributo "ALT" è sia

6. La risposta è stata data da ChatGPT 4 in data 28 dicembre 2023. Prompt: "Forniscimi una descrizione dell'immagine dal punto di vista della tecnica pittorica utilizzata".

7. Fabio Ciotti, post su Facebook del 14 ottobre 2023, raggiungibile all'indirizzo <https://www.facebook.com/Ciotti.Fabio/posts/pfbid0QkyB7QowRjzGhSMqWmfU3ooutoqfHBDHLW6HqhBca8m1Bn9vyJzGeEBkT8L3H83BI>. Per la descrizione del quadro di Klee a cui il post fa riferimento, cfr. <https://www.facebook.com/Ciotti.Fabio/posts/pfbid02Xm7xHumcbpHd5Pftzsd1suoEkawTU6GZpsntKqEireuob1HYauMRbT62TTGm3MWvI>.

8. Per fare solo un esempio, la app Be My Eyes, che usa la fotocamera dello smartphone per fornire a un utente non vedente la descrizione dell'ambiente circostante, sta passando dall'uso di volontari umani (dal 2015, più di 6 milioni di volontari avevano collaborato nel fornire descrizioni vocali di quanto inquadrato dallo smartphone delle persone che usavano la app) all'uso di ChatGPT 4, con notevoli vantaggi sia nella rapidità della descrizione, sia rispetto alla privacy della persona non vedente. Cfr. <https://www.bemyeyes.com/blog/introducing-be-my-eyes-virtual-volunteer>.

uno strumento di accessibilità, sia uno strumento per la search engine optimization (SEO). È facile capire che queste due differenti funzioni potrebbero suggerire l'uso di testi descrittivi almeno in parte diversi. D'altro canto, l'uso delle descrizioni testuali alternative non è affatto tanto diffuso quanto sarebbe auspicabile: nel maggio 2022, in occasione del Global Accessibility Awareness Day, Twitter (oggi X) ha fornito al riguardo un dato impressionante: solo lo 0,6% delle immagini usate nei Tweet è accompagnato da un "alt text"⁹. Questo significa che oltre il 99% dei tweet con immagini non è di fatto accessibile a una persona non vedente. La generazione di "alt text" usando sistemi di IA generativa, eventualmente anche a integrazione di contenuti già prodotti, può aiutare a capovolgere la situazione e rende anche possibile pensare all'uso di più descrizioni testuali, in funzione di scopi diversi. Il lavoro in questa direzione è appena iniziato, e nei primi sistemi image-to-speech doveva fare i conti con la difficoltà di far cogliere ai sistemi di IA generativa i contesti delle immagini da descrivere: la descrizione "uomo in camice bianco tiene fermo un gatto arancione su un tavolo bianco" è più dettagliata ma probabilmente meno informativa della più semplice "veterinario che visita un gatto"¹⁰; i sistemi nativamente multimodali dovrebbero permettere in questo campo risultati decisamente migliori.

Crono- e georeferenziazione

L'uso di strumenti digitali nella generazione di contenuti e la diffusione dei sistemi GPS ha determinato una vera e propria esplosione di contenuti crono- e georeferenziati: ogni immagine prodotta dal nostro smartphone, così come molta parte dei post su social network, sono oggi sia cronoreferenziati sia georeferenziati. Abbiamo dunque a che fare, in molte situazioni, con veri e propri "big data" spaziali o temporali, e spesso sia spaziali sia temporali, che pongono problemi specifici sia in termini di analisi sia in termini di visualizzazione.

Le applicazioni dell'intelligenza artificiale generativa a questo tipo di dati sono un campo su cui si è cominciato a lavorare di fatto solo da pochi mesi, ma il cui interesse potenziale è enorme: al di là delle ovvie applicazioni in sistemi come i navigatori per auto (migliorando le loro capacità di interazione con gli utenti attraverso il linguaggio naturale) o nella gestione, ad esempio, di un calendario di impegni, è tutto l'insieme delle situazioni in cui sono coinvolti dati spaziali o temporali che può essere coinvolto. Un esempio particolarmente rilevante nel campo delle digital humanities è rappresentato dalla dimensione storica e da quella geografica nell'uso del linguaggio. I dati di un LLM normalmente sono "piatti" rispetto a queste dimensioni, anche se il LLM include sicuramente nell'embedding la considerazione di alcune informazioni di questo tipo, ricavate dal contenuto dei testi stessi. Ovviamente è possibile costruire corpora limitati a testi di un particolare periodo storico o provenienti da una determinata area geografica, o chiedere al sistema, attraverso il prompt, di produrre un output calibrato su un periodo o un ambito geografico specifico, se il relativo corpus di addestramento include dati sufficienti a farlo. Ma includere esplicitamente e sistematicamente la dimensione geografiche e quella cronologica in un sistema nativamente multimodale ne migliorerebbe sicuramente le caratteristiche: un tema su cui stiamo riflettendo nell'ambito del gruppo

9. <https://twitter.com/XAIly/status/1527287962479869952>.

10. L'esempio è nel post *Be Careful When Using A.I. for Alternative Text* del blog del Bureau of Internet Accessibility, 26 ottobre 2023, <https://www.boia.org/blog/be-careful-when-using-ai-for-alternative-text>.

di ricerca dell'Università Roma Tre relativo allo spoke 3 del progetto PNRR "Changes" coordinato dall'Università La Sapienza di Roma¹¹.

Alcuni esempi di LLM allargati a dati spaziali sono già stati realizzati da ricercatori che lavorano nel campo della cosiddetta GeoAI, l'intelligenza artificiale con applicazioni alla geografia (Gengchen et al., 2023): un esempio è GeoLM, che usa Open Street Map come base di georeferenziazione. I pesi iniziali del modello sono quelli di BERT, uno dei LLM più noti, e sono poi modificati attraverso l'analisi dei dati di Open Street Map, incrociati con le informazioni spaziali contenute in Wikipedia/Wikidata. Nella sintesi del gruppo di ricercatori che stanno lavorando al progetto nella University of Southern California, «The training process of GEOLM aims to simultaneously learn the linguistic and geospatial context, aligning them in the same embedding space to obtain geospatially grounded language representations» (Zekun et al., 2023)

Un lavoro in parte analogo (per il momento, forse ancor meno sviluppato) riguarda i dati cronologici. La considerazione della dimensione temporale è importante in molte forme di ragionamento, e l'aspetto del "temporal reasoning" – orientato soprattutto alla spiegazione causale e alla predizione – è stato studiato ad esempio attraverso un'estensione specifica di LLaMA2, un LLM che essendo open source permette molta sperimentazione¹². Ma il temporal reasoning è solo un aspetto, per quanto rilevante, dei possibili usi di dati cronologici da parte degli LLM. Personalmente, non conosco al momento sistemi di intelligenza artificiale generativa multimodali che integrino nativamente sia dati geografici sia dati cronologici (anche se l'integrazione di questi dati è naturalmente già esplorata in una pluralità di applicazioni informatiche diverse dalla IA generativa in campi che vanno dai GIS alla ricerca storica, archeologica e linguistica); mi sembra tuttavia abbastanza facile prevedere che si tratti di una strada che sarà esplorata ben presto. Sistemi di questo tipo potrebbero – per fare solo qualche esempio – generare cartografia storica (o utilizzare cartografia storica esistente come fonte informativa), aiutare i ricercatori nell'esplorazione dell'evoluzione storica e nella distribuzione geografica dei linguaggi, ma anche nell'analisi, ad esempio, dei fenomeni migratori (umani e animali), nella costruzione di modelli relativi a processi fisici, biologici, chimici o geologici, in campo archeologico, negli studi di geografia sociale, fino ad arrivare – su scale geografiche e temporali assai più vaste – all'astronomia e alla cosmologia.

Naturalmente, non mancano i problemi e i rischi: ad esempio la difficoltà nel gestire i dati "sfumati" (non sempre disponiamo di dati di crono e georeferenziazione precisi), o il fatto che anche i set di dati geografici e cronologici su cui possiamo lavorare presentano assai spesso bias rilevanti (come la sovra-rappresentazione di certe aree geografiche, tipicamente l'Europa e il Nord America, a scapito di altre, o di certi periodi cronologici rispetto ad altri). Ma non è certo questa la sede per esplorare né questi problemi né le molte strade e possibilità sopra appena accennate: in questa sede, il mio obiettivo era solo quello, assai più limitato, di segnalare l'interesse dell'intelligenza artificiale generativa multimodale anche nel campo della ricerca umanistica, e raccomandare attenzione per un campo d'indagine recentissimo ma con la capacità di modificare radicalmente molti dei nostri metodi di lavoro e di indagine.

11. Per una presentazione del progetto cfr. <https://sites.google.com/uniroma1.it/changes/> e, per quanto riguarda la partecipazione del mio ateneo, <https://www.uniroma3.it/ricerca/progetti-pnrr/changes-cultural-heritage-active-innovation-for-sustainable-society/>.

12. Il modello – battezzato TimeLLaMA – è discusso in (Chenhan et al., 2023). Sul tema del temporal reasoning si veda anche (Yifan et al., 2023).

Bibliografia

Chenhan, Yuan et al. (2023), Back to the Future: Towards Explainable Temporal Reasoning with Large Language Models. *arXiv:2310.01074*, <https://arxiv.org/abs/2310.01074>.

Girdhar, Rohit et al. (2023). Imagebind: One embedding space to bind them all. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15180-15190, <https://arxiv.org/abs/2305.05665v2>.

Mack, Kelly et al. (2021). What Do We Mean by "Accessibility Research"? A Literature Survey of Accessibility Papers in CHI and ASSETS from 1994 to 2019, *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, Association for Computing Machinery, Article 371, pp. 1-18, <https://doi.org/10.1145/3411764.3444542>.

Gengchen Mai et al. (2023). On the opportunities and challenges of foundation models for geospatial artificial intelligence, *arXiv:2304.06798*, <https://arxiv.org/abs/2304.06798>.

O'Connor, Ryan (2022, June 23). How Imagen Actually Works, *AssemblyAI*, <https://www.assemblyai.com/blog/how-imagen-actually-works/>.

Roncaglia, Gino (2023), *L'architetto e l'oracolo. Forme digitali del sapere da Wikipedia a ChatGPT*, Laterza.

Saharia, Chitwan et al. (2022), Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, in *Advances in Neural Information Processing Systems* 35, 36479-36494, <https://arxiv.org/abs/2205.11487>.

Yifan, Wei et al. (2023), Menatqa: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. *arXiv:2310.05157*, <https://arxiv.org/abs/2310.05157>.

Zekun, Li et al. (2023). GeoLM: Empowering Language Models for Geospatially Grounded Language Understanding. *arXiv:2310.14478*, <https://arxiv.org/pdf/2310.14478.pdf>.



