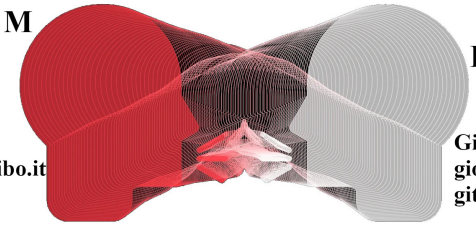
Deepfake:
A Game of Cat and Mouse

CYBERSECURITY M
REPORT

INGEGNERIA
INFORMATICA



Marco Motamed
marco.motamed@studio.unibo.it
github.com/MotaMarco

Giorgio Mocci
giorgio.mocci@studio.unibo.it
github.com/giorgio-mocci

ABSTRACT

In questo report verranno trattate le tematiche tecniche e sociali relative all'utilizzo dei deepfakes nella società odierna.
Verrà inoltre presentato un modello CNN, capace di individuare gli audio fake, da noi sviluppato.
Infine sarà presentata una possibile soluzione all'uso improprio dei deepfakes.

Professore:
Michele Colajanni

Dicembre , 2022

1 Introduzione

Il seguente report ha come obiettivo l'analisi e lo studio dello stato dell'arte relativo alle tecniche, di cybersicurezza, per stabilire l'autenticità di un audio e/o video.

Tramite l'utilizzo di reti neurali convoluzionali e dell'intelligenza artificiale si è ormai in grado di generare riproduzioni video e/o audio estremamente fedeli alla realtà. Fino a qualche anno fa tali strumenti risultavano essere accessibili solo ad un' "elite" di appassionati e ricercatori con accesso a macchine prestanti e con un'ampia conoscenza dell'argomento. Attualmente, in seguito al boom di queste tecnologie, queste tecniche sono applicabili da chiunque.

Nonostante queste tecnologie possano essere incredibilmente utili in molti ambiti, ad esempio medico, riabilitativo e sociale, è necessario interrogarsi sui problemi che un utilizzo improprio e scorretto possa causare nella società odierna.
Infatti, è possibile sintetizzare delle riproduzioni accurate senza il consenso dell'interessato andando a rubarne l'identità e utilizzandole per fini criminosi. Queste riproduzioni risultano quasi impossibili da distinguere per l'utente comune e necessitano di ausili tecnologici esterni per essere individuate.
Per queste ragioni negli ultimi anni sono nati diversi modelli che, sfruttando il machine learning e l'intelligenza artificiale, permettono di riconoscere una riproduzione generata artificialmente.

Il documento è diviso in tre sezioni sintetizzate come di seguito:

1. Che cos'è un deepfake? Problemi nella società odierna
2. Come individuare un deepfake (casi studio)
3. Sperimentazione e lavoro svolto
4. Conclusioni e considerazioni finali

2 Usi e abusi del Deepfake

Il termine deepfake è utilizzato per indicare una riproduzione digitale (foto, video, audio), estremamente realistica, di una persona, animale, macchina o oggetto animato creata al fine di ingannare il fruitore rappresentando una realtà falsa costruita artificialmente.

La parola deepfake è stata utilizzata per la prima volta nel 2017 da un utente di Reddit per riferirsi al software, da lui utilizzato, per creare filmati che partendo da video esistenti a sfondo sessuale andavano a modificare il volto dell'attrice con quello di una celebrità[1].

Figure 1: Frame del film "The Book of Boba Fett" rappresentante una riproduzione deepfake dell'attore Mark Hamill nelle vesti del giovane Luke Skywalker

Da allora il termine deepfake si è evoluto e ha anche assunto una parziale accezione positiva. Infatti, grazie a questa tecnologia è possibile creare materiale a supporto di numerosi nobili settori, come ad esempio quello medico, cinematografico ed educativo.[2]
Un altro esempio rilevante può essere l'assistenza alle persone che, in seguito alla perdita di un proprio caro, possano giovare di queste riproduzioni per aiutarle a superare il loro trauma[3].

Purtroppo però la stragrande maggioranza dei deepfake sono utilizzati in maniera dannosa, con lo scopo di screditare un personaggio pubblico, di mortificare e umiliare un individuo (solitamente tramite video pornografici che colpiscono maggiormente target femminili) o deviare l'opinione pubblica verso un determinato obiettivo in maniera simile a quanto avviene con le fakenews.
Un possibile esempio di come questi utilizzi possano diventare di importanza internazionale è quanto accaduto nel 2019 alla allora "house speaker" Nancy Pelosi, dove un video di questo tipo la ritraeva visibilmente alterata come se fosse in preda ai fumi dell'alcol[4].

Appare evidente come un utilizzo scorretto di queste tecnologie possa provocare ingenti danni economici e sociali alla collettività. Risulta quindi imperativa una maggiore attenzione da parte del governo e dell'opinione pubblica sull'argomento, al fine di sensibilizzare la popolazione e le aziende del settore. In questo modo si potrà combattere questa crescente piaga.

3 Come possiamo individuare un deepfake?

In questo capitolo verranno trattate le tecniche che possono essere usate per distinguere un video genuino da uno fasullo.
La prima categoria riguarda le metodologie basate sul riconoscimento di difetti e pattern comuni ai video deepfake.

Ad esempio una caratteristica comune ai suddetti video è la gestione innaturale degli occhi. Infatti, i soggetti sbattono le palpebre troppo spesso, troppo di rado o in casi estremi non le sbattono affatto.
Questo e altri simili problemi possono essere utili ad individuare riproduzioni sbrigative o poco curate. Purtroppo quelle di ultima generazione non presentano più queste problematiche ed è quindi necessario affidarsi a strumenti più sofisticati per raggiungere lo scopo.

Tramite l'utilizzo di tecniche basate su machine learning e intelligenza artificiale è possibile realizzare modelli in grado di individuare, con un certo grado di attendibilità, la maggior parte delle riproduzioni deepfake. Ultimamente, però, sono emerse nuove metodologie in grado di ingannare questi strumenti.
Un celebre esempio è GANs, una tipologia di rete neurale in grado di riconoscere automaticamente le possibili falle e migliorare la generazione del deepfake attraverso molteplici round, rendendo estremamente difficile l' individuazione da parte di software appositi.

Ovviamente anche gli esperti di cybersecurity hanno migliorato i propri sistemi, attraverso nuove tecniche che vanno a concentrarsi su parametri più specifici e precisi che attualmente risultano estremamente complessi da emulare.

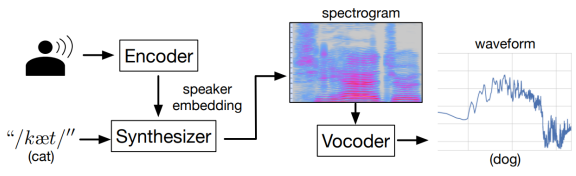


Figure 2

Un esempio virtuoso è il software sviluppato dai ricercatori dell'università della Florida, Gainesville in grado di riconoscere i deepfake audio utilizzando una tecnica basata sull'analisi dell' impronta generata dalle corde vocali [5].

Un'altra interessante tecnologia è quella sviluppata da Intel, in grado di individuare i deepfake video tramite deep-learning e lo studio dei flussi sanguigni attraverso i vasi presenti nella testa.
Intel dichiara un'accuratezza di individuazione del 96% e promette di generare un risultato in pochi millisecondi, rendendo il servizio utilizzabile anche per l'analisi real-time di un video.[6]

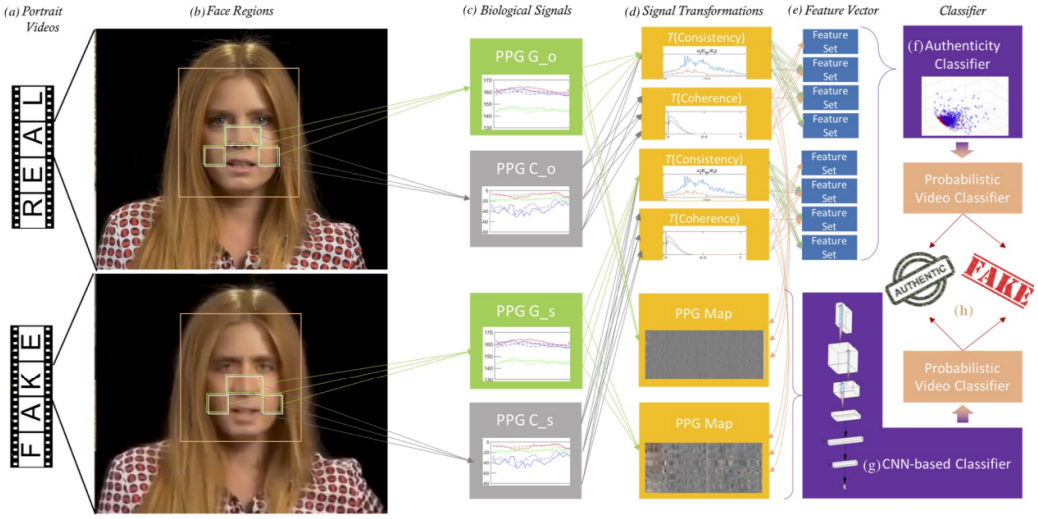


Figure 3: Partendo da una coppia reale e fake di frame(a) sono stati estratti segnali biologici (c) da diverse regioni facciali (b). Sono state applicate transformazioni (d) per calcolare coerenza spaziale e consistenza temporale, catturando segnali caratteristici(e) in un set di label e PPG maps e utilizzandoli per addestrare modelli di apprendimento SVM (f) e CNN (g). Infine vengono calcolate le probabilità (h) per stabilire l'autenticità dei frame iniziali (a).

Purtroppo questi software non sono stati ancora resi disponibili e risultano impossibili da testare. Per questo motivo nel prossimo capitolo verrà trattato il nostro personale approccio a questa problematica.

4 Il nostro modello per il deepfake's audio detection

In questa sezione verrà presentato il nostro modello di individuazione degli audio deepfake.
Per le motivazioni trattate nel precedente capitolo abbiamo deciso di sviluppare uno strumento che ci permettesse di effettuare test personalizzati approfondendo la problematica da un punto di vista più pratico.
A tal fine abbiamo addestrato una rete neurale convoluzionale. Questa tecnologia, emulando il funzionamento del nostro cervello e partendo da una base di dati, riesce attraverso diversi livelli a predire un risultato.

Per quanto riguarda la base di dati ci siamo affidati ai dataset di ASVspoof per gli audio fake [7] e a Common Voice per gli audio genuini [8]. Per poter utilizzare questi file è stato necessario estrarre dagli audio i coefficienti Mel-frequency cestral (rappresentazione spettro di potenza a breve termine di un suono)

Procedura di estrazione MFCC

```
def features_extractor(file_name):
    audio, sample_rate = librosa.load(file_name, res_type= 'kaiser_fast')
    mfccs_features = librosa.feature.mfcc(y=audio, sr=sample_rate, n_mfcc=40)
    mfccs_scaled_features = np.mean(mfccs_features.T,axis=0)
    return mfccs_scaled_features
```

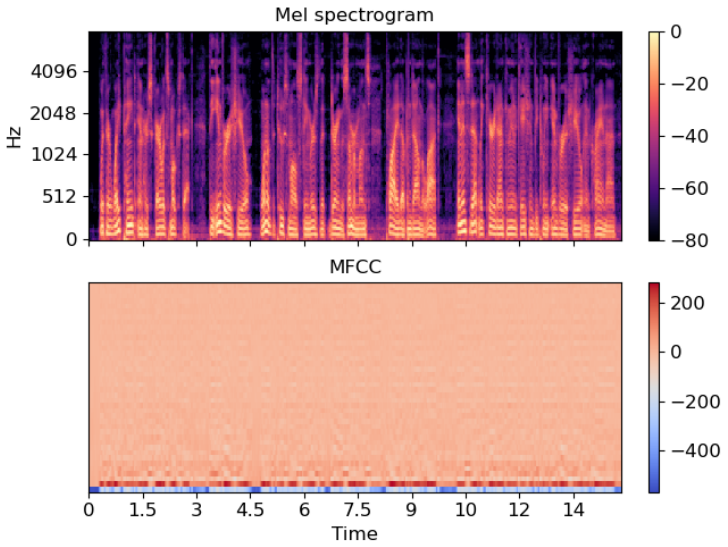
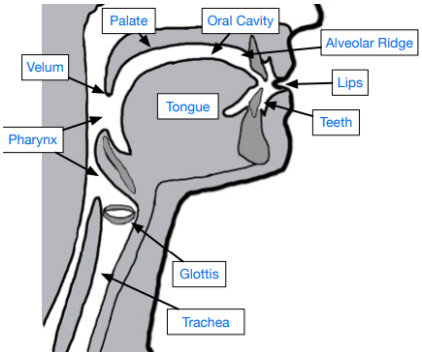


Figure 4: Esempio di spettrogramma Mel e relativo MFCC

Per il training del modello ci siamo affidati al framework di tensorflow.keras utilizzando una gpu Geforce Rtx 2060. Dopo una serie di test abbiamo constato che l'addestramento sequenziale fosse il più adatto alle nostre esigenze. Infatti, tale metodologia ci ha permesso di sfruttare al meglio le informazioni in nostro possesso ottenendo una maggiore accuratezza e precisione.
Di seguito vengono riportati i parametri che riteniamo essere più rilevanti ed i risultati ottenuti a seguito dell'addestramento della rete.

Parametri utilizzati nell'addestramento

```
model.compile(loss='categorical_crossentropy',metrics=['accuracy'],optimizer='adam')
lr_reduction = ReduceLROnPlateau(monitor = 'val_loss',
                                 patience = 3,
                                 verbose = 1,
                                 factor = 0.2,
                                 min_lr = 0.001)
callbacks = [lr_reduction]
num_epochs = 50
num_batch_size = 32
history = model.fit(X_train,
          y_train,
          batch_size=num_batch_size,
          epochs=num_epochs,
          validation_data=(X_val, y_val),
          callbacks=callbacks,
          verbose=1)
```



I risultati restituiti si basano sulle divergenze esistenti tra gli audio reali, che sono limitati dalla struttura organica delle corde vocali, e gli audio deepfake *GAN-generated* che non hanno queste limitazioni. Il tratto vocale è composto da vari componenti che agiscono insieme per produrre un suono. I distinti suoni sono articolati in base al percorso dell'aria, determinato da come sono posizionati i vari componenti [5]. Tali tratti non posso essere replicati facilmente da parte di un'intelligenza artificiale.

In conclusione ci riteniamo soddisfatti dei risultati ottenuti. Abbiamo scelto di rendere diffidente il nostro riconoscitore, in caso non sia certo del risultato predetto, in modo da evitare il più possibile i falsi positivi.
Il nostro modello riesce a raggiungere una precisione superiore al 95% con gli audio appartenenti alla stessa tipologia dei dataset utilizzati durante l'addestramento; mentre ha più difficoltà a riconoscere i deepfake generati con altre tecnologie.
Questa problematica verrà approfondita nel prossimo ed ultimo capitolo.

5 Conclusioni

Come anticipato nel capitolo precedente i riconoscitori, basati sul ML come il nostro, presentano diverse problematiche.
Come afferma uno studio pubblicato su software engeneering institute [9] la maggior parte dei deepfake detector, nonostante affermi di avere un accuratezza di oltre il 99%, di fatto ha un success rate che oscilla tra il 30% e il 97%. Le difficoltà principali nell'individuazione dei video deepfake dipendono da diversi fattori come il livello di compressione delle immagini, la loro risoluzione e le caratteristiche del dataset utilizzato durante l'addestramento.

Un esperimento condotto congiuntamente dalle università di Berlino e Amsterdam [10] dimostra che la maggior parte delle persone intervistate non è assolutamente in grado di riconoscere un deepfake. Inoltre, queste ultime tendono a sovrastimare le proprie capacità di riconoscimento e, se non sicure, vanno a identificare i deepfakes come veri piuttosto che viceversa.
La non capacità di individuare un deepfake non appare quindi come una mancanza di attenzione o motivazione, ma come una questione di incapacità.

Quanto detto, nel prossimo futuro, ci obbligherà a non poterci più fidare dei classici strumenti di comunicazione ,come video e audio, e andrà a minare la nostra capacità di informazione.
Personalmente, pensiamo che una possibile soluzione possa essere raggiunta tramite uno sforzo congiunto di impegno e tecnica. Infatti, a fianco degli ingegneri e dei loro software, è importante che l'intera società impari a sensibilizzarsi sull'argomento; andando a fidarsi solo delle fonti considerate attendibili.
In questo modo riusciremo a limitare i lati negativi dei deepfake arrivando infine a coesistere con essi e a beneficiare dei loro aspetti positivi.

References

[1] Origine deepfake. https://www.cybersecurity360.it/nuove-minacce/deepfake-in-tempo-reale-cosa-sono-come-funzionano-e-quali-tutele-per-prevenire-la-minaccia.

[2] Luke deepfake. https://www.gq-magazine.co.uk/culture/article/boba-fett-luke-skywalker.

[3] Superare un trauma. https://www.wired.com/story/deepfake-death-grief-hologram-photography-film/.

[4] Video fake di nancy pelosi. https://www.washingtonpost.com/technology/2020/08/03/nancy-pelosi-fake-video-facebook/.

[5] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, and Patrick Traynor. Who are you (i really wanna know)? detecting audio DeepFakes through vocal tract reconstruction. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2691–2708, Boston, MA, August 2022. USENIX Association.

[6] Umur Aybars Ciftci and Ilke Demir. Fakecatcher: Detection of synthetic portrait videos using biological signals. *CoRR*, abs/1901.02212, 2019.

[7] Fake audio dataset. https://www.asvspoof.org/index2021.html.

[8] Real audio dataset. https://commonvoice.mozilla.org/it/datasets.

[9] C. Bernaciak and D. Ross. How easy is it to make and detect a deepfake? Carnegie Mellon University's Software Engineering Institute Blog, Mar. 14, 2022. [Online].

[10] Nils C. Köbis, Barbora Doležalová, and Ivan Soraperra. Fooled twice: People cannot detect deepfakes but think they can. *iScience*, 24(11):103364, 2021.