Modelli AI GPT 3 e GPT4 a confronto

Introduzione:

L'Intelligenza Artificiale (AI) si sta rapidamente affermando come una importante innovazione tecnologica in numerosi campi. Tra le numerose applicazioni di questa tecnologia, la generazione del linguaggio naturale sta giocando un importante ruolo. In questo contesto, il sistema di generazione del testo GPT (Generative Pretrained Transformer) rappresenta un notevole passo avanti. Tale sistema fu inizialmente sviluppato dall'azienda OpenAI, e poi riprodotto da altre società con fini analoghi.

Il sistema GPT, basato sulle reti neurali, si preoccupa di apprendere il contenuto del linguaggio da grandi quantità di testo, utilizzando diversi metodi di apprendimento automatico. Gli fu attribuito il nome di GPT-1 nel 2018, gli successe poi GPT-2 nel 2019 e nel 2020 GPT-3. Quest'ultimo è quello che ha ricevuto maggiore attenzione anche dai media, poiché è stato in grado di dimostrare una grande capacità e versatilità nella produzione di testo, paragonabile a quello umano.

Per superare i limiti di GPT3, OpenAI ha annunciato di avere messo in cantiere un nuovo progetto ancora più avanzato: il sistema GPT4. In questa relazione, ci concentreremo dunque sulla differenza tecnica e pratica fra questi ultimi due sistemi.

Cosa è una GPT:

La GPT, letteralmente Generative Pretrained Transformer, è un modello di rete neurale artificiale sviluppato per la generazione del linguaggio naturale, ampliando la famiglia di modelli di deep neural network. Tale sistema, su base di apprendimento automatico, è in grado di prevedere e generare il testo sequenziale a seguito di un input.

L'architettura di una GPT di base prevede un codificatore, una decodificatore e un separatore, permettendo all'AI di prevedere e generare linguaggio naturale non solo basandosi sui dati in entrata, ma anche sulla conoscenza acquisita dal training. Il sistema era stato in origine progettato per migliorare le prestazioni dell'AI in molte applicazioni del linguaggio naturale, come la traduzione automatica, la sintesi del discorso e la raccolta automatica di informazioni.

GPT-1:

Il modello GPT-1, creato come prototipo dello sviluppo, aveva 117 milioni di parametri. Nonostante fosse in grado di produrre testi coerenti, la qualità dei testi non era ancora al livello delle aspettative.

GPT-2:

GPT-2 è stata la seconda iterazione di questo modello di AI, lanciato nel 2019. Questo sistema ha dimostrato notevoli capacità: seppur con solo 1,5 miliardi di parametri, è stato in grado di scrivere testi coerenti, generare sequenze di testo sintatticamente e semanticamente corrette e distintive dalle produzioni umane. Ciò ha suscitato notevole attenzione anche al di fuori del mondo dell'informatica. L'AI ha mostrato di avere una notevole capacità di generare molti tipi di testo, anche molto lungo e complesso, ma allo stesso tempo poteva (in ambito di pura associazione statistica di parole) produrre affermazioni false e radicarsi in pregiudizi di genere o di razzismo.

GPT-3:

Nel 2020, sono state presentate le potenzialità di GPT-3, risultando essere un ulteriore grande passo avanti, dotato di circa 175 miliardi di parametri. Ciò ha permesso alla AI di dimostrare una grande

capacità di comprensione e produzione di testo accurato e generale. GPT-3 è stato in grado di generare testo di qualità molto simile ai testi scritti da un essere umano.
Si è diffuso quindi un grande entusiasmo per il potenziale dell'utilizzo di questa AI in numerosi settori, come ad esempio la traduzione automatica, la produzione automatizzata di testi commerciali, la creazione di storie e le antologie digitali.

GPT-4:

Nel 2021, la società OpenAI ha annunciato il nuovo sviluppo nell'ambito delle generative pre-trained transformers: il sistema GPT-4. Dotato di un trilione di parametri, il sistema avrà la capacità di elaborare una grande quantità di dati a una velocità senza precedenti.
GPT-4 presenterà molte innovazioni rispetto alle generazioni precedenti, come la comprensione avanzata contestuale e la capacità di elaborazione in multitasking. Inoltre, sarà in grado di generare testo di alta qualità e personalizzati per una vasta gamma di contesti, compresi report scientifici e documenti accademici.

Contesto generale:

L'AI è un campo in rapida evoluzione, che non smette di presentare significativi progressi. L'utilizzo di questi dispositivi copre un'ampia gamma di settori, tra cui la salute, l'istruzione, il settore bancario, il marketing e l'industria automobilistica. Molte di queste applicazioni sono già in uso, e sono in continuo miglioramento. Tra le ultime innovazioni in questo settore, GPT3 si è dimostrato un'importantissima evoluzione nel campo della generazione del linguaggio naturale.
Il sistema GPT3, come detto prima, è in grado di elaborare il linguaggio umano con una dettagliata comprensione semantica delle frasi. Grazie al suo ampio vocabolario di parole e alla sua espressività, è possibile utilizzarlo per rispondere a domande, tradurre i linguaggi, scrivere parole e frasi in contesti accademici e commerciali.
Tuttavia, nonostante gli ottimi risultati e l'entusiasmo suscitati dal sistema GPT3, secondo un feedback mondiale non siamo ancora a livelli di generazione di testo propriamente "umano", da qui la necessità di evolverlo in GPT4. Ma quali sono le differenze tecniche e pratiche tra questi due sistemi? In seguito, ci concentreremo sulla comparazione di questi due importanti sviluppi dell'AI.

Differenze tecniche:

Attenzione: in questa sezione, ci concentreremo sui dettagli tecnici di GPT3 e GPT4. Per il pubblico non esperto in informatica, questi termini possono risultare complessi, ma sono molto importanti per capire le differenze tra questi due progetti.
Il sistema GPT3 utilizza una piattaforma di elaborazione parallela su GPU, grazie alla quale è possibile elaborare fino a 175 miliardi di parametri. La quantità di parametri è il fattore-base per misurare la potenza di elaborazione dei sistemi di AI basati sulle reti neurali. In questo ambito, GPT3 rappresenta un miglioramento rispetto alla precedente iterazione, ovvero il sistema GPT2, che ne conteneva solo 1,5 miliardi.
La quarta versione del sistema, chiamata GPT4, da un punto di vista di parametri ne possiederà ancora di più rispetto alla versione GPT3: infatti, il sistema GPT4 utilizzerà una piattaforma di elaborazione su supercomputer, in grado di elaborare fino a un trilione di parametri. Questi progressi rappresentano un

considerevole passo avanti rispetto alle generazioni precedenti, e si prevede che tale sviluppo possa avere importanti applicazioni a livello commerciale ed accademico.
Ma le innovazioni non si fermano qui: parallelamente, GPT4 presenterà anche innovazioni nell'architettura informatica, con l'obiettivo di migliorare l'esplicazione e la risposta ai comandi. In particolare, si passerebbe da un'architettura lineare singola, ovvero in cui i "token" (le più piccole unità testuali con significato specifico) vengono immessi in un'unica sequenza lineare, ad un'architettura multipla, migliorando la capacità di elaborazione.
Un ulteriore ambito di progresso consiste nella capacità di comprensione contestuale. Nel nuovo sistema, infatti, la comprensione sintattica e semantica sarebbero integrate con tutti i contenuti precedenti e successivi, fornendo una comprensione più completa e profonda del contesto in cui la frase viene utilizzata. Ciò permetterebbe al sistema di produrre risposte più complete e coerenti.
Infine, GPT4 sarebbe anche in grado di elaborare multitasking, ovvero di svolgere più attività contemporaneamente, come la risposta a diverse domande, traduzioni simultanee di più lingue, e la generazione di testi in più lingue e stili.

Differenze pratiche:
Mentre le differenze tecniche tra GPT3 e GPT4 sono rilevanti a livello di programmazione e elaborazione dei dati, le differenze pratiche tra questi due sistemi possono influenzare l'esperienza dell'utente e le possibilità offerte dal sistema di AI.
Uno dei principali vantaggi di GPT3 è la natura molto intuitiva dell'interfaccia, che permette all'utente di fare richieste e di ottenere risposte rapide ed accurate. Grazie alla grande capacità di elaborazione dei dati di cui è dotato il sistema, è possibile ottenere da GPT3 una risposta completa e dettagliata anche ad una domanda vagamente formulata.
Tuttavia, la capacità di generare frasi complete da sola non rappresenta un'indicazione del grado di accuratezza o utilità di un sistema: può infatti risultare limitato in quanto non può totalmente comprendere il contesto della conversazione, non può fare domande di follow-up o riconoscere l'ironia o il sarcasmo. In altre parole, è importante guardare oltre la sola capacità di generare frasi complete per valutare l'efficacia delle risposte Ad esempio, potrebbe essere difficile per il sistema produrre una risposta adeguata ad una domanda che implica un'approfondita conoscenza specifica e non generale su un particolare argomento accademico. In questo ambito, GPT3 potrebbe rischiare di produrre risposte che, sebbene tecnicamente e strutturalmente corrette, non rispondono al quesito in modo efficace.
Allo stesso tempo, l'efficacia di una risposta dipende strettamente dal contesto in cui viene applicato. Il suo utilizzo in ambito commerciale potrebbe essere molto utile per la creazione di nuovi contenuti di marketing, ma la sua utilità risulta limitata per attività più complesse, che vadano oltre la pura nozionistica statistica. Difatti, dal punto di vista dell'esperienza d'uso, l'utilizzo di GPT3 a fini personali potrebbe avere qualche limite dovuto alla presenza di risposte predefinite e all'eccessiva attenzione sulle parole chiave di ricerca. Questo fenomeno causa un inevitabile blocco creativo nella comprensione e interpretazione dei risultati, e limiterebbe l'esplorazione delle possibilità offerte dal sistema di AI.
Tuttavia, con GPT4, queste restrizioni dovrebbero essere superate grazie alla sua capacità di generare contenuti che rispondano più efficacemente ai quesiti e alla comprensione degli schemi linguistici. L'utilizzo di questo sistema potrebbe portare a significativi miglioramenti nel campo della ricerca e nell'elaborazione di dati complessi.

Conclusioni:

In sintesi, possiamo affermare che il sistema GPT3 è un notevole passo avanti in termini di capacità di generazione di linguaggio naturale, e che viene utilizzato in diversi settori. Tuttavia, nonostante la sua grande potenza, il sistema presenta alcune limitazioni che si riflettono sull'esperienza dell'utente e sulla capacità di produrre risposte personalizzate ed efficaci.

Con GPT4, OpenAI promette importanti innovazioni a livello di elaborazione dati, che dovrebbero rendere questo nuovo sistema ancor più adattabile e versatile. In particolare, la maggiore capacità di multitasking e la comprensione più completa del contesto linguistico potrebbe rappresentare una svolta fondamentale per AI, e aprire nuove vie di controllo e personalizzazione di questo dispositivo. Tuttavia, dal punto di vista pratico, è ancora difficile valutare l'efficacia delle nuove funzionalità, e si dovrà attendere la presentazione ufficiale di GPT4 per avere un quadro esaustivo delle sue effettive potenzialità e limitazioni.

Generativamente vostro… Mike Yoshi