Un’intelligenza artificiale senza pregiudizi

“Una donna felice e un uomo serio che si abbracciano in un parco.” Che il tuo computer sia già in grado di descrivere le tue foto in questo modo è fantastico. E per molte persone che navigano in Internet utilizzando i lettori di schermo, è anche una tecnologia quasi essenziale. Tutto questo grazie all’intelligenza artificiale.

Certo, le macchine non sono perfette, a volte si guastano. Ma ultimamente stiamo scoprendo qualcosa di preoccupante: a volte i computer falliscono come fallisce un essere umano. Per una donna e un uomo con la stessa espressione, i sistemi di intelligenza artificiale possono tendere a credere che lei sia felice e che lui sia di cattivo umore. Chiamiamo questo tipo di errori pregiudizi e includono tendenze razziste, sessiste, abiliste… che possono finire per fare del male alle persone.

Per studiare questi pregiudizi prendiamo in esame un’applicazione specifica, il riconoscimento automatico delle emozioni nelle fotografie. Innanzitutto, dobbiamo chiarire al computer cosa intendiamo per “emozione”. La classificazione più utilizzata si basa su 6 emozioni di base: paura, tristezza, felicità, rabbia, disgusto e sorpresa. Questa classificazione è stata proposta dallo psicologo Paul Ekman negli anni ’70. Queste emozioni hanno dimostrato di essere più o meno universali e riconosciute da tutti. Tuttavia, è stato anche dimostrato che si riconoscono meglio tra persone dello stesso gruppo sociale, sesso, età… Non tutti noi ci esprimiamo esattamente allo stesso modo, né leggiamo le espressioni allo stesso modo. Anche senza rendercene conto, siamo di parte.

Queste differenze si riscontrano in molti contesti e talvolta si trasformano in stereotipi e pregiudizi. Ad esempio, ci aspettiamo che le donne siano più felici che arrabbiate,

mentre per gli uomini vale il contrario. E questo si riflette su internet, dove le foto tendono a includere soprattutto donne sorridenti.

D'altra parte, affinché un sistema di intelligenza artificiale impari a distinguere queste emozioni, dobbiamo anche pensare a come le persone le comprendono. In realtà, la faccia è solo una parte di un puzzle molto complesso. Contribuiscono anche i gesti, la postura, le nostre parole… Nonostante si stia lavorando per risolvere tutte queste modalità con l'intelligenza artificiale, la forma più popolare e versatile è il riconoscimento basato sulle foto dei volti.

La creazione di un'intelligenza artificiale priva di pregiudizi è piuttosto una sfida. E tutto inizia da come facciamo "imparare" questa tecnologia. Chiamiamo il campo dell'intelligenza artificiale dedicato a questo apprendimento *automatico*. Sebbene esistano diverse forme di apprendimento, la più comune è l'apprendimento supervisionato.

L'idea è semplice: impariamo dagli esempi. E l'intelligenza artificiale ha bisogno di sapere per ogni esempio cosa vogliamo ottenere. Per imparare a riconoscere le emozioni, abbiamo bisogno di un mucchio di foto di volti con emozioni diverse: felici, tristi, ecc. La chiave è che per ogni foto, dobbiamo sapere quale emozione appare. Successivamente, passiamo le foto e le emozioni associate all'intelligenza artificiale. Attraverso un algoritmo di apprendimento, il sistema imparerà "da solo" a mettere in relazione le foto con le emozioni che appaiono. Immagine per immagine, chiediamo di prevedere un'emozione: se è giusta, andiamo avanti, e se è sbagliata, aggiustiamo il modello per correggere questo caso. A poco a poco, l'intelligenza artificiale imparerà e fallirà sempre meno. Se ci pensiamo, non è così diverso da come noi umani impariamo.

Come si può vedere, gli esempi sono essenziali in questo processo. Sebbene ci siano progressi che ci consentono di

apprendere con pochi esempi o esempi con errori, un insieme ampio e ben catalogato di esempi è fondamentale per ottenere una buona intelligenza artificiale. Sfortunatamente, in pratica è comune avere esempi con errori, ad esempio volti etichettati con l'emozione sbagliata a foto senza volti o con volti di animali. Ma ci sono altri problemi, a volte più sottili e preoccupanti: il razzismo, il sessismo, l'abilismo…

Se i nostri esempi sono distorti, la macchina imparerà e riprodurrà questi pregiudizi. A volte moltiplicherà anche l'effetto dei pregiudizi. Ad esempio, se nelle nostre foto abbiamo solo persone arrabbiate dalla pelle scura e persone felici dalla pelle chiara, è molto probabile che l'intelligenza artificiale finisca per confondere il colore della pelle con l'umore. Tenderà a prevedere la rabbia ogni volta che vede persone dalla pelle scura.

Sfortunatamente, questa non è solo una teoria. È già stato dimostrato, ad esempio, che i sistemi di analisi facciale per riconoscere il genere falliscono più per le donne nere che per gli uomini bianchi e commettono regolarmente errori con persone che sono di aspetto trans.

Uno degli esempi più noti è stato quando nel 2018 un sistema di intelligenza artificiale ha erroneamente identificato 28 membri del Congresso degli Stati Uniti come criminali_. Dei politici identificati, il 40% erano persone di colore, sebbene rappresentassero solo il 20% del Congresso. Tutto questo perché il sistema era stato addestrato principalmente sui bianchi e confondeva le persone di colore tra loro.

Rilevare e ridurre questi pregiudizi è un campo di ricerca molto attivo con un grande impatto sociale. Molti di questi pregiudizi sono sottili e correlati a diversi fattori demografici allo stesso tempo, il che rende l'analisi difficile. Inoltre, tutte le fasi dell'apprendimento devono essere riviste, dalla raccolta dei dati e le relative misurazioni all'applicazione finale. E normalmente non sono le stesse persone che lavorano in ogni fase.

Su internet ci sono molti database di emozioni già etichettati. Sfortunatamente, i database più grandi spesso hanno anche forti pregiudizi di sesso/genere, razza ed età. È necessario che a poco a poco si sviluppino database diversificati ed equilibrati su cui lavorare. Cioè, dobbiamo includere tutti i tipi di persone nei nostri database. Inoltre, tutti devono essere ben rappresentati in ogni emozione. Infine, se vogliamo raccogliere dati senza pregiudizi, dobbiamo pensare all'intero processo. Tutte le fasi, dalla raccolta dei dati ai test finali di un'intelligenza artificiale, devono essere svolte in modo attento e accessibile. Ed è necessario coinvolgere persone che sappiano riconoscere e segnalare possibili pregiudizi in ognuno di loro.

L’intera faccenda del riconoscimento delle emozioni può sembrare astratta, ma ha già importanti applicazioni. La più comune è la tecnologia assistiva, come la descrizione automatica delle foto per i non vedenti. È già utilizzato anche nei robot domestici. Può essere applicato anche in medicina, dove è stato possibile riconoscere automaticamente il dolore nei neonati che a volte non lo esprimono attraverso il pianto.

In ogni caso, lo studio dei pregiudizi nell’intelligenza artificiale va oltre le emozioni. Le tecnologie che sviluppiamo hanno un enorme impatto sulla vita delle persone. Abbiamo il dovere morale di assicurarci che siano equi, che il loro impatto sul mondo sia positivo.

Vogliamo costruire un’intelligenza artificiale di cui possiamo fidarci e che ci faccia sorridere. *(Se vuoi partecipare alla creazione di un database di emozioni più equo e diversificato, puoi collaborare con il progetto Emotional Films)*

Ricerca a cura di Mikel Galar Idoate, Daniel Paternain e Iris Dominguez Catena, pubblicata su The Conversation