17 Gennaio

Deep Learning: pappagalli stocastici e cavalieri inesistenti

il più grande pericolo dell'Intelligenza Artificiale
è che le persone concludano troppo presto di averla compresa.
Eliezer Yudkowsky

Le nuove *macchine pensanti* dotate degli algoritmi di ***deep learning*** sono in grado di ascoltare, parlare e guardare, ma si limitano a stare nel mondo delle funzioni, cioè quello dei compiti da eseguire: l'***AI*** ascolta ma non capisce, parla ma non sa, guarda ma non vede. Noi invece pensiamo, percepiamo, facciamo esperienza del nostro vissuto

Basta che accendiate il vostro cellulare e lui sarà in grado di riconoscervi, ovvero di guardare il vostro viso e sbloccare le sue funzioni soltanto se siete veramente voi. Eppure, nello svolgimento di questa funzione, *il cellulare vi ha veramente visto?*

Il cellulare risolve una funzione, il vostro riconoscimento, ma non è una persona, un soggetto cosciente, che vi vede. Il cellulare vi riconosce, ma non vi vede.

In modo analogo, tutti i recenti *bot* e *algoritmi*, che promettono di riconoscere la lingua naturale e la nostra voce, svolgono il loro compito in modo sempre più simili agli esseri umani.
Casi famosi come **Replika**

Una specie di "*Her*" dei poveri produce testi che farebbero invidia a molti giovani laureati, ma non conoscono il significato delle parole che elaborano. Questi *bot* elaborano le parole in modo da produrre risposte che assomigliano in tutto e per tutto a quelle di un essere umano, ma non ne conoscono il significato.

REPLIKA vive ed opera esclusivamente nel *mondo delle funzioni* ascolta *ma non capisce*, parla *ma non sa,* guarda *ma non vede.*

Frank Jackson professore emerito di filosofia presso ***l'Australian National University***

è il maggior esponente del ***Fisicalismo*** secondo cui l'intero contenuto del mondo è solo materia, atomi e vuoto. Le entità immateriali , come la mente, sono semplicemente epifenomeni , un sottoprodotto dell'attività cerebrale, senza alcun effetto causale sul mondo materiale.

Ma cos'è allora la conoscenza umana ?
La conoscenza non influenza il nostro mondo ogni giorno?

L'idea è che l'universo sia costituito interamente dai tipi di entità descritte dalla fisica. Il suo primo esempio dell'argomento della conoscenza è *uno scienziato di nome Mary* che vive in una stanza in bianco e nero e vede il mondo esterno attraverso un monitor televisivo in bianco e nero.

Mary è una brillante scienziata che, per qualsiasi motivo, è costretta a indagare sul mondo da una stanza in bianco e nero attraverso un monitor televisivo in bianco e nero. È specializzata nella neurofisiologia della visione e acquisisce, supponiamo, tutte le informazioni fisiche che c'è da ottenere su ciò che accade quando vediamo pomodori maturi, o il cielo, e usa termini come "rosso", "blu" e così via. su. Scopre, ad esempio, esattamente quali combinazioni di lunghezze d'onda dal cielo stimolano la retina, ed esattamente come questa produce attraverso il sistema nervoso centrale la contrazione delle corde vocali e l'espulsione dell'aria dai polmoni che si traduce nella pronuncia della frase 'Il cielo è blu'...

Cosa accadrà quando *Mary* uscirà dalla sua stanza in bianco e nero o riceverà un monitor televisivo a colori? Imparerà qualcosa o no? Sembra ovvio che imparerà qualcosa sul mondo e sulla nostra esperienza visiva di esso.

Ma poi è inevitabile che la sua precedente conoscenza fosse incompleta anche se possedeva tutte le informazioni fisiche.

In altre parole Immaginate di non avere mai visto alcun colore: siete un non-vedente totale. Tuttavia, siete sempre vissuti in una comunità di vedenti e sapete che gli oggetti sono dotati di una proprietà misteriosa che tutti chiamano colore.

Conoscete anche le relazioni tra i colori (per esempio sapete che il rosso e il verde sono molto diversi tra loro, mentre il rosso e il rosa sono in qualche modo vicini).

Quindi siete in grado di parlare in modo appropriato dei colori, ma non sapete che cosa sia l'esperienza del colore.

Ecco, oggi l'intelligenza artificiale è un po' come quel cieco congenito che parla di colori. Solo che non è solo cieca dalla nascita, ma è anche sorda, priva di sensazioni tattili, non ha mai sentito un sapore o percepito un odore.

Questa è la situazione in cui si trova l'intelligenza artificiale al momento. Grazie ai nostri cellulari e terminali, riceve miliardi di informazioni ogni secondo, ma questa informazione è priva di significato, come la parola "rosso" non ha alcun significato per un non vedente congenito.

Ovviamente, questa condizione dell'intelligenza non è solo un problema filosofico, ma ha grosse conseguenze pratiche.

L'elaborazione del linguaggio naturale è alla base dell'attività sia di Google che di Facebook (e da poco anche di Microsoft che ha acquistato OpenAI).

Due ricercatori di Google **Timmit Gebru** **e** **Margaret Mitchell**

sono state licenziate perché hanno pubblicato un articolo dove mettevano in discussione i rischi di creare intelligenze artificiali senza una reale comprensione dell'informazione che elaborano ovvero:
come possiamo essere sicuri che le macchine non prendano decisioni eticamente pericolose visto che non fanno esperienza del mondo, di noi e delle nostre emozioni?
Come prendere decisioni sugli esseri umani se non si è umani?

L'articolo di Gebru e Mitchell ha un titolo suggestivo: ***I pericoli dei pappagalli stocastici***

I ***pappagalli stocastici*** non sono altro che le *intelligenze artificiali* che producono linguaggio naturale senza conoscerne il significato (come i pappagalli appunti, o forse anche peggio di loro).

Come il *cavaliere inesistente di Calvino*, che era una armatura infallibile che sconfiggeva ogni altro essere umano ma dentro di sé non aveva altro che il vuoto, così l'intelligenza artificiale di oggi è un guscio di funzioni, sempre più sofisticate, che non hanno niente dentro.

Il **cavaliere inesistente** appare come un inquietante simbolo dell'uomo moderno, che è talmente in crisi da sembrare privo di identità, quasi **inesistente**. L'uomo del nostro tempo appare infatti incerto, smarrito, privo di orientamenti e sicurezze.

C'è qualcosa in lui di vuoto, come è vuota la bianca armatura di Agilulfo

Quando conversiamo con le *macchine pensanti* che il mercato dell'informatica ci propina abbiamo l'impressione di parlare con una persona in carne e ossa, come se, dietro i suoi messaggi, ci fosse una persona che "capisce". Ma sappiamo che non c'è alcuna persona. C'è solo un modello statistico del linguaggio naturale degli esseri umani, abbastanza complesso da imitare come un pappagallo le risposte di una persona vera.

Si tratta sicuramente di enunciati ben formati e grammaticalmente corretti. È anche possibile che la loro conversazione, ai nostri occhi, abbia un senso. Ma ecco il punto. I messaggi che si sono scambiati, ai loro occhi, non hanno alcun senso, perché loro, le due intelligenze artificiali, non hanno occhi con i quali capire il senso delle parole.

Allo stesso modo, altre intelligenze artificiali, presentate come se fossero in grado di immaginare forme e figure, non vedono quello che producono.

WALL-E è uno spin off clonato da **GPT-3**, che è in grado di generare immagini nuove a partire da parole e frasi. l sistema è una versione costruita da 12 miliardi di parametri di **GPT-3:**

Ha una serie diversificata di capacità, tra cui la creazione di versioni antropomorfizzate di animali e oggetti, la combinazione di concetti non correlati in modi plausibili, il rendering del testo e l'applicazione di trasformazioni a immagini esistenti.

Se gli chiediamo di mostrarci, un **cetriolo verde** con la gonna e le ruote, ecco che compare un disegno a colori sgargianti proprio di un simpatico cetriolo dotato di ruote e gonnellino, anzi molte versioni della stessa idea.
E così via, perché **Wall-E** è un modello che ha immagazzinato un modello di milioni di parametri che collega immagini e parole prese dai Big Data della rete. Ma quello che non dobbiamo dimenticare è che, mentre noi vediamo il cetriolo con le ruote, **Wall-E** non vede niente.

Wall-E è "solo" un insieme di miliardi di parametri che attivano altri parametri.

Ciò gli consente di creare immagini completamente nuove esplorando la struttura di un suggerimento, inclusi oggetti fantastici che combinano idee non correlate che non sono mai state alimentate durante l'addestramento.

E lo stesso vale per tante altre intelligenze artificiali che oggi vengono presentate come se fossero in grado di riprodurre altrettanti processi della mente umana, dal sogno alla creatività.

La differenza con gli esseri umani è che noi facciamo esperienza del senso delle nostre creazioni, l'intelligenza artificiale, per ora no.

Credere, come fanno molti utenti (e anche molti esperti), che per capire qualcosa sia sufficiente formulare frasi corrette e riconoscere volti e suoni, è un errore molto pericoloso che mette sullo stesso piano le parole di un non vedente e le parole di un vedente per quanto riguarda colori e sfumature. Non sono la stessa cosa.

Il secondo accede al mondo del significato, mentre l'intelligenza artificiale, almeno per ora, si limita al mondo delle funzioni. Noi funzioniamo (a volte nemmeno bene), ma soprattutto noi siamo e noi percepiamo. L'intelligenza artificiale ancora no.
L'intelligenza artificiale ascolta *ma non capisce*, parla *ma non sa*, guarda *ma non vede.*

Allegato:

A proposito di Oper AI

Nel giugno 2020, una nuova e potente intelligenza artificiale (AI) ha iniziato ad abbagliare i tecnologi nella Silicon Valley. Chiamato **GPT-3 (Generative Pretrained Transformer** 3) e creato dalla società di ricerca **OpenAI** di San Francisco, in California, è stato l'ultimo e il più potente di una serie di "modelli linguistici di grandi dimensioni": IA che generano flussi di testo fluenti dopo aver assorbito miliardi di parole da libri, articoli e siti web. **GPT-3** era stato addestrato su circa 200 miliardi di parole, a un costo stimato di decine di milioni di dollari.

Tutto quello che devi fare è scrivere un prompt e aggiungerà il testo che pensa possa seguire in modo plausibile. Ce l'ho fatta a scrivere canzoni, racconti, comunicati stampa, tablature per chitarra, interviste, saggi, manuali tecnici. È divertente e spaventoso.

Il team di **OpenAI** ha riferito che **GPT-3** era così buono che le persone trovavano difficile distinguere le sue notizie dalla prosa scritta dagli umani . Potrebbe anche rispondere a domande trivia, correggere la grammatica, risolvere problemi di matematica e persino generare codice per computer se gli utenti gli dicessero di eseguire un'attività di programmazione. Anche altre IA potrebbero fare queste cose, ma solo dopo essere state specificamente addestrate per ogni lavoro.

I grandi modelli linguistici sono già proposte commerciali. Google li utilizza per migliorare i suoi risultati di ricerca e la traduzione linguistica; Facebook, Microsoft e Nvidia sono tra le altre aziende tecnologiche che le realizzano.

OpenAI mantiene segreto il codice di **GPT-3** e offre l'accesso ad esso come servizio commerciale. (OpenAI è legalmente una società senza scopo di lucro, ma nel 2019 ha creato una sottoentità a scopo di lucro chiamata OpenAI LP e ha collaborato con Microsoft, che ha investito un miliardo di dollari nell'azienda.)

Gli sviluppatori stanno ora testando la capacità di **GPT-3** di riassumere documenti legali, suggerire risposte alle richieste del servizio clienti, proporre codice informatico, eseguire giochi di ruolo basati su testo o persino identificare individui a rischio in una comunità di supporto tra pari etichettando i post come richieste di aiuto.

Nonostante la sua versatilità e scalabilità, **GPT-3** non ha superato i problemi che hanno afflitto altri programmi creati per generare testo. Tuttavia presenta gravi debolezze e talvolta commette errori molto sciocchi, Funziona osservando le relazioni statistiche tra le parole e le frasi che legge, ma non ne comprende il significato.

Di conseguenza, proprio come i ***chatbot*** più piccoli, può vomitare incitamento all'odio e generare stereotipi razzisti e sessisti, se richiesto, riflettendo fedelmente le associazioni nei suoi dati di addestramento. A volte darà risposte prive di senso ("Una matita è più pesante di un tostapane") o risposte decisamente pericolose.

I modelli linguistici metafoiricamente simili a “**pappagalli stocastici**" fanno eco a ciò che sentono, remixati dalla casualità.

I ricercatori hanno idee su come affrontare pregiudizi potenzialmente dannosi nei modelli linguistici, ma instillare nei modelli buon senso, ragionamento causale o giudizio morale, come molti vorrebbero fare, è ancora un'enorme sfida di ricerca. Quello che abbiamo oggi, è essenzialmente **una loquace bocca senza cervello".**

A proposito dei pappagalli:un pensiero di **Friedrich Engels:**

L'uccello che ha la voce più sgradevole, il pappagallo, è quello che parla meglio. Non si dica che egli non comprende quello che dice. Senza dubbio, ripeterà ciarliero tutto il suo patrimonio di parole per ore ed ore, per il semplice gusto di parlare e per il fatto che sta in compagnia di uomini. Ma entro i limiti delle cose che comprende può imparare anche a capire quello che dice. Si insegnino a un pappagallo delle ingiurie, in modo che si faccia una idea del loro significato (è uno dei sommi piaceri dei marinai che tornano veleggiando dai paesi tropicali); lo si stuzzichi, e si vedrà ben presto che sa far uso dei suoi insulti non meno appropriatamente di un'erbivendola berlinese. Lo stesso si dica per quel che riguarda la richiesta di leccornie.

PostScriptum

consiglio un bel gioco: se avete un pappagallo, addestratelo a ripetere

"Qualcuno mi aiuti, sono stato trasformato in un pappagallo!".

Un anno fa… Baedeker/Replay del 17 gennaio 2022

La pandemia è finita! siamo in piena infodemia alimentata dalla rete invisibile delle "fake-paper factories"

Nel 2020, nonostante la pandemia di COVID, gli scienziati hanno prodotto e sottoposto a revisione paritaria 6 milioni di pubblicazioni, un aumento del 10% rispetto al 2019. A prima vista questo grande numero sembra una buona cosa, un indicatore positivo del "progresso della scienza" e della diffusione della conoscenza. Purtroppo tra questi milioni di articoli, tuttavia, ci sono migliaia di articoli inventati, di sana pianta prodotti da molti di accademici e ricercatori in carriera che si sentono obbligati a produrre in base al principio del "pubblish or perisch" pubblicare o morire, anche se ciò li costringe a barare. Molti "scienziati" disonesti stanno usando questi strumenti per copiare il testo da varie fonti autentiche, parafrasarlo e incollare il risultato "torturato" nei propri fogli.

Come facciamo a saperlo? Una prova è che si possono riprodurre la maggior parte delle frasi torturate inserendo termini consolidati in un software di parafrasi. La presenza di questa letteratura spazzatura pseudoscientifica mina la fiducia dei cittadini negli scienziati e nella scienza, soprattutto quando viene trascinata e strumentalizzata nei dibattiti di politica pubblica. Ma in una questa escalation della frode accademica amplificata dalla pandemia i moderni plagiatori stanno facendo uso di software e forse anche di tecnologie di intelligenza artificiale emergenti per redigere impuniti articoli su esperimenti mai eseguiti e, impunemente se la stanno cavando senza evidenti conseguenze. La crescita delle pubblicazioni di ricerca, combinata con la disponibilità di nuove tecnologie digitali, suggerisce che la frode mediata da computer nella pubblicazione scientifica potrebbe aumentare vertiginosamente . Insieme ad una maggioranza di ricercatori di qualità una minoranza di "scienziati" disonesti stanno usando questi strumenti per copiare il testo da varie fonti autentiche, parafrasarlo e incollare il risultato "torturato" nei propri fogli. Una frode come questa non colpisce solo i ricercatori e le pubblicazioni coinvolte, ma può complicare la collaborazione scientifica e rallentare il ritmo e le acquisizioni della ricerca, in particolare inquina le conoscenze faticosamente ottenute acquisite deformandola realtà scientifica come sta avvenendo in maniera sistematica e continua in questa pandemia ormai divenuta una infodemia. La presenza di questa letteratura pseudoscientifica spazzatura mina anche la fiducia dei cittadini negli scienziati e nella scienza, soprattutto quando viene utilizzata nei dibattiti di politica pubblica, e condiziona pesantemente le scelte dei decisori, la nostra salute e mette in pericolo la vita delle future generazioni Uno dei modi per individuare ricerche fraudolente è di individuare all'interno di un lavoro utilizzando uno dei tanti software anti plagio disponibili con l'applicazione del Problematic Paper Screener, una applicazione capace di segnalare lavori sospetti attraverso la ricerca delle le "frasi torturate"

Cosa sono le frasi torturate? Una frase torturata è un concetto scientifico consolidato parafrasato in una sequenza di parole senza senso. L "intelligenza artificiale" diventa "coscienza contraffatta". . "Segnale al rumore" diventa "Bandiera a clamore", "Cancro al seno" diventa "pericolo al seno". Tutti i relatori di tesi di laurea che utilizzano soft-ware anti-plagio per la revisione dei draft dei laureandi e/o di specializzandi

trovano una abbondanza di queste frasi per mascherare il plagio (taglia e incolla) espressione di una creatività fantasiosa. Frasi recentemente torturate sono emerse nella letteratura scientifica sulla pandemia di COVID19. Un articolo pubblicato a luglio 2020, da quando è stato ritirato, è stato citato 52 volte a partire da questo mese, nonostante riportasse la frase "sindrome respiratoria estremamente intensa (SARS)", che è chiaramente un riferimento alla sindrome respiratoria acuta grave, la malattia causata dal coronavirus SARS-CoV-1. Altri documenti contenevano la stessa “frase torturata”. Una volta scoperti e identificati documenti fraudolenti, farli ritirare non è un compito facile. Gli editori e gli editori che sono membri del Comitato per l'etica della pubblicazione devono seguire linee guida complesse prestabilite quando trovano documenti problematici.

Ma il processo ha una scappatoia. Gli editori che "indagano il problema" per mesi o anni perché dovrebbero aspettare risposte e spiegazioni dagli autori per un periodo di tempo indefinito. L' intelligenza artificiale aiuterà a intercettare i documenti privi di significato, errati o contenenti frasi torturate . Ma questo sarà efficace solo a breve e medio termine. Gli strumenti di controllo dell'intelligenza artificiale potrebbero paradossalmente finire per provocare….

(per continuare vai all’originale)

La salute dipende più dalle precauzioni che dalle medicine.
Jacques-Bénigne Bossuet