L’importanza dei dati nell’IA

Nel panorama dell’IA una figura di spicco è sicuramente quella di Andrew Ng, pioniere nell’addestramento dei modelli di deep learning attraverso l’utilizzo della GPU, durante il suo corso di studi alla Stanford University, nonché cofondatore del progetto Google Brain. In un’intervista con IEEE Spectrum, Ng ha condiviso le sue prospettive, a parer suo, sul prossimo grande cambiamento nell’ambito dell’IA, esaltando il movimento chiamato "Data-centric AI”.

In questa relazione esamineremo le idee proposte da Andrew Ng e l’importanza di questo nuovo approccio nel contesto dell'evoluzione dell’intelligenza artificiale. Approfondiremo come la raccolta dei dati possa diventare un punto cruciale nello sviluppo di modelli basati su piccoli dati, per poter ottenere un’ottima efficienza, accuratezza e bias. Attraverso le dovute analisi, cercheremo di identificare il ruolo dei dati all’interno del mondo dell’IA.

Attraverso una serie di domande Andrew Ng presenta l’approccio Data-centric, i principali punti evidenziati nell’articolo sono:

- **Sviluppo dei modelli di Deep Learning**: La costruzione ed il miglioramento dei modelli di Deep Learning è stato un punto cruciale negli ultimi 15 anni, questo processo è tuttora fondamentale per alcuni tipi di problemi basati su dati di grandi dimensioni, bensì non sempre opportuno, per alcune circostanze è infatti necessario lavorare su dati di piccole dimensioni.
- **Foundation model**: Un foundation model è un modello addestrato su dati di grandi dimensioni, perfezionato su alcune applicazioni. Ng, prende come esempio il modello GPT-3.
- **Foundation model per i video**: Come spiegato nell’intervista, per un gran numero di immagini è necessaria una potenza di calcolo molto elevata, ancora al giorno d’oggi difficile da ottenere, proprio per questo motivo i primi foundation model sono basati sul NLP (Natural Language Processing).
- **Approccio “Big Data” a “qualità dei dati”**: Solo grandi aziende con milioni di utenti sono in grado di ottenere un numero massiccio di dati, per tutti gli altri, con un numero di dati limitato, va cambiato il metodo di approccio, basandosi di più sulla qualità dei dati in possesso.
- **Il movimento data-centric**: In passato il paradigma principale era quello di scaricare i dati ed adattare il modello ad essi concentrandosi principalmente su quest’ultimo. Grazie a questo approccio i modelli sono notevolmente migliorati negli ultimi anni, così da rendere questo problema praticamente risolto. Per questo motivo, afferma Ng, che è il momento di concentrarsi di più sul migliorare la qualità dei dati.
- **Aumentare i dati è la scelta migliore?**: Negli ultimi anni per problemi con un gran quantitativo di dati con del rumore all’interno, la soluzione è sempre stata aumentare i dati affinché il modello possa fare una media su di essi, nonostante si sapesse chiaramente che il problema era di un’altra natura.
- **Utilizzo dei dati sintetici**: Ng specifica come l’utilizzo dei dati sintetici sia uno strumento molto utile per l’approccio data-centric, anche questo però deve essere utilizzato correttamente.

Andrew Ng presenta delle idee molto chiare sull' utilizzo dell'approccio data-centric. In particolare specifica come l'attenzione debba essere spostata sulla qualità dei dati. Questo nuovo campo è molto efficiente in termini di spesa economica, in quanto un modello con pochi dati a disposizione ma di ottima qualità è comunque molto performante.
L'ingegneria dei dati diventa quindi fondamentale per lo sviluppo di modelli efficienti, concentrandosi sui dati per cui il modello tende ad avere problemi, riducendo il rumore ad essi associato, ad esempio con un corretto labeling di questi ultimi.

Ng fornisce alcuni esempi dove il processo data-centric ha funzionato con ottimi risultati. Durante la costruzione di un modello di speech-recognition, Andrew notò che il modello faceva fatica particolarmente, in tutte le parti di discorso dove in sottofondo era presente il rumore di un'automobile. Per risolvere questo problema fu necessario semplicemente aggiungere dei dati con all'interno il problema stesso, in questo caso il rumore in sottofondo, invece che collezionare più dati generici, i quali sarebbero stati anche più costosi economicamente.

L'utilizzo dei dati sintetici può essere un'ottima soluzione al problema della limitatezza dei dati, anche questo approccio però va usata con cautela e ragionamento. Ng fornisce un esempio per far capire a pieno il significato delle sue parole:
Prendendo in considerazione un modello che vuole identificare i difetti esterni di uno smartphone, questi possono essere di varia natura: graffi, pixel malfunzionanti, ammaccature, decolorazione dei materiali ed altri tipi di difetti. Se durante l'addestramento del modello, svolgiamo un'analisi degli errori e ci accorgiamo che quest'ultimo non è molto performante sotto il punto di vista delle ammaccature, è qui che ci può venire in aiuto l'utilizzo dei dati sintetici, in quanto basterà fornire al modello più dati di quella natura in modo da poter risolvere il problema.

Lo studio di Andrew ha quindi delle ottime fondamenta, il mondo dell'IA potrebbe a breve passare da un approccio model-centric ad uno data-centric, in quanto non tutti disponiamo di dataset si grandi dimensioni. Ng inoltre, spiega molto bene quando e come sarà necessario intervenire sui dati. Questo tipo di approccio in futuro sarà sempre più fondamentale, soprattutto adesso che il mondo dell'intelligenza artificiale si sta aprendo a tutti grazie all'hardware in continua evoluzione ed ai foundation models utilizzabili da qualsiasi utente. Sarà sicuramente necessario nei prossimi anni concentrare le proprie forze su questo nuovo movimento ed accantonare per adesso l'approccio model-centric, in quanto come anche riportato da Andrew, già parzialmente perfetto grazie agli studi degli ultimi anni su di esso.