Che cosa sono i dati?

Il ruolo dei dati in un sistema di IA

Nel senso digitale generale, i dati corrispondono alle informazioni utilizzate, elaborate e generate dal software in un sistema informatico.

Non esiste AI senza dati. I dati svolgono un ruolo centrale in tutti i processi di apprendimento automatico, poiché vengono utilizzati sia per l'addestramento che per i test. Si presentano anche sotto forma di parametri utilizzati per gestire i processi di formazione. Infine, il sistema di IA è una combinazione di una certa architettura software con tutti i parametri addestrati, il cosiddetto modello, che è anche un dato.

Comprendere il ruolo dei dati nei sistemi di IA e il modo in cui vengono selezionati, documentati e diffusi è essenziale per poter valutare il comportamento di un sistema di IA. Questo è importante in termini di riproducibilità o per confrontare due diversi sistemi di IA.

Nel mondo dell'elaborazione del linguaggio naturale, ad esempio, la disponibilità di grandi quantità di dati parlati e scritti è fondamentale per le prestazioni dei correttori ortografici, della predizione nei motori di ricerca o, naturalmente, della traduzione automatica. I dati vengono utilizzati per costruire i cosiddetti modelli linguistici, che alimentano ulteriori processi con rappresentazioni statistiche di combinazioni di parole o frasi.

La sostenibilità dei sistemi di IA dipende quindi strettamente dai metodi di gestione dei dati utilizzati nella loro progettazione.

Dati per sistemi di IA supervisionati e non supervisionati

Come abbiamo già visto, i sistemi di IA sono di due tipi, a seconda del modo in cui i dati vengono utilizzati per addestrarli. I sistemi supervisionati si basano sulla fornitura di input e dei corrispondenti output previsti. L'addestramento consiste quindi nell'insegnare al sistema a generare l'output più probabile a partire da input sconosciuti. I dati possono essere ottenuti in vari modi. Ad esempio, un database di immagini in cui ogni immagine è associata a parole chiave o una raccolta di documenti digitalizzati che sono stati trascritti da annotatori (cfr. figura seguente).

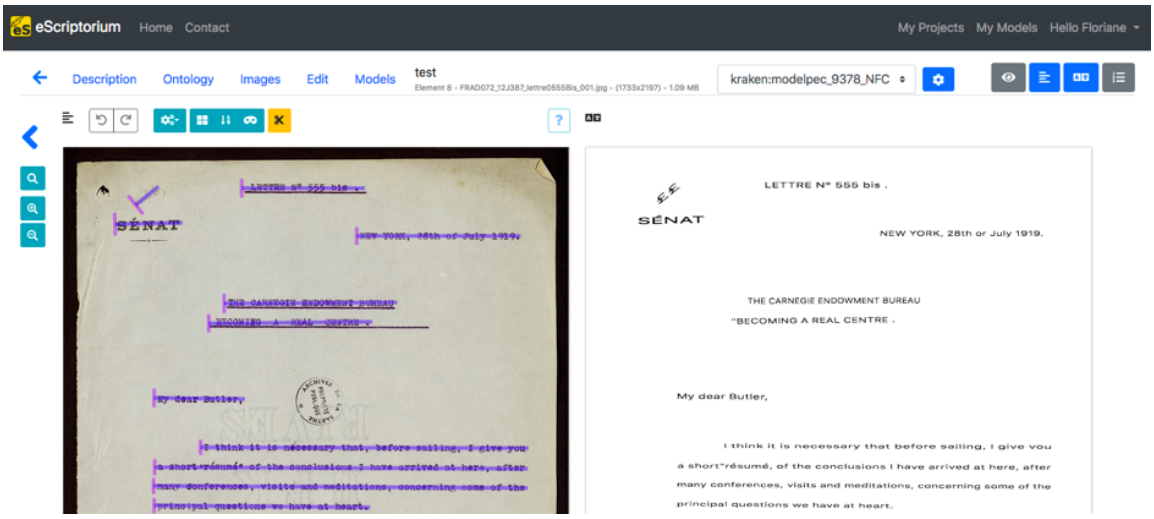

Figura: Trascrizione automatica di una lettera di Paul D'Estournelles (per gentile concessione di F. Chiffoleau, Coll. Archives de la Sarthe)

I sistemi di intelligenza artificiale basati sull'addestramento non supervisionato non saranno progettati per un comportamento specifico, ma per sopportare le proprietà statistiche dei dati di addestramento. È il caso, ad esempio, di modelli linguistici come il BERT, che tendono ad associare posizioni simili in uno spazio matematico a parole che hanno lo stesso comportamento sintattico o semantico, osservato grazie a un gran numero di frasi campione per ogni parola. Tali modelli sono ad esempio molto bravi a prevedere i sinonimi o le parole successive di una determinata sequenza.

Fonti - selezione, documentazione, preparazione, annotazione

La progettazione di un sistema di IA si basa essenzialmente sulla corretta progettazione dell'insieme di dati utilizzati per l'addestramento. Tra i vari fattori che entrano in gioco possiamo citare la rilevanza dei dati per il compito da svolgere, la dimensione dei dati che dovrebbe corrispondere alla complessità dell'architettura del software di IA --- più parametri matematici si devono addestrare, più dati sono necessari --- e la varietà di campioni che dovrebbe riflettere la complessità del compito.

A seconda delle fonti di dati, questi devono essere selezionati e spesso ripuliti prima di essere immessi nel processo di addestramento. Se prendiamo ad esempio l'addestramento di un modello linguistico sulla base di contenuti web, i vari campioni devono essere ordinati in base alla lingua effettiva, privati del codice web che li accompagna (HTML, Javascript ecc.) ed eventualmente rimescolati per evitare violazioni del copyright. Un buon esempio di preparazione dei dati è la progettazione del corpus OSCAR[1].

La progettazione di dati annotati per sistemi di intelligenza artificiale supervisionati è più complessa, poiché comporta la progettazione di uno schema di annotazione, l'organizzazione

di campagne di annotazione e il controllo della qualità dei dati annotati, ad esempio valutando l'accordo tra gli annotatori sugli stessi dati.

Nel complesso, è essenziale che il processo di progettazione sia ben documentato per poter risalire alla fonte di eventuali comportamenti errati nel sistema addestrato risultante.

Sistemi di intelligenza artificiale distorti

Come appena accennato, il comportamento di un sistema di intelligenza artificiale riflette da vicino la natura dei dati che sono stati utilizzati per addestrarlo. Questo, a sua volta, può indurre tutta una serie di possibili distorsioni che possono derivare dalla selezione dei set di dati. Per esempio, un modello linguistico addestrato solo su articoli di giornale coprirà tipi di espressioni e argomenti completamente diversi rispetto a uno per il quale sono stati scelti contenuti di letteratura o di social network. Allo stesso modo, i sistemi di generazione di immagini rifletteranno la dimensione e la varietà dei database di immagini di partenza (ad esempio, opere artistiche) che sono stati considerati.

Nel caso dei sistemi supervisionati, un pregiudizio specifico può derivare dal modo in cui sono state progettate le etichette di annotazione e dal modo in cui gli annotatori, con il loro background culturale, interpreteranno i dati. Se, ad esempio, si vuole identificare l'incitamento all'odio sui social network, il modo in cui i sentimenti vengono interpretati dagli annotatori può variare in base all'età, alla cultura e ai sentimenti personali degli annotatori nei confronti del materiale da annotare.

In definitiva, bisogna sempre tenere presente che i sistemi di IA sono per loro natura molto conservativi per quanto riguarda i dati di addestramento e quindi le osservabili esistenti. Non ci si può aspettare alcun tipo di creatività da un sistema di IA.

Hosting, pooling, distribuzione dei dati

A causa delle dimensioni e della possibile complessità dei dati di addestramento dei sistemi di IA e dei modelli risultanti, sono state avviate diverse iniziative per consentirne l'hosting e la distribuzione.

Gli insiemi di dati e i modelli aperti possono essere ospitati in archivi specializzati (ad esempio, l'image data resource[2]) o in archivi generici nazionali o internazionali (ad esempio, Zenodo[3]). Tali archivi forniscono solitamente l'infrastruttura necessaria per la gestione della paternità, delle licenze, delle versioni e dell'archiviazione dei contenuti.

Nel caso di attività complesse, in cui diversi team lavorano in parallelo all'annotazione di una varietà di campioni di dati, alcune iniziative fungono da cataloghi per le fonti di dati corrispondenti. È il caso, ad esempio, dell'iniziativa HTR United[4] che raccoglie i metadati dei documenti annotati per il riconoscimento del testo (scritto a mano).

1. OSCAR corpus: https://oscar-corpus.com/ ↩
2. Image data resource: https://idr.openmicroscopy.org/ ↩
3. Zenodo: https://zenodo.org/ ↩
4. HTR United: https://htr-united.github.io ↩