La fallace percezione delle macchine intelligenti

- Arturo Di Corinto, 08.08.2019

Hacker's dictionary. Ecco come trarre in inganno i sistemi di riconoscimento visivo guidati da algoritmi di intelligenza artificiale

L'Intelligenza Artificiale (IA) può migliorare la nostra vita guidando al posto nostro, prendendosi cura delle persone anziane e dei malati, svolgendo lavori pericolosi e usuranti, e ottimizzando la gestione di grandi quantità di dati. Tutto questo è possibile grazie ai recenti sviluppi sia della robotica che delle tecniche di IA come, ad esempio, la capacità delle reti neurali di apprendere che ha apportato molti benefici al settore della *computer vision*, con applicazioni come il riconoscimento degli oggetti, il *video labelling*, eccetera.

La *machine perception* è forse il settore dell'IA su cui l'avvento del deep learning ha più inciso. Il basso costo dell'informatica computazionale, la disponibilità di grandi quantità di dati e l'affinamento di reti di algoritmi neurali ha permesso all'IA di eseguire compiti di classificazione visiva anche meglio degli esseri umani. Poiché le "macchine intelligenti" che riconoscono al posto nostro luoghi e persone sono utilizzate in ambito di sicurezza e sorveglianza negli aeroporti, negli *smart buildings* o al telefono, è ora di chiedersi se funzioni per davvero.

Durante la seconda guerra mondiale i segnali stradali venivano spostati o modificati per disorientare il nemico. Questa alterazione iconografica dell'ambiente ha tutta una serie di equivalenti nel mondo digitale. In un saggio dal titolo Ma le reti neurali sognano pecore elettriche? pubblicato nella raccolta *State Machines dall'Institute for Network Culture di Amsterdam* e che fa il verso al famoso libro da cui è stato tratto il film *Blade Runner* (*Do Androids Dream of Electric Sheep?*) Janelle Shane spiega come si possono ingannare le reti neurali deputate al riconoscimento visivo.

Il primo trucco per confonderle è basato sul foto collage, ma un altro, più interessante, lo ha sperimentato di persona. Janelle ha chiesto agli amici su twitter di mandarle foto di pecore. Una foto che rappresentava delle pecore dipinte di arancione data in pasto a una rete neurale veniva interpretata come un gruppo di fiori di campo. L'artista James Bridle ha realizzato un video e una performance in cui si veda un'automobile che ruota su se stessa all'interno di un cerchio dipinto come una rotatoria stradale dimostrandosi capace di ingannare il sistema di riconoscimento di pattern delle macchine a guida autonoma che usano appunto la segnaletica orizzontale, le strisce bianche continue e i led sull'asfalto per arrivare a destinazione.

Altri scienziati hanno applicato degli adesivi ai segnali di Stop degli incroci stradali facendo fallire i sistemi artificiali di riconoscimento visivo basati su reti neurali ben allenate mentre un'altra ricerca ha dimostrato come il riconoscimenti facciale fallisca grazie all'uso di minuscole luci a infrarossi montate su occhiali da vista oppure con l'uso di una gabbia biometrica, cioè piccoli oggetti di metallo applicati sul viso che rendono alle macchine impossibile il compito assegnato come quello di aprire un cancello dopo aver riconosciuto la faccia del proprietario di casa.

Insomma le intelligenze artificiali basate sul deep learning possono essere ingannate cambiando posizione, forma, taglia e luminosità di ogni oggetto, compresi i volti umani.

Parliamo di tecniche nate per hackerare l'intelligenza artificiale, note in letteratura come *adversarial attack tecnique* che consistono nella modificazione di un elemento per rendere complesso o errato il

riconoscimento del tutto. Quindi, prima di affidare compiti importanti alle intelligenze artificiali è bene progettarle a prova di imbroglio.