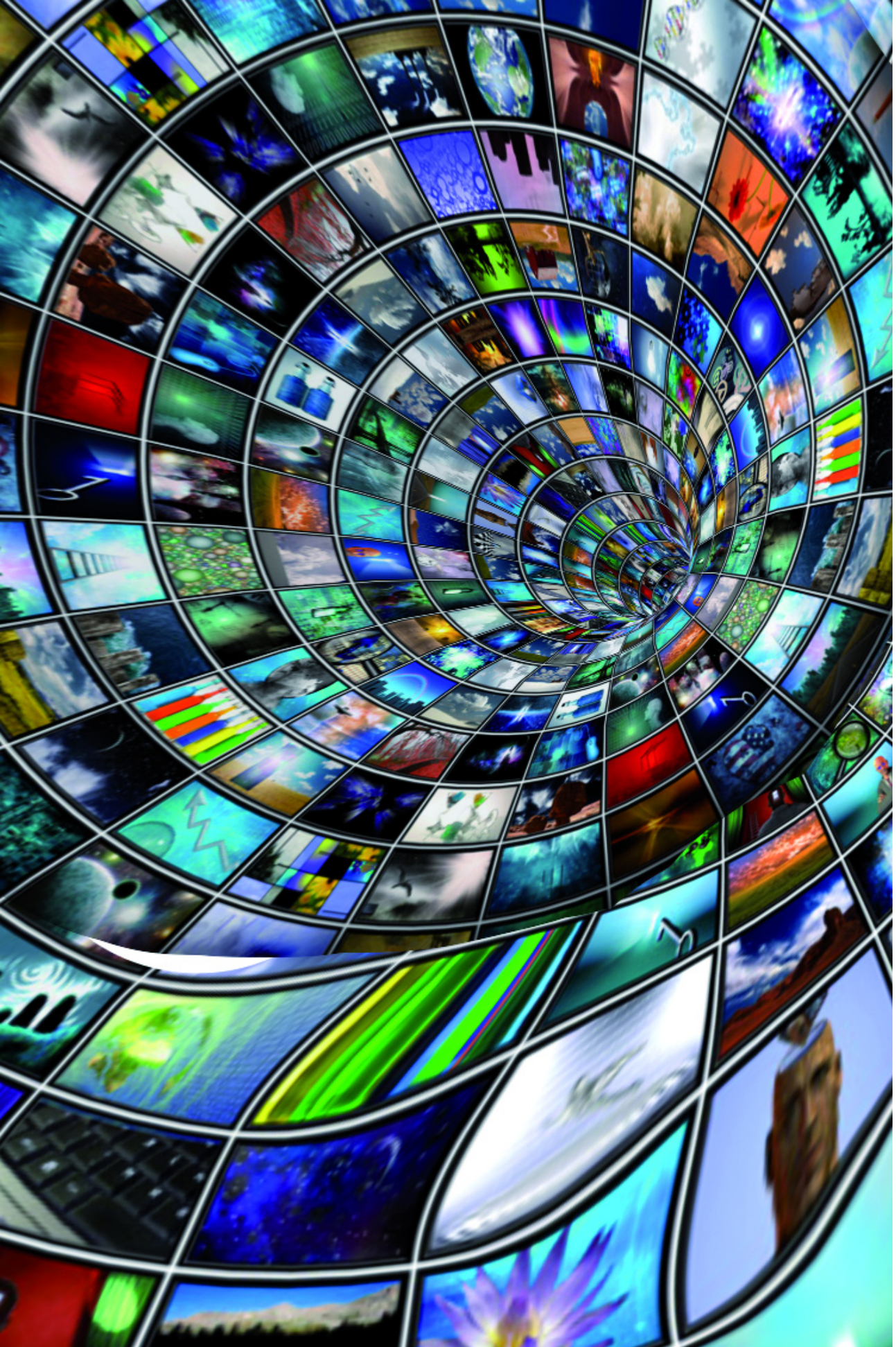


LA VISIONE COMPUTERIZZATA DELL'ERA DEL DEEP LEARNING

SFIDE E OPPORTUNITÀ PER LA CYBERSECURITY

Barbara Caputo

Data un'immagine, la visione computerizzata ambisce a sviluppare algoritmi in grado di rispondere in maniera autonoma alla domanda: «che cosa vedi?». Questa sfida si sta declinando in maniera innovativa con l'avvento del deep learning, tecnologia capace di trarre vantaggio dalla presenza di grandi quantità di riproduzioni. L'articolo offre una breve storia della visione computerizzata: com'è cambiata negli ultimi anni grazie all'utilizzo dei Big Data*, quali obiettivi significativi per la cybersecurity stiano divenendo raggiungibili e quali siano le sfide da affrontare a medio termine.*

Viviamo in una società in cui foto e video vengono acquisiti, visti e condivisi in maniera pressoché continua. Questa tendenza irreversibile è dovuta all'ampia diffusione di dispositivi elettronici dotati di telecamere, come telefoni cellulari di ultima generazione, telecamere digitali a basso costo, laptop ecc. Inoltre, la diffusione di piattaforme di social network come Facebook, Pinterest, Instagram e molte altre ha reso socialmente accettabile la condivisione d'immagini, a vari livelli appartenenti alla propria sfera privata, tra utenti e la loro rete di conoscenze digitali.

Alcuni numeri consentono di avere una percezione più esatta del fenomeno. Si consideri, infatti, che: Flickr ospita più di sei miliardi di foto; ogni giorno vengono caricate su Facebook circa 300 milioni di foto; la Bbc converte più di 400 ore di programmazione per il suo servizio internet iPlayer; YouTube fornisce più di tre miliardi di ore di visualizzazioni di video ogni mese e così via.

Il fenomeno della diffusione massiva di immagini, e in generale di dati visivi, non è limitato solo ai social media. Anche la dimensione delle collezioni d'immagini e di video digitali dei privati cittadini sta crescendo vertiginosamente. Grazie alle aumentate capacità di memoria di smartphone, tablet e laptop si stima che in media ogni famiglia abbia immagazzinato, nella memoria dei propri dispositivi elettronici dotati di telecamere, foto e video per circa tre terabytes.

Davanti a questa enorme mole d'immagini, l'intelligenza artificiale è chiamata a fornire nuove soluzioni che rendano possibile la catalogazione, l'analisi e la comprensione automatica dei dati visivi. In particolare, l'incontro recente tra la visione computerizzata e gli algoritmi di apprendimento profondo (deep learning) sta portando a risultati rivoluzionari nella comprensione e nell'analisi automatica d'immagini: dal riconoscimento automatico di oggetti in una foto alla possibilità di stimare l'età e la classe sociale di una persona a partire unicamente da una sua immagine, al comprendere, analizzando automaticamente un video, chi tra due persone è in una posizione di potere rispetto all'altra. In questo articolo si fornisce una breve introduzione storica alla visione computerizzata e al progresso raggiunto in questo campo – da quando l'utilizzo di tecniche di deep learning ha permesso di adoperare in maniera efficiente le grandi quantità di immagini disponibili gratuitamente e liberamente sul web – e si descrivono gli ambiti di applicazione nel settore della sicurezza – dove l'utilizzo di metodi moderni di visione computerizzata può essere più efficace – e le sfide tecnologiche e applicative ancora aperte.

LA VISIONE COMPUTERIZZATA

La visione computerizzata è un campo di ricerca interdisciplinare, il cui scopo è sviluppare algoritmi in grado di ottenere in maniera autonoma una comprensione delle immagini, pari o superiore a quella degli esseri umani, ovvero la capacità di capire, a partire dalla visualizzazione di un'immagine, quali oggetti essa contenga; in quale posizione si trovino nella composizione; quali siano le proprietà degli oggetti selezionati e che relazioni leghino gli oggetti tra loro e così via. Ad esempio, l'analisi automatica di un'immagine scattata in una qualsiasi area di Roma da parte di un algoritmo di visione computerizzata potrebbe determinare se il paesaggio nella foto rappresenti la veduta di una città o di una spiaggia. Essendo una foto realizzata in un centro abitato, potrebbe riconoscere la città in cui la foto è stata acquisita, localizzare e riconoscere gli edifici di rilevanza storica in essa contenuti, separare gli immobili dalla strada, dalle macchine e dal marciapiede; potrebbe localizzare i pedoni e le macchine presenti nella foto, contare per ciascuna categoria quanti sono e, se la risoluzione dell'immagine lo permette, identificarli a partire da particolari del loro viso (per le persone) e dalla loro targa (per le macchine). Questo campo di ricerca è nato negli anni Sessanta, inizialmente come attività collaterale dell'intelligenza artificiale. Nell'ambito del rilevante sforzo innovativo che in quegli anni portò a grandi progressi nel campo dei sistemi autonomi intelligenti, il problema della comprensione automatica delle immagini fu inizialmente sottovalutato, tanto che nel 1966 Seymour Papert, professore al Massachusset Institute of Technology, pubblicizzò un progetto estivo sulla visione per studenti in cerca di tirocini estivi in vari laboratori, il cui scopo era di costruire una mappa tra l'input di una telecamera collegata a un computer e una descrizione testuale degli oggetti visualizzati e del contesto in cui essi si trovavano. Questo sarebbe poi stato integrato in un sistema autonomo intelligente, come un robot, capace di ragionare ed elaborare piani e strategie sulla base delle informazioni raccolte. Più di sessant'anni dopo la sfida è ancora aperta, con sistemi automatici di visione non in grado ancora di raggiungere, in molte applicazioni le capacità di comprensione di un'immagine propria di un bambino di tre anni.

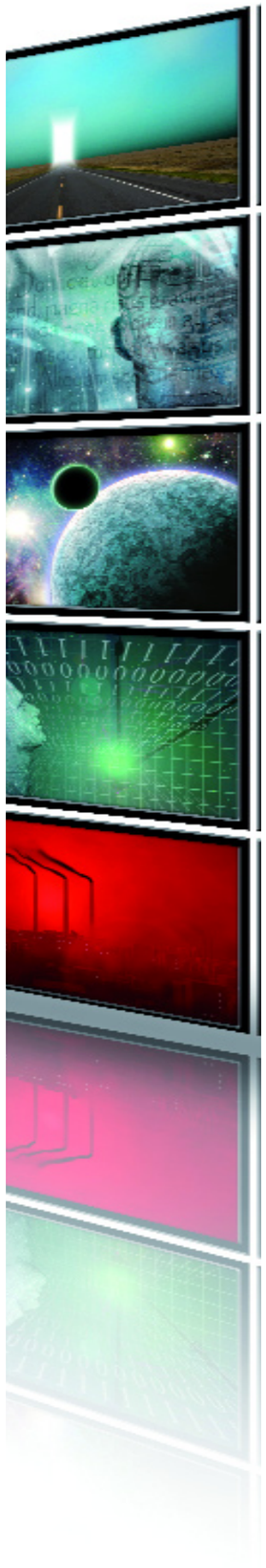
LA VISIONE COMPUTERIZZATA NELL'ERA DEI BIG DATA

Nel 2012, per la prima volta, un algoritmo basato sul deep learning partecipò alla ImageNet Large Scale Recognition Challenge, competizione aperta ai ricercatori del campo della visione computerizzata di tutto il mondo, il cui obiettivo è quello di classificare correttamente, tramite un algoritmo, immagini raffiguranti 1.000 oggetti differenti, utilizzando per l'apprendimento dei modelli un database di circa un milione e mezzo di immagini. Questo algoritmo, sviluppato dai ricercatori dell'università di Toronto e basato sulle reti neurali convolutive di tipo profondo, vinse la competizione con uno scarto rispetto al metodo secondo classificato di più del 10%. In tutte le competizioni degli anni precedenti, lo scarto medio tra il primo e secondo classificato non aveva mai superato il 5%. Da quell'anno, nessun algoritmo non basato sul deep learning è più riuscito a piazzarsi tra i primi tre classificati. Questo ricordo offre una stima quantitativa dell'impatto che il deep learning ha avuto sulla visione artificiale nell'ultimo quinquennio. Le reti neurali convolutive sono algoritmi di classificazione introdotti negli anni Ottanta, che cercano d'imitare il funzionamento delle strutture neurali dei sistemi biologici. L'idea di renderle molto profonde, insieme ai progressi impressionanti raggiunti nel campo computazionale, ha reso possibile l'utilizzo di questi algoritmi con enorme profitto.

Un ingrediente fondamentale del successo del deep learning è la capacità, unica nel panorama dell'intelligenza artificiale, di apprendere da grandissime quantità di immagini senza incorrere in problemi di memoria, convergenza o complessità computazionale, imparando in maniera dinamica come rappresentare l'informazione racchiusa nelle illustrazioni (descrittori), insieme a come usarla per riconoscere il contenuto delle stesse (classificatori). Questo permette di imparare in maniera ottimale entrambe le componenti allo stesso tempo, ottenendo descrittori più espressivi di quelli sviluppati nel passato, combinati a classificatori capaci di sfruttarli al meglio. Oggi il deep learning è il paradigma di apprendimento dominante nella visione computerizzata, l'ingrediente fondamentale dietro i recenti sviluppi nel campo e il fiorire di prodotti, startup e investimenti importanti nel settore da parte di giganti come Microsoft, Google, Facebook e Amazon.

LA VISIONE COMPUTERIZZATA PER LA SICUREZZA

Grazie alla crescente notorietà, affermazione e semplicità d'uso, social network e piattaforme di condivisione dei contenuti multimediali (Youtube, Instagram ecc.) sono state utilizzati da singoli soggetti e da organizzazioni terroristiche per diffondere i propri messaggi di propaganda. Questa piazza digitale svolge spesso la funzione di cassa di risonanza per raggiungere in maniera diretta e spettacolare i principali canali di comunicazione televisivi e influenzare la narrativa dell'informazione nei paesi occidentali, accrescendo la sensazione di accerchiamento 'percepita' dai cittadini. Nello stesso tempo, la diffusione di questi contenuti può rappresentare una possibilità per chi, per motivi di difesa e sicurezza, ha l'esigenza di rilevare tempestivamente ogni possibile indicatore di rischio. Ad esempio, nelle immagini e nei filmati condivisi dagli utenti nei canali social come Facebook, Instagram, Twitter ecc. si potrebbero identificare, tramite tecniche automatiche di controllo basate su algoritmi di visione artificiale, singoli individui attenzionati o rilevare la presenza di materiale di propaganda. I processi automatici potrebbero segnalare agli operatori umani le situazioni di potenziale interesse così da avviare appropriate iniziative. Similmente, l'analisi automatica d'immagini e di video trovati nei dispositivi elettronici di un sospetto può fornire informazioni rilevanti ai fini d'indagine. Nel caso d'immagini che siano state private dei loro metadati, ovvero dell'informazione normalmente associata al file che permette d'identificare in quale parte del mondo ciascuna foto sia stata scattata, la visione computerizzata permette di riconoscere, all'interno delle immagini, informazioni riguardo al tipo di vegetazione, allo stile degli edifici, ai cartelloni pubblicitari nonché alla lingua e ai caratteri adottati ecc. L'insieme consente di formulare ipotesi dettagliate sul luogo dove esse sono state acquisite e ne agevola la geolocalizzazione automatica. Infine, la presenza oramai capillare di telecamere di sorveglianza nelle città e nei



luoghi pubblici (stazioni, aeroporti, centri commerciali ecc.) offre la possibilità di identificare comportamenti o individui sospetti, anche grazie a caratteristiche biometriche della postura, della corporatura e della camminata di una persona. Anche qui, la visione computerizzata offre grandi opportunità attraverso la possibilità di filtrare terabytes di informazione non rilevante in maniera efficace ed efficiente, per poi concentrare le risorse computazionali esclusivamente sulle parti dei filmati che hanno alta probabilità di contenere dettagli rilevanti, e su tali segmenti di video applicare algoritmi specializzati nella re-identificazione di soggetti, in grado di stabilire, con una probabilità sufficientemente alta, se un individuo ripreso da una telecamera nell'aeroporto X sia o meno lo stesso ritratto da un'altra telecamera nella stazione Y. Questi sono solo alcuni esempi delle possibili applicazioni della visione computerizzata moderna nel campo della sicurezza. I progressi ottenuti grazie all'utilizzo di algoritmi di deep learning, fortemente legati alla possibilità di servirsi di grandi quantità di dati, permettono oggi di affrontare queste sfide con ragionevoli probabilità di successo.

GUARDANDO AVANTI: NUOVE SFIDE PER LA VISIONE COMPUTERIZZATA

I progressi ottenuti e l'accelerazione seguita all'avvento del deep learning non devono far dimenticare quali sfide rimangono aperte per la ricerca di base in questo campo. Limitandosi a quelle più rilevanti, si ricordano:

– RICONOSCIMENTO DI CLASSI MOLTO SIMILI TRA LORO. Mentre possiamo considerare in via di risoluzione la capacità di localizzare e riconoscere in maniera efficace oggetti a livello generico (ad esempio cani, gatti, automobili, motociclette ecc.), non è stato ancora risolto il problema su come distinguere classi molto simili tra loro. Si pensi alla difficoltà di cogliere automaticamente la differenza tra tipi diversi di arbusti sulla base del loro fogliame, o di distinguere diversi tipi di uccelli o motociclette. Queste categorie hanno tendenzialmente molte caratteristiche visive in comune e la sfida è riuscire a disegnare algoritmi in grado di evidenziare automaticamente i tratti distintivi;
– RE-IDENTIFICAZIONE DI PERSONE SU LARGA SCALA TEMPORALE. I sistemi di re-identificazione moderni, basati sulla visione computerizzata, tendono a concentrarsi su scale temporali molto ristrette, in cui è ipotizzabile che il soggetto d'interesse stia indossando gli stessi indumenti e che l'illuminazione in cui sono acquisite le immagini sia sostanzialmente costante. Si tratta di rilassare queste assunzioni e arrivare a sistemi in grado d'identificare individui anche a distanza di tempo (giorni, settimane o mesi), ovvero in presenza di differenze significative nell'abbigliamento e nelle condizioni di illuminazione, nel caso in cui le immagini siano state acquisite da differenti sistemi di sorveglianza, dotati di telecamere con distinte caratteristiche tecniche.

La chiave per affrontare queste sfide risiederà sempre di più nell'utilizzo di grandi quantità di dati, congiuntamente allo sviluppo di architetture di deep learning finalizzate a modellarle. Ciò indica che il discrimine fondamentale per il progresso in questi campi sarà sempre più legato, oltre che al possesso di dati, alla capacità computazionale dell'hardware a disposizione