IL FUTURO dell'INTELLIGENZA ARTIFICIALE

La recente letteratura tratta l'argomento delle previsioni di trasformazione della IA in campi quali il natural language processing, il knowledge graph, il machine learning e il deep learning. Ecco i principali punti:

- Fotografia computazionale. Elabora le foto di uno smartphone correggendone gli errori più grossolani. E' basata sugli algoritmi del machine learning
- Assistenti digitali. Sono basati sugli algoritmi di natural language processing, e interpretano i nostri comandi vocali
- Motori di ricerca per il Web. Come Google, dipendono dai knowledge graph, sorta di mappe relazionali alle informazioni disponibili su dati non strutturati di persone, luoghi ed avvenimenti. In futuro aumenteranno quando la IA entrerà in più processi produttivi ed organizzativi dell'industria. Tutto questo grazie ad un forte sviluppo della capacità di calcolo dei dati
- IA cambierà le nostre vite quando si focalizzeranno i due concetti di training e di inferenza negli algoritmi di deep learning
- Il deep learning è un processo al termine del quale un calcolatore, dopo aver imparato ad elaborare dei dati grazie agli umani (training) riesce a fornire risposte in autonomia e senza necessità di correzioni (inferenza) applicando il modello precedentemente imparato. Un caso tipico è il riconoscimento facciale
- Il training consiste quindi nell'imparare una nuova capacità partendo da dati esistenti e solitamente sotto la guida di un umano, mentre l'inferenza è la capacità di applicare il processo appreso a nuovi dati. Sia il training che l'inferenza presentano grosse esigenze di calcolo. In tal caso si potrebbe ricorrere al cloud, cioè ad una rete esterna ottimizzata per questo tipo di calcoli
- Assistenti vocali. Gli ultimi modelli integrano nello smartphone chip dotati di Neural Engine, un coprocessore dedicato agli algoritmi di IA. Anche la privacy risulterebbe migliorata
- Potenza richiesta. Non occorre potenza superiore a quella oggi disponibile, ma di tipo diverso, con processori per nuovi compiti. E' aumentata anche la corsa ai GHz, cioè alla maggiore frequenza possibile per smaltire in fretta la coda delle operazioni. L'operazione all'interno del chip dovrebbe essere parallela anziché sequenziale, cioè su più dati contemporaneamente. Per questo negli ultimi anni i risultati migliori sono stati ottenuti con i GPU, processori grafici che offrono parallelismo spinto. Il passo successivo è l'uso di processori ASIC (Application Specific Integrated Circuits) ed FPGA (Field Programmable Gate Array)
- I due problemi principali sono la difficoltà di scala, cioè adattare automaticamente le risorse disponibili alla più ampia mole di dati per soluzioni più complesse e i limiti di performance dell'hardware

- Altri problemi sono causati dallo smaltimento del calore dei computer, l'insufficienza di memoria e la lentezza dell'incamerare i dati
- Un'importante trasformazione dell'hardware in una nuova generazione è attesa per il 2024
- SUMMIT, uno dei computer più potenti al mondo, è dotato dell'hardware Power System AC922 di IBM, con un processore POWER9 ad alto parallelismo che riesce ad elaborare 5,6 volte più istanze rispetto ai concorrenti. Il Power System AC922 rappresenta la base anche per l'altro supercomputer Sierra del Lawrence Livermore National Laboratory, usato dalla National Nuclear Security Administration. Gran parte dei calcoli è svolta anche dalle GPU NVIDIA TESLA V 100 dotate di NVLink che permette a più GPU di scambiarsi dati ad altissima velocità. IBM ha creato anche un set di librerie software, dalle quali gli scienziati possono richiedere il numero di GPU necessario per risolvere complessi calcoli.

Commenti

Quanto sappiamo oggi della IA rappresenta solo una parte degli attesi sviluppi, che vengono previsti per il prossimo 2024. E' la traccia del cammino che accompagna l'integrazione tra umani e macchine.