Deepfake detection through facial dynamics

Corso di Fondamenti di Visione Artificiale e Biometria

2024/2025

Dott.ssa Lucia Cimmino email: lcimmino@unisa.it

Dott. Benedetto Simone email: bsimone@unisa.it

Deepfake

- Con il termine deepfake ci si riferisce a contenuti audiovisivi generati sinteticamente, prodotti utilizzando tecniche di deep learning, come Generative Adversarial Networks (GAN), autoencoder o Stable Diffusion. Questi modelli consentono la simulazione realistica dell'aspetto facciale di un individuo target, dell'articolazione del parlato e delle dinamiche facciali, spesso con elevata fedeltà visiva.

Motivazioni

I deepfake rappresentano una minaccia significativa per l'integrità e l'affidabilità dei media digitali, consentendo:

- Impersonificazione dell'identità tramite rievocazione facciale o manipolazione del lip-sync.
- Campagne di disinformazione, in particolare in contesti politici o sociali.
- Attacchi ai sistemi di autenticazione biometrica (spoofing del volto o della voce).

Challenges in Detection

I moderni deepfake presentano caratteristiche come:

- Realismo visivo ad alta risoluzione, con artefatti di compressione minimi
- Aspetto coerente per ogni frame, che rende meno efficace il rilevamento delle immagini statiche
- Coerenza temporale, che spesso richiede l'analisi di sottili incongruenze spaziotemporali, specialmente in regioni localizzate come occhi, bocca o mascella

Case Study

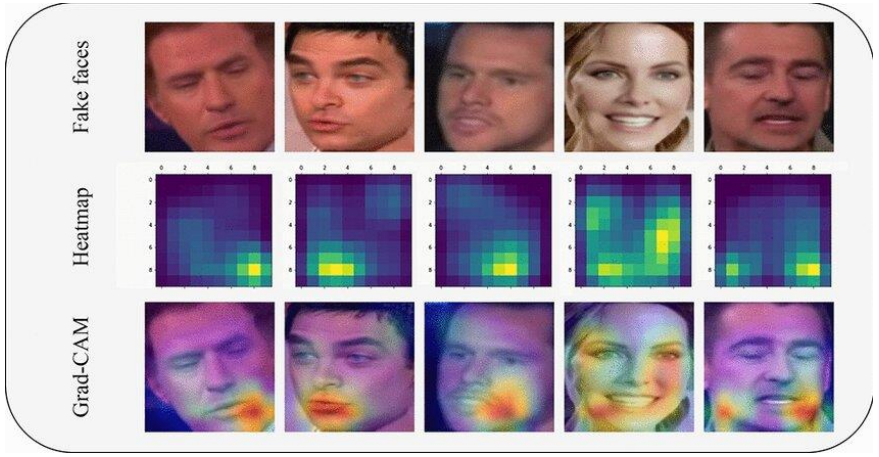
- **OBIETTIVO:** Riconoscere l'autenticità di un video dalla dinamica del volto durante la pronuncia di frasi che inducono a dei movimenti facciali.

- **IDEA:** Sintetizzare le informazioni spazio-temporali del volto per identificare anomalie nei movimenti facciali che risultano difficili da riprodurre in modo realistico nei video deepfake.

- **PROPOSTA:** Utilizzare modelli di deeplearning che tengono conto della sequenza temporale dei video per poter distinguere un video reale da un video fake.

Motivazioni

- Le dinamiche temporali del movimento facciale, specialmente nelle regioni coinvolte nel linguaggio (ad esempio, movimenti labiali, coordinazione dei muscoli periorali), sono difficili da modellare in modo convincente in contenuti sintetizzati. Sfruttare queste dinamiche offre una direzione promettente per sistemi di rilevamento deepfake robusti e generalizzabili.

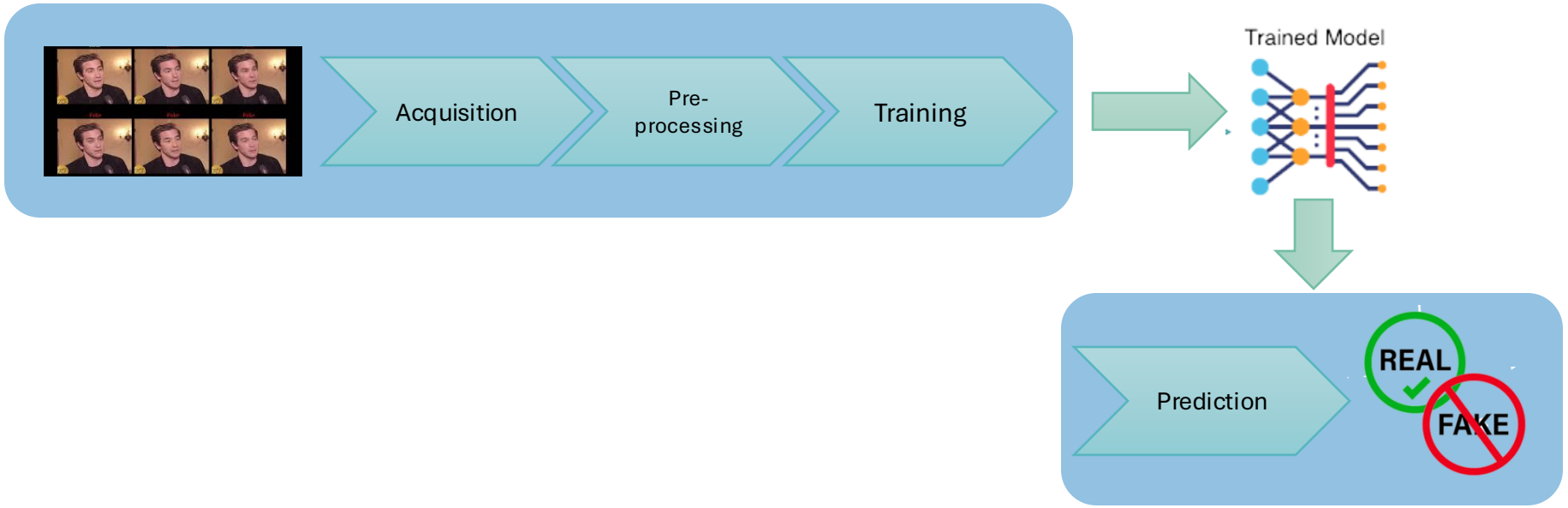


Contesti applicativi

- Verifica dell’autenticità nei media digitali
- Sicurezza e **biometria forense**
- Contrasto alla disinformazione audiovisiva

How to..

- Utilizzo di modelli di deep learning che prendano in considerazione la componente temporale per estrarre le caratteristiche temporali dai video

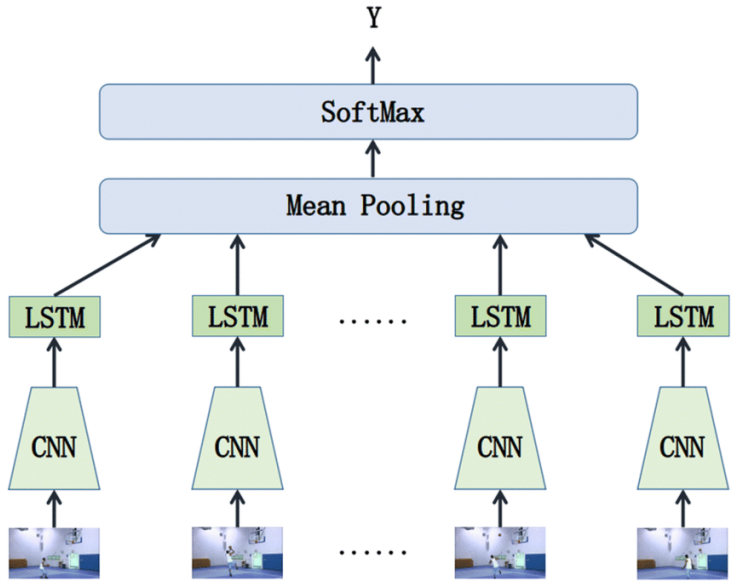


Recurrent Neural Network (RNN)

- Una rete neurale ricorrente (o RNN da Recurrent Neural Network) è una rete neurale in cui esistono cicli:
  - i valori di uscita di un layer di livello superiore (più vicino all'uscita) vengono utilizzati come ingresso per un layer di livello inferiore (più vicino all'ingresso).
- Esistono diversi modelli di RNN:
  - LSTM
  - GRU
- Una LSTM (Long Short-Term Memory) è un tipo di RNN progettata per modellare dipendenze temporali a lungo raggio in dati sequenziali. A differenza delle RNN standard, le LSTM incorporano meccanismi di gating (porte di input, di forget e di output) che regolano il flusso di informazioni, mitigando problemi come vanishing/exploding gradient.

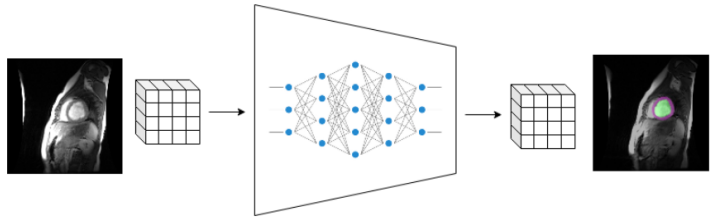
Approccio 1: CNN-LSTM

- Utilizzo di un'architettura CNN-LSTM come possibile soluzione per modellare in maniera efficace i **movimenti naturali del volto**
- L'architettura **CNN-LSTM** è una rete neurale ibrida che combina due componenti complementari:
  - **CNN (Convolutional Neural Network):** utilizzata per l'estrazione automatica di **caratteristiche spaziali** da ogni frame del video (es. struttura e texture del volto).
  - **LSTM:** rete neuralea ricorrente progettata per apprendere **dipendenze temporali** all'interno di sequenze, in grado di modellare l'evoluzione dei **movimenti facciali** nel tempo.

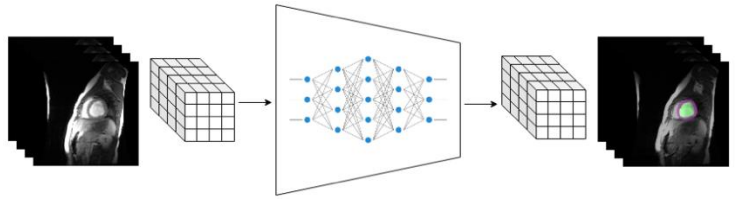


Approccio 2: 3DCNN

- Utilizzo di una rete 3D-CNN come possibile soluzione per apprendere simultaneamente caratteristiche spaziali e temporali dai video facciali
- L'architettura **3D-CNN** estende le tradizionali reti convoluzionali 2D introducendo un asse temporale, permettendo l'elaborazione diretta di **volumi video** (frame × altezza × larghezza).
- Le 3DCNN usano filtri convoluzionali tridimensionali che vengono applicati su **blocchi temporali di frame**, catturando **pattern dinamici locali** che coinvolgono sia la struttura del volto che il suo movimento nel tempo.



2D-CNN



3D-CNN

Environment e Librerie

AMBIENTI DI SVILUPPO

IMAGE E VIDEO PROCESSING

DEEP LEARNING

CASALab

Materiale da produrre

- Archivio contente il codice prodotto (.py o .ipynb) opportunamente commentato e corredato di tutti i files utili all'esecuzione
- Short-paper strutturato come di seguito:
  1. Abstract
  2. Introduzione e descrizione del problema
  3. Stato dell'arte/Related Works
  4. Sistema proposto (sezione in cui esporre la strategia implementativa adottata)
  5. Risultati sperimentali (sezione in cui vengono presentati i risultati raggiunti in termini di metriche come accuracy, F1-score, MSE etc..)
  6. Discussione dei risultati (sezione in cui discutere e analizzare i risultati ottenuti)
  7. Conclusioni
- Presentazione finale

Consigli Utili

- E’ strettamente consigliato l’utilizzo di strumenti di condivisione del codice (e.s. file jupyter, github…) in modo da agevolare eventuali correzioni in corso d’opera.
- E’ consigliabile utilizzare Google Colab in quanto i blocchi note colab possono eseguire il codice sui server cloud di Google consentendovi di utilizzare la potenza computazionale di Google (RAM, GPU e TPU), a prescindere dalla vostra macchina.
- E’ consigliabile inviare il materiale (short paper e codice) almeno 5 giorni prima della data d’esame in modo da avere il tempo di apportare eventuali correzioni.
- Per qualsiasi dubbio non esitate a contattarmi.

CASALab

GRAZIE A TUTTI E BUON LAVORO!

Dott.ssa Lucia Cimmino email: lcimmino@unisa.it

Dott. Benedetto Simone email: bsimone@unisa.it