Emotions Don’t Lie: An Audio-Visual Deepfake Detection Method using Affective Cues

Concetto teorico di reti neurali siamesi

Le reti neurali siamesi sono un tipo di architettura di rete neurale che sono progettate per apprendere rappresentazioni simili per input simili. Questo tipo di rete neurale è comunemente utilizzato per problemi di confronto o misura di similarità tra coppie di dati.

L'architettura di una rete neurale siamese si compone di due o più sotto-reti identiche (noti come "bracci siamesi") che condividono gli stessi pesi e le stesse architetture. Questi bracci siamesi prendono in input coppie di dati simili (ad esempio due immagini di volti) e generano rappresentazioni latenti o embedding di questi dati. Le rappresentazioni latenti sono quindi confrontate per misurare la loro similarità o distanza.

Durante l'addestramento, le coppie di dati di input vengono fornite ai bracci siamesi e le loro rappresentazioni latenti vengono generate. Successivamente, viene calcolata una misura di similarità o distanza tra le rappresentazioni latenti, ad esempio utilizzando la distanza euclidea o la distanza coseno(cioè il nostro triplet loss, lo spiego nei paragrafi). Questo valore di similarità o distanza viene quindi utilizzato come base per calcolare una perdita (loss) che viene retropropagata attraverso la rete per aggiornare i pesi e ottimizzare l'apprendimento delle rappresentazioni.

Durante la fase di inferenza, la rete neurale siamese può prendere in input una singola coppia di dati e generare le rispettive rappresentazioni latenti. Queste rappresentazioni possono quindi essere utilizzate per calcolare la similarità o distanza tra i due dati, ad esempio per il riconoscimento di pattern o la misura di similarità tra due oggetti. Questo tipo di architettura è particolarmente utile quando è necessario confrontare coppie di dati e misurare la loro similarità o distanza in uno spazio di rappresentazione appreso.

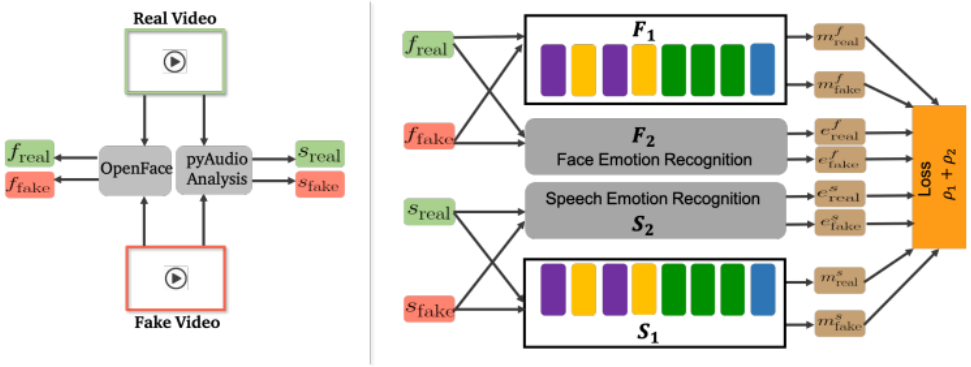
Introduzione

Ci sono molte altre applicazioni di elaborazione video che utilizzano e combinano modalità multiple per il riconoscimento audio-visivo del linguaggio del corpo, il riconoscimento delle emozioni, e compiti di linguaggio e visione. Queste applicazioni mostrano che la combinazione di modalità multiple può fornire informazioni complementari e portare a inferenze più accurate. Anche per la rilevazione dei contenuti deepfake, è possibile estrarre molte modalità come le indicazioni facciali, le indicazioni vocali, il contesto di sfondo, i gesti delle mani e la postura del corpo da un video. Quando combinate, molteplici indicazioni o modalità possono essere utilizzate per rilevare se un determinato video è reale o falso.

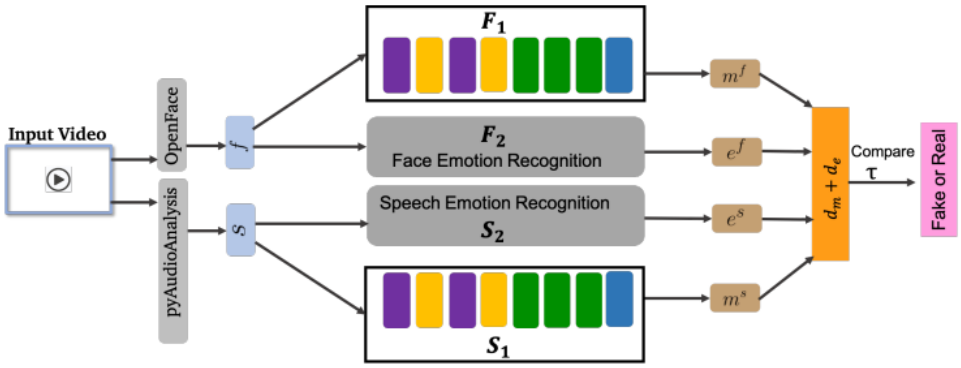
Studi hanno evidenziato una forte correlazione tra diverse modalità dello stesso soggetto [56]. Più specificamente, suggeriscono una correlazione positiva tra le modalità audio-visive, che sono state sfruttate per il riconoscimento multimodale delle emozioni percepite. Ad esempio, suggerisce che quando diverse modalità vengono modellate e proiettate in uno spazio comune, dovrebbero indicare indicazioni affettive simili. Le indicazioni affettive sono caratteristiche specifiche che trasmettono ricche informazioni emotive e comportamentali agli osservatori umani e li aiutano a distinguere tra diverse emozioni percepite. Questi segnali affettivi includono diverse caratteristiche di posizione e movimento, come la dilatazione degli occhi, sopracciglia alzate, volume, velocità e tono della voce. Sfruttiamo questa correlazione tra modalità e segnali affettivi per classificare video "veri" e "falsi".

Il loro lavoro:
Presentiamo un approccio innovativo che sfrutta contemporaneamente le modalità audio (parole) e video (volto) e le caratteristiche delle emozioni percepite estratte da entrambe le modalità per rilevare eventuali falsificazioni o alterazioni nel video in ingresso. Per modellare queste caratteristiche multimodali e le emozioni percepite, il nostro metodo di apprendimento utilizza un'architettura basata su una rete Siamese. Durante il training, passiamo un video reale insieme al suo deepfake attraverso la nostra rete e otteniamo vettori di incorporamento per il volto e la voce del soggetto, relativi alle modalità e alle emozioni percepite. Utilizziamo questi vettori di incorporamento per calcolare la funzione di perdita al fine di minimizzare la somiglianza tra le modalità del video falso e massimizzare la somiglianza tra le modalità del video reale.

APPROCCIO:
overview
Il nostro obiettivo è rilevare se un video è una deepfake o meno, utilizzando un input video con modalità audio-visive. Una panoramica dei nostri routine di addestramento e test è illustrata in Figura 1a e Figura 1b, rispettivamente.



(a) Training Routine: *(left)* We extract facial and speech features from the raw videos (each subject has a real and fake video pair) using OpenFace and pyAudioAnalysis, respectively. *(right)* The extracted features are passed to the training network that consists of two modality embedding networks and two perceived emotion embedding networks.



(b) Testing Routine: At runtime, given an input video, our network predicts the label (real or fake).

Durante l'addestramento, selezioniamo un video "reale" e uno "falso" che contengono lo stesso soggetto. Estraiamo le caratteristiche visive del viso e le caratteristiche del discorso, indicate rispettivamente come $f$real e $s$real, dal video di input reale. In modo simile, estraiamo le caratteristiche del viso e del discorso (utilizzando OpenFace e pyAudioAnalysis), indicate rispettivamente come $f$fake e $s$fake, dal video falso. Le caratteristiche estratte, $f$real, $s$real, $f$fake, $s$fake, formano gli input alle reti neurali ($F1$, $F2$, $S1$ e $S2$), rispettivamente. Alleniamo queste reti utilizzando una combinazione di due funzioni di perdita triplet basate

sulla similarità, indicate come $\rho 1$ e $\rho 2$. $\rho 1$ rappresenta la similarità tra le modalità del viso e del discorso, mentre $\rho 2$ è la similarità tra gli indizi dell'affetto (specificamente, l'emozione percepita) dalle modalità di entrambi i video reali e falsi. Il nostro metodo di addestramento è simile a una rete Siamese perché utilizziamo gli stessi pesi della rete ($F1$, $F2$, $S1$, $S2$) per elaborare due input diversi, un video reale e un video falso dello stesso soggetto. A differenza delle normali reti neurali basate sulla classificazione, che eseguono la classificazione e propagano tale perdita all'indietro, utilizziamo invece metriche basate sulla similarità per distinguere i video reali e falsi. Se si lavora sulla similarità ovvimente Modelliamo questa similarità tra queste modalità utilizzando la triplet. (La Triplet Loss richiede tre campioni di dati: un'ancora (o esempio di query), un positivo (un esempio simile all'ancora. quindi trovare una rappresentazione dei dati in cui le distanze tra esempi simili siano ridotte e le distanze tra esempi dissimili siano aumentate. Ciò aiuta a creare un'incorporazione compatta e discriminativa dei dati) e un negativo (un esempio dissimile dall'ancora). L'obiettivo è minimizzare la distanza tra l'ancora e il positivo e massimizzare la distanza tra l'ancora e il negativo. Durante il test, ci viene fornito un singolo video di input, dal quale estraiamo i vettori delle caratteristiche del viso e del discorso, indicati rispettivamente come f e s. Passiamo f a $F1$ e $F2$ e passiamo s a $S1$ e $S2$, dove $F1$, $F2$, $S1$ e $S2$ vengono utilizzati per calcolare le metriche di distanza, indicate come $dist1$ e $dist2$. Utilizziamo una soglia τ, appresa durante l'addestramento, per classificare il video come reale o falso. F1 e S1 sono reti di incorporamento di modalità e F2 e S2 sono reti di incorporamento delle emozioni percepite per il volto e la voce, rispettivamente.

Training routine

Durante il tempo di addestramento, utilizziamo un video falso e un video reale con lo stesso soggetto come input. Successivamente, passiamo le caratteristiche estratte dai video grezzi (freal, ffake, sreal, sfake) attraverso F1, F2, S1 e S2, ottenendo le incapsulazioni di modalità e emozione percepite normalizzate. Considerando un video reale e uno falso in input, confrontiamo prima freal con ffake e sreal con sfake per capire quale modalità è stata manipolata di più nel video falso. Considerando ciò, identifichiamo la modalità del volto come quella manipolata di più nel video falso, basandoci su queste incapsulazioni calcoliamo la prima similarità tra le incapsulazioni reali e false del discorso e del volto, tramite delle formule ma sti cazzi. In termini più semplici, L1(cioè la formula) sta calcolando la distanza tra due coppie, d(ms_real, mf_real) e d(ms_real, mf_fake). Ci aspettiamo che ms_real e mf_real siano più vicini tra loro rispetto a ms_real e mf_fake, poiché contiene una modalità di volto falsa. Pertanto, ci aspettiamo di massimizzare questa differenza. Per utilizzare questa metrica di correlazione come funzione di perdita per addestrare il nostro modello, la formuliamo utilizzando la notazione di Triplet Loss.

Testing

Durante il test, abbiamo solo un singolo video di input che deve essere etichettato come reale o falso. Dopo l'estrazione delle caratteristiche, f e s dai video grezzi, eseguiamo un passaggio in avanti attraverso F1, F2, S1 e S2, come illustrato nella Figura 1b, per ottenere le rappresentazioni delle modalità e delle emozioni percepite.

Per fare un'infertenza su reali e falsi, calcoliamo i seguenti due valori di distanza:

Distanza 1: dm = d(mf, ms),

Distanza 2: de = d(ef, es).

Per distinguere tra reali e falsi, confrontiamo dm e de con una soglia, cioè τ, appresa empiricamente durante l'addestramento come segue: Se dm + de > τ, etichettiamo il video come un video falso.

Calcolo di τ: Per calcolare τ, utilizziamo il modello addestrato migliore e lo eseguiamo sull'insieme di addestramento. Calcoliamo dm e de sia per i video reali che falsi dell'insieme di addestramento. Facciamo la media di questi valori e troviamo un numero equidistante, che funge da buon valore di soglia. Sulla base dei nostri esperimenti, il valore di τ calcolato è risultato essere quasi consistente e non è variato molto tra i dataset.

Implementazione

Dataset ed i paramentri di training poco rilevanti al fine della comprensione degli algoritmi, quindi li ometto ma ci serviranno in un secondo momento perché saranno utili per migliorare le metriche e cercare di avere risultati migliori.

Feature Extraction

Per estrarre le caratteristiche del volto e della voce dai video di input reali e falsi. Utilizziamo metodi di stato dell'arte (SOTA) già esistenti a tale scopo. In particolare, utilizziamo OpenFace per estrarre caratteristiche facciali a 430 dimensioni, comprese le posizioni dei landmark 2D, l'orientamento del volto e le caratteristiche dello sguardo. Per estrarre le caratteristiche della voce, utilizziamo pyAudioAnalysis per estrarre 13 coefficienti cepstrali delle frequenze mel (MFCC) come caratteristiche della voce.

Conclusioni

Presentiamo un metodo basato sull'apprendimento per rilevare video falsi. Utilizziamo la somiglianza tra le modalità audio-visive e la somiglianza tra le indicazioni affettive delle due modalità per inferire se un video è "reale" o "falso". Abbiamo valutato il nostro metodo su due set di dati di deepfake audio-visivi di riferimento, DFDC e DF-TIMIT.

Il nostro approccio ha alcune limitazioni. In primo luogo, potrebbe risultare in classificazioni errate su entrambi i set di dati, rispetto a quelli presenti nei video reali. Dato che esistono diverse rappresentazioni per l'espressione di emozioni percepite, il nostro approccio potrebbe trovare una discrepanza nelle modalità dei video reali e (in modo errato) classificarli come falsi. Inoltre, molti dei set di dati di deepfake contengono principalmente più di una persona per video. Potremmo dover estendere il nostro approccio per considerare lo stato emotivo percepito di più persone nel video e sviluppare uno schema possibile per il rilevamento di deepfake.

In futuro, vorremmo esaminare la possibilità di incorporare più modalità e persino il contesto per inferire se un video è un deepfake o meno. Vorremmo inoltre combinare il nostro approccio con le idee esistenti per il rilevamento di artefatti visivi come la sincronizzazione labiale, l'orientamento della posa della testa e gli artefatti specifici in denti, naso e occhi attraverso i fotogrammi per una migliore performance. Inoltre, vorremmo sviluppare metodi migliori per utilizzare le indicazioni audio.