Homework 1

Davide Gattini

14/10/2022

Andrew NG, globalmente riconosciuto nel mondo dell'intelligenza artificiale, ha identificato un grande cambiamento in questo ambito: sostiene che l'adozione dell'approccio *data-centric*, per cui è sufficiente avere a disposizione pochi dati ma di qualità, sia una soluzione migliore rispetto ad avere un dataset con molti dati, usati per addestrare un modello che successivamente si pensa a modificare per ottenere risultati migliori (approccio *model-centric*). Quindi, secondo Andrew NG, collezionare sempre più dati risulta costoso e meno efficace che occuparsi dell'ingegnerizzazione dei dati (che quindi prevede la pulizia di quest'ultimi) e, poiché è un approccio già adottato, per evitare che rimanga solo un'intuizione, crede necessario che esso diventi sistematico.

L'utilizzo di quei modelli che Andrew NG chiama *foundation models*, potrebbe risultare vantaggioso qualora si disponesse di una quantità di dati sufficiente per addestrare il modello e renderlo utile ed efficace per differenti tasks. Però non sempre questo requisito viene soddisfatto, infatti come Andrew NG sottolinea, ci sono realtà in cui il dataset a disposizione è "povero", quindi è conveniente e produttivo seguire l'approccio data-centric. Di contro, è necessario trovare del personale qualificato, che lo stesso Andrew NG chiama MLOps, che deve curarsi di questo aspetto per bilanciare una bassa quantità di dati con una buona qualità degli stessi, ma molto spesso tutto ciò viene già fatto, in quanto è sempre prevista l'acquisizione dei dati e la loro organizzazione, in modo che siano utilizzabili e coerenti.

E' dunque furbo ed efficace risolvere i problemi specifici cercando di dare in pasto al modello dati specifici con cui colmare l'inconsistenza dovuta ad un addestramento errato. Quindi continuare a raccogliere un quantità indefinita di dati (a volte non disponibili) potrebbe risultar come una perdita di tempo e risorse che potremmo impiegare a migliorare dati che già abbiamo e conosciamo. Se ci si dovesse domandare cosa fosse meglio, se migliorare il modello o i dati, penserei a migliorare i dati: a parità di dati, migliorare il modello potrebbe sì risolvere il problema in maniera più o meno efficace, ma il modello adottato potrebbe comunque comportarsi diversamente da quello che ci aspettiamo, poiché, a causa del rumore con cui mal addestriamo l'architettura (qualsiasi essa sia), non sarà efficace quanto lo **stesso modello** che alimentiamo con gli **stessi dati** ma privati del rumore.