Credi a ciò che vedi? Il sempre più sottile confine fra reale e virtuale

Author : Giulia Boato

Date : 5 Ottobre 2020

La diffusione di strumenti di manipolazione di immagini e video facilmente accessibili a un sempre più vasto numero di persone ha fatto crescere, negli ultimi anni, i problemi legati alla verifica di autenticità e alla **credibilità dei dati multimediali**. Il noto detto popolare “credo a ciò che vedo” non può più essere considerato un paradigma valido: oggi, quando si osserva un contenuto multimediale, non è affatto scontato il fatto di essere davanti a qualcosa di autentico e reale, quanto piuttosto al frutto di una rielaborazione o addirittura di una sintesi artificiale. Per questo è fondamentale disporre di tecniche che consentano di rivelare possibili alterazioni dei contenuti, specie quando queste sono realizzate in maniera talmente accurata da essere non percepibili ad occhio nudo.

L’ingegneria forense applicata alla multimedialità si occupa esattamente di questo. L’idea è quella di produrre algoritmi il più possibile automatizzati in grado di scoprire una serie di **tracce nascoste nei dati**, che a loro volta consentano di determinare la sorgente dell’informazione (ad esempio quale modello di fotocamera è stata usato nell’acquisizione), così come eventuali elaborazioni effettuate sui dati a valle dell’acquisizione (fotomontaggi, inserimento di contenuti sintetici, filtraggi, ecc.).

I progressi ottenuti nell’ambito del *multimedia forensics* negli ultimi decenni sono estremamente rilevanti, ma vanno di pari passo con i progressi delle tecniche di generazione e manipolazione di immagini, in una sorta di sfida continua. In questo senso, un grosso *asset* nelle mani dei manipolatori è oggi costituito dall’intelligenza artificiale **(AI)**.

Attualmente esistono tecniche basate su AI capaci di generare immagini e video falsi di eccezionale qualità, senza la necessità di grandi competenze da parte dell’utilizzatore. Studi recenti dimostrano come la generazione di visi umani basata su tecnologie di AI (es. StyleGAN2 [1]) sia in grado di produrre immagini che il nostro cervello non riesce a distinguere da visi reali. Queste tecnologie sono alla base dei cosiddetti ***deepfakes***, video in cui una persona viene inserita in maniera estremamente realistica in situazioni mai avvenute nella realtà [2]. Le vittime abituali di questi *tool* sono le persone maggiormente popolari (politici, attori, sportivi, ecc.), sia per l’impatto e la visibilità ottenuti dalla manipolazione, sia per il fatto che di tali persone sono disponibili grandi moli di dati e immagini, necessarie ai motori di AI in fase di addestramento. È

così possibile creare sequenze video particolarmente accurate in cui alla vittima vengono fatte dire frasi completamente inventate, sovrapponendo espressioni facciali impostate dal creatore o semplicemente copiate da un modello umano (un attore).

Tutto ciò pone domande cruciali rispetto al confine esistente fra mondo reale e mondo virtuale e alla possibilità, per un utente umano, di distinguere tra i due. Per evitare che tali metodologie possano essere utilizzate in maniera non controllata, l'ingegneria forense studia sempre nuovi metodi di difesa, ideando tecniche via via più sofisticate. Ad esempio, è oggi possibile imparare i modelli facciali di una persona, per poi costruire una sorta di **firma digitale unica e robusta** (una sorta di impronta digitale) che possa essere utilizzata per identificare, con elevata affidabilità, sequenze manipolate o create *ad hoc* come dati non reali [3][4].

Un altro aspetto molto importante da tenere in considerazione è l'abitudine sempre più diffusa di condividere e scambiare dati multimediali e altre informazioni tramite social media e canali web. Questi strumenti possono diventare armi potentissime nelle mani dei manipolatori di informazioni, con lo scopo malevolo di influenzare l'opinione degli utenti [5][6]. Si ricordi, ad esempio, il caso del video della Presidente della Camera dei Rappresentanti statunitense Nancy Pelosi, condiviso su Twitter e diventato immediatamente virale, in cui un semplice ritardo introdotto nella sequenza suggeriva volutamente uno stato di ubriachezza. Il video - inizialmente scambiato per vero anche dalle persone vicine a Pelosi - aveva scatenato, forse per la prima volta in questa misura, un'accesa discussione sull'impatto diffamatorio dei video manipolati. Dal punto di vista tecnico, la **condivisione dei *fake*** via social media ha due impatti principali: da un lato genera una diffusione rapidissima, virale e poco controllabile dei contenuti, e dall'altro rischia di far perdere traccia della sorgente delle informazioni manipolate, a seguito delle elaborazioni insite nei passaggi da un social all'altro.

Nasce quindi un altro ruolo importante dell'ingegneria forense, consistente nella capacità di ricostruire la sequenza di passaggi subiti da un contenuto multimediale durante il suo ciclo di vita. Possiamo immaginare a tal proposito una sequenza in cui il video viene acquisito (ad esempio da un'emittente televisiva), manipolato (ad esempio ricodificato per caricarlo sulla pagina web dell'emittente) e scaricato da un utente, che lo elabora per manipolarne maliziosamente il contenuto per poi condividerlo sul social A, da cui viene potenzialmente ri-condiviso sul social B. La versione finale del video sul social B si porterà dietro **una serie di tracce** (dati, metadati, forma, contenuto) lasciate dai vari passaggi subiti che, almeno teoricamente, potrebbero consentire di ricostruirne il ciclo di vita a partire dalla creazione.

Allo stato attuale, la ricerca ha portato alla definizione di soluzioni valide per la rilevazione di molte di queste tracce quali, ad esempio, quelle lasciate dal sistema di acquisizione, la codifica e ri-codifica del segnale, alcune tecniche comuni di elaborazione, alcuni meccanismi di condivisione su social. Molto più complesso (e, al momento, risolto solo in piccola parte) è il problema di valutare queste cose all'interno di una catena di elaborazione come quella sopra esemplificata, in cui ogni passaggio può in parte coprire le tracce dei passaggi precedenti. Molte delle tecniche ideate ad oggi, inoltre, hanno il problema di uscire dalla realtà controllata dei laboratori di ricerca, per affrontare l'analisi *in the wild*, dove gli scenari diventano estremamente complessi e variabili.

L’estensione dell’**analisi forense** a questi nuovi scenari richiede l’abilità di affrontare problemi tecnologici significativi, richiedendo metodologie che possano lavorare in maniera affidabile in condizioni molto generali. D’altra parte, la capacità di affrontare problemi quali il recupero di informazioni sul ciclo di vita del dato multimediale, in termini di provenienza, manipolazioni e condivisioni subite, rappresenterebbe un supporto fondamentale per i servizi di *intelligence*, per la polizia postale e per tutti gli attori preposti a tracciare contenuti maliziosi. In generale, questi strumenti potranno aiutare a garantire la veridicità dei contenuti multimediali, ripristinando il tradizionale concetto di affidabilità legato all’informazione visiva.

Riferimenti

[1] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: *Analyzing and Improving the Image Quality of StyleGAN*, arXiv:1912.04958 (2019)

[2] Verdoliva, L.: *Media forensics and deepfakes: an overview*. IEEE Journal of Selected Topics in Signal Processing, vol 14, n. 5 (2020)

[3] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: *Protecting world leaders against deep fakes. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2019)

[4] Marra, F., Saltori, C., Boato, G., Verdoliva, L.: *Incremental learning for the detection and classification of GAN-generated images*. IEEE International Workshop on Information Forensics and Security (2019)

[5] Cao, J., Qi, P., Sheng, Q., Yang, T., Guo, J., Li, J.: *Exploring the role of visual content in fake news detection. Disinformation, Misinformation, and Fake News in Social Media*, (2020)

[6] Phan, Q., Boato, G., Caldelli, R., Amerini, I.: *Tracking multiple image sharing on social networks*. IEEE International Conference on Acoustics, Speech and Signal Processing (2019)

Articolo a cura di **Giulia Boato** e **Francesco De Natale**