Università degli Studi di Camerino

Scuola di Scienze e Tecnologie

Corso di Laurea in Informatica (Classe L-31)

Esplorazione del fenomeno dei deepfake generati dall'IA: analisi tecnica ed etica

Studente
Micarelli Simone

Mat. 104945

Relatore
Prof. Fausto Marcantoni

A.A. 2022/2023

Indice

1. Introduzione

False informazioni, disinformazione e propaganda sono caratteristiche della comunicazione umana almeno sin dai tempi dei Romani.

Quando Marco Antonio incontrò Cleopatra, Ottaviano condusse una campagna di propaganda contro di lui, pensata per diffamare la sua reputazione. Questa assunse la forma di *"slogan brevi e taglienti scritti su monete, nello stile di un Tweet arcaico"*[1]; i quali dipinsero Marco Antonio come un donnaiolo e un ubriaco, implicando che egli fosse diventato il burattino di Cleopatra, corrotto dalla sua relazione con lei. Successivamente, Ottaviano divenne Augusto, primo imperatore Romano.
Le *"fake news permisero a Ottaviano di hackerare il sistema repubblicano una volta per tutte"*[2].

Ad un passato più recente invece, esattamente nel 1888, risale il *primo filmato registrato della storia*[3], o per essere più precisi, il più vecchio al quale si può attualmente risalire. Nel 1898, dieci anni dopo, già si ebbero casi di disinformazione diffusa tramite video. La Edison Manufacturing Company voleva filmare la guerra ispano-americana, ma le goffe telecamere del XIX secolo lo rendevano impegnativo. La società di produzione ha scelto di intrecciare filmati reali di soldati ed armi in marcia, con filmati in scena di soldati americani che sconfiggono reggimenti nemici. Oscurando la verità dietro le quinte, le scene hanno alimentato il patriottismo tra i telespettatori americani.

Sebbene l'incidente del 1898 non sia stato necessariamente un **deepfake**, dimostra come la manipolazione dei dati possa diffondere intenzionalmente false informazioni.

1.1 Deepfake: di cosa parliamo?

Il vocabolario Treccani dà all'espressione *deepfake* il seguente significato: *"Filmato che presenta immagini corporee e facciali catturate in Internet, rielaborate e adattate a un contesto diverso da quello originario tramite un sofisticato algoritmo"*[4]. Il termine deriva *"dall'espressione inglese deepfake, che incrocia la locuzione s.le* ***deep learning*** *('insieme di tecniche che permettono all'Intelligenza artificiale di imparare a riconoscere le forme') con il s. fake ('falso, notizia falsa')"*[5].

[1] [Kam17]
[2] *ibidem.*
[3] [Rou]
[4] [Deeb]
[5] *ibidem.*

Sebbene il dizionario Treccani parli soltanto di filmato nella sua definizione, il fenomeno del deepfake non riguarda solo quelli. Nel provare a dare una definizione in senso più generale del termine, potremmo dire che: *"i deepfake sono foto, video e/o audio creati grazie a software di intelligenza artificiale (AI), i quali, partendo da contenuti reali (immagini e audio), riescono a modificare o ricreare, in modo estremamente credibile, le caratteristiche e i movimenti di un volto o di un corpo e imitare una determinata voce"*[6].

La **Computer Vision** (o visione artificiale) è l'ombrello accademico sotto al quale ricadono i deepfake; e il suo continuo sviluppo, non ha fatto altro che renderli popolari e più accessibili.

1.2 Abstract

L'obiettivo di questa tesi è quello di esplorare il fenomeno dei deepfake, ovvero la creazione di video manipolati mediante l'utilizzo di tecniche di intelligenza artificiale. Il lavoro si articola in sei capitoli, che affrontano la storia, le tecnologie e gli aspetti etico-sociali dei deepfake, nonché le strategie per riconoscerli e difendersi da essi.

Nel primo capitolo si definisce il termine "deepfake" e in quello successivo si analizza il contesto storico e tecnologico dal quale sono emersi. Nel capitolo dedicato agli aspetti tecnologici, si analizzano le tecniche di deep learning e di elaborazione delle immagini alla base dei deepfake. Nel quarto capitolo si esaminano i possibili impatti dei deepfake sulla società, suddividendo i casi d'uso in negativi e positivi, e illustrando le loro conseguenze nonché i possibili aspetti legali-normativi che li riguardano. Nel quinto capitolo vengono elencate alcune soluzioni software che è possibile utilizzare per la creazione di media deepfake e nel successivo, si descrive il processo di creazione di un deepfake mediante l'utilizzo del software DeepFaceLab e si analizzano le problematiche connesse alla sua realizzazione. Infine, nel settimo capitolo, si analizzano le tecniche di rilevamento dei deepfake e si forniscono alcune linee guida per difendersi da questi video manipolati. In sintesi, la tesi offre una panoramica completa e approfondita del fenomeno dei deepfake, analizzando le implicazioni tecnologiche, sociali ed etiche di questa tecnologia emergente.

[6] [Deea]

2. Background

2.1 New Tech, Old Concept

Per essere in grado di capire da dove nascono i deepfake, dobbiamo in primis esaminare gli accademici che hanno posato le sue basi.

Nel 1997 un articolo[1] scritto da Christoph Bregler, Michele Covell, e Malcolm Slaney sviluppò un programma innovativo e davvero unico che sostanzialmente automatizzò ciò che alcuni studi cinematografici potevano fare. Video Rewrite Program poteva sintetizzare nuove animazioni facciali da un output audio. Il programma era stato costruito sulle fondamenta di tecnologie già esistenti: interpretazione di volti, sintetizzazione di audio da testo, e modellazione di labbra nello spazio 3D; ma è stato il primo ad unire il tutto e animarlo in modo convincente.

Questa è stata una delle più importanti opere nello sviluppo dei deepfake. Infatti molti degli effetti video comuni di oggi che vengono raggruppati in programmi come *Adobe Premiere Pro* o *Final Cut* utilizzano filosofie algoritmiche aggiornate, basate su questo articolo.

I primi anni 2000, sono stati abbastanza silenziosi, poiché la computer vision (visione artificiale) si è spostata nell'approfondire il mondo del riconoscimento facciale. Gli sviluppi in questo campo hanno apportato drastici miglioramenti a cose come il motion tracking, rendendo i deepfake odierni più convincenti.

Active appearance models (AAM) è un algoritmo di computer vision, che ha debuttato in un articolo[2] di Timothy F. Cootes, Gareth J. Edwards e Christopher J. Taylor nel 2001. L'articolo ha mantenuto la sua popolarità dell'epoca. Usare un modello statistico approfondito per abbinare una forma ad un'immagine si è rivelato un grande passo in avanti, rendendo l'abbinamento e il riconoscimento facciale significativamente più efficiente.

Figure 2.1: AAM che trova i parametri facciali dopo 3 iterazioni e una posizione di partenza scarsa[CET01]

[1] [BCS97]
[2] [CET01]

Nel 2016 e nel 2017, uscirono due nuove ricerche scientifiche che rendevano possibile la realizzazione di deepfake con hardware di livello consumer. Rispettivamente, queste erano: il progetto Face2Face[3] dell'Università Tecnica di Monaco e il progetto Synthesizing Obama[4] dell'Università di Washington.

Sebbene completamente diversi negli obiettivi che stavano cercando di raggiungere, hanno drasticamente migliorato i tempi di elaborazione e rendering aggiornando la fedeltà grafica in modo da sembrare fotorealistica.

Face2Face permise di creare un'animazione in tempo reale, sostituendo l'area della bocca nel video target con quella di un "attore". Il programma non forniva alcun audio.

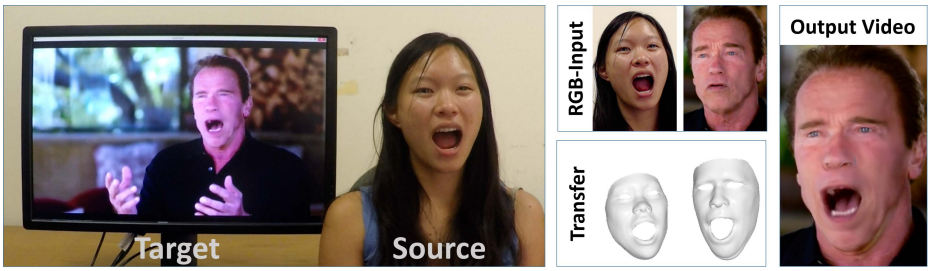

Figure 2.2: Funzionamento Face2Face[Thi+16]

Synthesizing Obama puntò invece ad essere un Video Rewrite[5] 2.0, con animazioni, texture ed espressioni migliorate. Ha aggiunto rughe e fossette e ha cambiato i colori per abbinare meglio l'illuminazione e il tono della pelle. Sebbene questi miglioramenti grafici forniscano effettivamente un modello più realistico, il più grande sviluppo di questo progetto è stata la sua capacità di alterare temporalmente sia l'audio che il video in modo convincente; nel senso, le sopracciglia del soggetto si muovono in base a ciò che sta dicendo. Non c'erano più momenti in cui il soggetto smetteva di parlare, ma le sue sopracciglia continuavano a muoversi.

Una delle più importanti invenzioni che ha posto la base tecnica dei deepfake come li conosciamo oggi, sono state le **GAN (Generative Adversarial Networks)**. Nel 2014, Ian Goodfellow, informatico e ricercatore, al tempo studente, pubblicò insieme ai suoi colleghi, una ricerca scientifica[6] che introduceva il concetto di GAN per la prima volta. Queste reti sono costituite da due agenti di intelligenza artificiale: uno falsifica un'immagine, l'altro cerca di rilevare il falso. Se l'agente scopre il falso, l'IA del falsario si adatta e migliora. In questo modo, entrambi gli agenti diventano sempre più efficienti nelle rispettive discipline nel corso della formazione, e le immagini generate diventano sempre più credibili.

L'uso di questa tecnologia di apprendimento automatico è stato limitato alla comunità scientifica di ricerca sull'intelligenza artificiale per diversi anni, fino a quando, nel 2017, non venne pubblicato da un utente di Reddit un fake video pornografico sotto il nome utente "deepfake". Fu la prima volta che questo termine venne utilizzato. Da allora il termine si è ampliato per includere le "applicazioni multimediali sintetiche" che esistevano già prima della pagina Reddit e tutte le future creazioni a riguardo.

[3] [Thi+16]
[4] [SSKS17]
[5] [BCS97]
[6] [Goo+14]

2.2 I deepfake di oggi

L'enorme picco di deepfake può essere in gran parte attribuito a Reddit e alla pornografia (come già introdotto sopra), portati a maggiore attenzione dagli articoli[7] di Samantha Cole su Vice. Un subreddit, ora cancellato, opportunamente chiamato r/deepfake, era nato dopo la pubblicazione dei suddetti articoli. Al momento del suo divieto, la comunità contava quasi 90.000 membri e presentava deepfake pornografici di una già vasta varietà di attori.

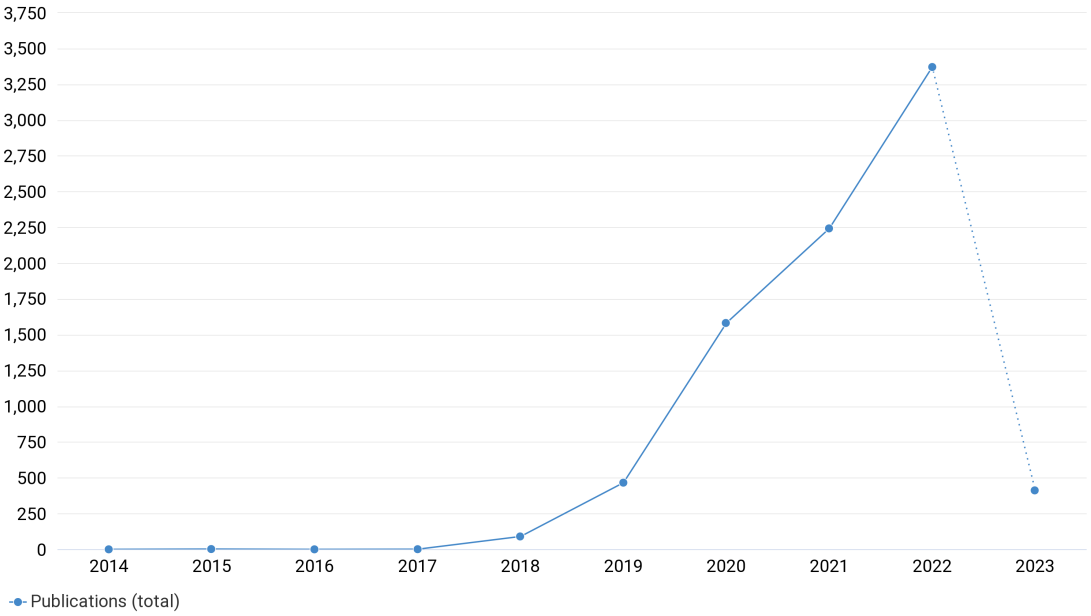

Figure 2.3: Il grafico mostra il numero di documentazioni relative al deepfake pubblicate ogni anno[Dim]

Attualmente, c'è una grande varietà di risorse pubbliche disponibili per lo sviluppo di deepfake. L'utente di Reddit u/deepfake citò la libreria Python Keras[8] e il progetto Github Tensorflow[9] come fonti per il suo software, ma ad oggi, ci sono molti altri progetti deepfake su Github, alcuni dei quali contenenti eseguibili pronti per l'uso immediato. È ormai facile dunque, anche per un dilettante, creare deepfake; l'unico ostacolo è la pazienza. Detto questo, i guadagni di efficienza in arrivo dallo sviluppo di hardware e software li renderanno solo più diffusi.

[7] [Col17; Col18]
[8] [Ker]
[9] [Ten]

3. Aspetti tecnologici

Come già introdotto in precedenza, il deepfake è una fenomeno che pone le sue basi sul campo dell'intelligenza artificiale. Più nello specifico, avevamo anche già detto sotto quale ombrello accademico i deepfake ricadessero, ovvero quello della computer vision.

Ai sistemi operanti in questa branca dell'intelligenza artificiale, viene assegnato il compito di capire ed interpretare il contenuto di un'immagine digitale, di una sequenza video o di un altro input visivo. Il tutto è reso possibile grazie all'avvento di tecniche sempre più avanzate di **machine learning**, e quindi, più nello specifico di **deep learning**.

Procediamo ora ad esaminare ognuno dei campi sopra citati in modo approfondito, con un occhio di riguardo anche per le tecnologie che operano dietro di questi e il loro funzionamento.

3.1 AI: Artificial intelligence

"Definiamo AI come lo studio di agenti che ricevono percezioni dall'ambiente ed eseguono azioni. Ciascuno di questi agenti implementa una funzione che associa alle sequenze di percezioni delle azioni"[1].

In senso lato, l'intelligenza artificiale comprende qualsiasi tecnica che consenta ai computer di imitare il comportamento umano e riprodurre o eccellere il suo processo decisionale, per risolvere compiti complessi in modo indipendente.

La capacità di tali sistemi per la risoluzione avanzata dei problemi, si basa su dei modelli analitici che generano previsioni, regole, risposte, raccomandazioni o risultati simili. I primi tentativi di costruire modelli analitici si basavano principalmente su istruzioni **hard-coded** in linguaggi formali, mediante le quali un computer poteva quindi ragionare automaticamente in base a regole di inferenza logica. Questo è anche noto come *approccio basato sulla conoscenza*[2].

Tuttavia, il paradigma ha subito riscontrato delle limitazioni, poiché i programmatori si sono presto resi conto di non riuscire ad esplicare tutta la loro conoscenza tacita sotto forma di linguaggio formale, necessario alla macchina per eseguire i suoi compiti. Le difficoltà affrontate dai sistemi IA basati su conoscenze hard-coded, hanno quindi suggerito la necessità, per i suddetti sistemi, di dover acquisire le proprie conoscenze da per sé, estraendo dei patterns dai dati grezzi. Questa capacità è nota come machine learning.

[1] [Rus10]
[2] [GBC16]

Negli ultimi decenni, il campo del machine learning ha prodotto una serie di notevoli progressi nei sofisticati algoritmi di apprendimento e nelle efficienti tecniche di pre-elaborazione. Uno di questi progressi è stata l'evoluzione delle **reti neurali artificiali (ANN, Artificial Neural Network)**, verso architetture di reti neurali sempre più profonde con capacità di apprendimento migliorate, riassunte come deep learning.

Riassumendo quindi, la distinzione tra AI, machine learning e deep learning può essere così rappresentata:

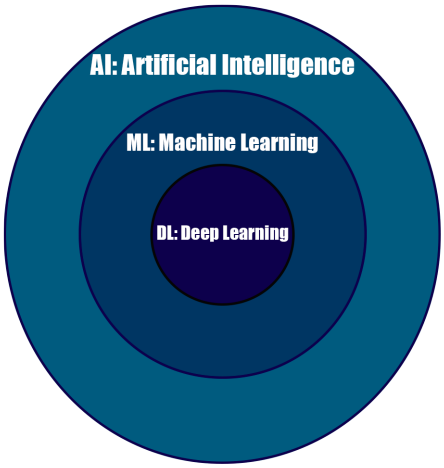

Figure 3.1: Distinzione tra AI, ML e DL

3.2 Machine learning

Il machine learning solleva l'essere umano dall'onere di spiegare e formalizzare la propria conoscenza in una forma accessibile alla macchina e consente di sviluppare sistemi intelligenti in modo più efficiente, automatizzando il processo di costruzione del modello analitico e consentendo alla macchina di prendere decisioni autonome rispetto al problema trattato. Ciò si ottiene applicando algoritmi che apprendono in modo iterativo dai dati di addestramento specifici del problema, consentendo così ai sistemi di trovare intuizioni nascoste e patterns complessi senza essere esplicitamente programmati.

Sulla base del problema dato e dei dati disponibili, possiamo distinguere tre tipi di machine learning: supervised learning, unsupervised learning e reinforcement learning.

1. **Supervised learning**: richiede un set di dati di addestramento che copra degli esempi pratici, nonché risposte o valori target per ciascuno di essi. In ogni esempio quindi, sono indicate le variabili di input (x) e il risultato corretto come output (y). La macchina imparando dagli esempi, elabora un suo modello predittivo. Le coppie di dati di input e output nel training set vengono quindi utilizzate per calibrare i parametri aperti del modello. Una volta che il modello è stato addestrato con successo, può essere utilizzato per prevedere le variabili target y, su dati, delle features di input x, nuovi e mai visti dalla macchina.

   Per quanto riguarda il Supervised learning, possiamo ulteriormente distinguere tra **problemi di regressione**, in cui è previsto un apprendimento di valori numerici (apprendimento quantitativo), ad esempio il numero di utenti di una certa piattaforma; e **problemi di classificazione**, dove il risultato della previsione è

invece un'affiliazione di classe categoriale (apprendimento qualitativo), come ad esempio "email spam" o "email non spam".

2. **Unsupervised learning**: ha luogo quando il sistema di apprendimento deve rilevare dei patterns senza specifiche pre-esistenti. Pertanto, i dati di addestramento consistono solo di variabili di input (x), con l'obiettivo di trovare informazioni strutturali di interesse, come gruppi di elementi che condividono proprietà comuni (noto come **clustering**), o rappresentazioni di dati che vengono trasformate da spazi ad alta dimensione in uno inferiore, mantenendo in quest'ultimo alcune proprietà significative dei dati originali (noto come **riduzione della dimensionalità**).

   Esempi attuali di applicazione dell'unsupervised learning possono essere: l'uso di tecniche di clustering per raggruppare clienti o mercati in segmenti allo scopo di una comunicazione più specifica per il gruppo target; il riconoscimento vocale utilizzato da assistenti come Siri, Alexa o Google Assistant; oppure ancora, l'ordinamento automatico delle gallerie fotografiche sugli smartphone, dove foto contenenti le stesse persone o le stesse posizioni vengono raggruppate vicine tra di loro; ecc.

3. **Reinforcement learning**: in questo sistema di apprendimento, invece di fornire coppie di dati di input e output, si descrive lo stato attuale del sistema, si specifica un obiettivo da raggiungere, si fornisce un elenco di azioni consentite e si lascia che il modello sperimenti da per sé il processo di raggiungimento dell'obiettivo, utilizzando il principio di tentativi ed errori, cercando di massimizzare una ricompensa nel lungo periodo. La restituzione di una ricompensa come effetto dell'azione eseguita dal sistema, è il "rinforzo" di questa modalità di apprendimento, che si distingue dalle altre per il concetto di interazione della macchina con l'ambiente circostante.

   Il reinforcement learning si utilizza nei casi in cui è necessario raggiungere un obiettivo in un ambiente incerto, quando non è possibile prevedere tutte le variabili, quando non c'è un solo modo per eseguire un compito, ma occorre osservare delle regole: un esempio è la guida autonoma con il codice stradale..

   Uno degli esempi più famosi di reinforcement learning è **AlphaGo**, il software per il gioco cinese del Go realizzato da DeepMind:

   "AlphaGo is the first computer program to defeat a professional human Go player, the first to defeat a Go world champion, and is arguably the strongest Go player in history."[3]

   Altro campo in cui viene utilizzato, come già anticipato sopra, è quello dei **sistemi avanzati di assistenza alla guida (ADAS, Advanced Driver Assistance Systems)**. Ciascuna azione di guida autonoma può infatti corrispondere a una policy: il parcheggio automatico, il cambio corsia, il sorpasso, il control cruise automatico, ecc.

[3] [Alp]

Figure 3.2: AWS DeepRacer, automobile da corsa autonoma in scala 1:18 progettata da Amazon per gli sviluppatori di modelli di reinforcement learning[Aws]

A seconda del compito di apprendimento, il campo del machine learning offre varie classi di algoritmi. Di particolare interesse è la famiglia delle reti neurali artificiali (ANN), poiché la loro struttura flessibile consente loro di essere modificate per un'ampia varietà di contesti in tutti e tre i tipi di machine learning.

3.2.1 ANN: artificial neural networks

Ispirate al principio dell'elaborazione delle informazioni nei sistemi biologici, le reti neurali artificiali sono costituite da rappresentazioni matematiche di unità di elaborazione connesse chiamate neuroni artificiali. Come sinapsi in un cervello, ogni connessione tra neuroni trasmette segnali la cui forza può essere amplificata o attenuata da un **peso** che viene continuamente aggiustato durante il processo di apprendimento.

I segnali in input ad un neurone vengono così di seguito elaborati: una funzione sommatoria raccoglie i valori degli ingressi "pesati" e li somma, aggiungendo un ulteriore valore fisso interno chiamato **bias**. Quest'ultimo è un offset che può servire a trovare il punto di lavoro ottimale del neurone. Successivamente, solo se viene superata una certa soglia, una **funzione di attivazione** produce un valore in uscita dal neurone parteno dal risultato della sommatoria precedente.

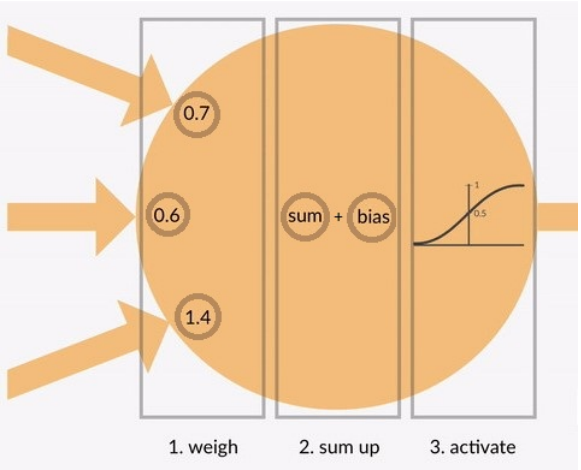

Figure 3.3: Rappresentazione di un neurone artificiale[Art]

Tipicamente, i neuroni sono organizzati in reti con diversi livelli:

- un livello di input di solito riceve l'input dei dati (ad esempio, un immagine);
- un livello di output produce il risultato finale (ad esempio, il riconoscimento di un viso, di un oggetto o di una voce);
- nel mezzo, ci sono zero o più livelli nascosti che sono responsabili dell'estrazione di patterns dai dati forniti in input, apprendendo così una mappatura non lineare tra input ed output.

Il numero di livelli e neuroni, insieme ad altre scelte, come la funzione di attivazione, non possono essere appresi dall'algoritmo di apprendimento. Costituiscono gli iperparametri di un modello e devono essere impostati manualmente.

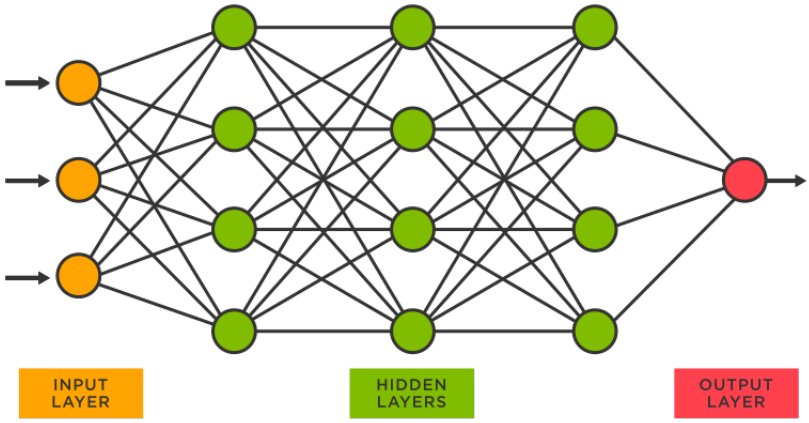

Figure 3.4: Struttura di una rete neurale artificiale[Neu]

Le reti neurali che consistono in più di uno strato nascosto sono chiamate **reti neurali profonde**. Queste sono organizzate in architetture di rete profondamente annidate e di solito contengono neuroni avanzati, cioè, possono utilizzare operazioni complesse (ad esempio, convoluzioni) o più attivazioni in un neurone, piuttosto che utilizzare una semplice funzione di attivazione. Queste caratteristiche consentono alle reti neurali profonde di essere alimentate con dati di input grezzi e di scoprire automaticamente la rappresentazione necessaria per l'attività di apprendimento corrispondente. Questa è la capacità principale di queste reti, che è comunemente nota come deep learning.

Le ANN semplici (con un solo strato nascosto) ed altri tipi di algoritmi di machine learning, possono essere inclusi nel termine *shallow machine learning (apprendimento automatico superficiale)*, poiché non forniscono tali funzionalità.

3.2.2 Pattern recognition

In generale, una feature descrive una proprietà individuale e misurabile di un fenomeno osservato, derivata dall'input di dati grezzi fornito alla macchina, che può essere sfruttata per la costruzione di modelli analitici efficienti.

L'insieme, inizialmente grezzo, delle caratteristiche (features), potrebbe essere ridondante e troppo vasto per essere gestito efficientemente. Di conseguenza, un tipico passo preliminare, importante per l'identificazione automatizzata di patterns e relazioni da

grandi data sets, consiste nel filtrare le features più rilevanti mediante l'utilizzo di alcune forme speciali di riduzione della dimensionalità, applicate al cosiddetto *input space* ("spazio di ingresso", data set di input).

Le tecniche disponibili a tale scopo, sono essenzialmente 3:

1. **Feature Selection**: la selezione delle caratteristiche, è il processo di selezione di un sottoinsieme di caratteristiche rilevanti dai dati grezzi, da utilizzare nella costruzione del modello;

2. **Feature Extraction**: l'estrazione delle caratteristiche, si riferisce al processo di trasformazione dei dati grezzi in caratteristiche numeriche che possono essere elaborate, preservando le informazioni nel data set originale;

3. **Feature Engineering**: l'ingegneria delle funzionalità, si riferisce al processo di utilizzo della propria conoscenza di dominio, per selezionare e trasformare le caratteristiche più rilevanti dai dati grezzi durante la creazione di un modello.

Feature engineering e feature extraction sono molto simili tra loro; entrambe si riferiscono alla creazione di nuove caratteristiche partendo dai dati grezzi in input, con la differenza tra le due, essere la seguente: l'ingegneria delle funzionalità si riferisce alla creazione di una nuova feature quando avremmo potuto utilizzare anche i dati non elaborati, mentre l'estrazione delle caratteristiche si riferisce alla creazione di nuove features quando non avremmo potuto utilizzare i dati grezzi nell'analisi (esempio, la conversione di un'immagine in valori RGB).

L'ingegneria delle funzionalità è l'abilità più importante quando si desidera ottenere buoni risultati per la maggior parte delle attività di previsione. Tuttavia, è difficile da apprendere e padroneggiare poiché diversi set di dati e diversi tipi di dati richiedono approcci di feature engineering diversi. La sua difficoltà e lo sforzo richiesto, sono le ragioni principali per cui si sono ricercati algoritmi in grado di apprendere le caratteristiche; ovvero algoritmi che possano progettare automaticamente le features.

Gli algoritmi di **feature learning** trovano i patterns comuni che sono importanti da distinguere tra le classi e li estrae automaticamente per essere utilizzati in un processo di classificazione o regressione. Il feature learning può essere pensato come l'ingegneria delle funzionalità eseguita automaticamente dagli algoritmi.

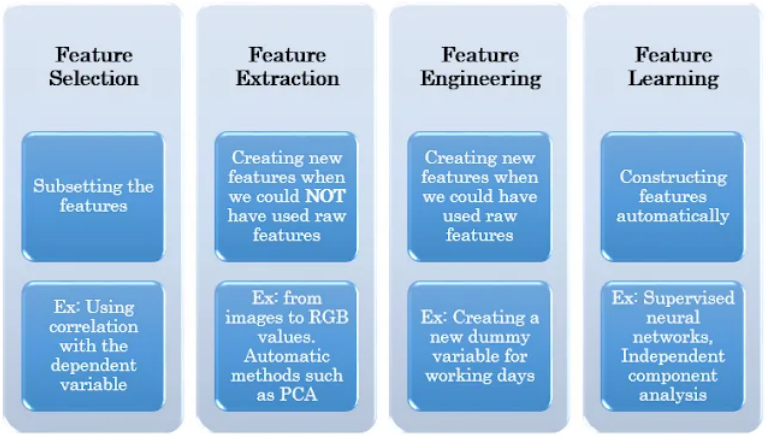

Figure 3.5: Tecniche di trasformazione dell'input data set in features[Bha19]

3.3 Deep learning

3.3.1 Feature Learning

Le reti neurali profonde superano le difficoltà dell'ingegneria delle funzionalità. La loro architettura avanzata offre loro la capacità di feature learning, per estrarre carattestiche discriminanti con il minimo sforzo umano.

Per questo motivo, il deep learning gestisce meglio dati su larga scala, rumorosi e non strutturati.

Il processo di feature learning generalmente procede in modo gerarchico, estraendo più livelli di features non lineari e passandole ad un classificatore che le combina per fare previsioni.

È necessario impilare gerarchie così profonde di features non lineari poichè non è possibile apprenderne di complesse da pochi strati. Si può dimostrare matematicamente che, per le immagini come esempio, le migliori caratteristiche per un singolo strato sono bordi e blob(macchie), perché contengono la maggior parte delle informazioni che è possibile estrarre da una singola trasformazione non lineare. Per generare features che contengono più informazioni non è possibile operare direttamente sugli input, ma è necessario trasformare nuovamente le prime features estratte (bordi e blob) per ottenerne di più complesse, che contengano più informazioni da distinguere tra le classi.

È stato dimostrato che il cervello umano fa esattamente la stessa cosa: la prima gerarchia di neuroni che riceve informazioni nella corteccia visiva è sensibile a bordi e macchie specifiche mentre le regioni del cervello più in basso nella pipeline visiva sono sensibili a strutture più complesse come i volti.

Sebbene il feature learning gerarchico fosse utilizzato prima che esistesse il campo del deep learning, queste architetture soffrivano di problemi importanti come il **problema della scomparsa del gradiente (vanishing gradient problem)**, in cui i gradienti diventavano troppo piccoli per fornire un segnale di apprendimento significativo per gli strati molto profondi delle reti, rendendo così queste architetture poco performanti rispetto agli algoritmi di shallow machine learning.

Il termine deep learning ha avuto origine dai nuovi metodi e strategie progettati per generare queste gerarchie profonde di caratteristiche non lineari, superando il problema della scomparsa del gradiente.

3.3.2 Architetture deep learning

Varie architetture deep learning sono emerse nel tempo. Sebbene praticamente ogni architettura possa essere utilizzata per ogni attività, alcune sono più adatte di altre per dati specifici come immagini o serie temporali. Le varianti si differenziano tra di loro dai tipi di strati, unità neurali e connessioni che utilizzano.

Di seguito, ne vengono riassunte alcune delle più comuni. È però importante tener presente che le seguenti architetture delineano una forma base di rete neurale, dalla quale molte sottovarianti sono successivamente state sviluppate; e rappresentano quindi una minoranza di tutte quelle attualmente esistenti.

- **Convolutional neural networks (CNN, reti neurali convoluzionali)**: Le CNN sono utilizzate principalmente per attività relative alla visione artificiale e al riconoscimento vocale. Sono in grado di affrontare attività che coinvolgono set di dati con relazioni spaziali, in cui le colonne e le righe non sono intercambiabili

(ad esempio, dati immagine). La loro architettura di rete comprende una serie di fasi che consentono l'apprendimento gerarchico delle caratteristiche. Ad esempio, quando si considera il riconoscimento degli oggetti nelle immagini, i primi strati della rete sono responsabili dell'estrazione delle caratteristiche di base sotto forma di bordi e angoli. Questi vengono quindi aggregati in modo incrementale in caratteristiche più complesse negli ultimi strati che assomigliano agli oggetti reali di interesse, come animali, case o automobili. Successivamente, le funzionalità generate automaticamente vengono utilizzate a scopo di previsione per riconoscere oggetti di interesse in nuove immagini.

- **Recurrent neural networks (RNN, reti neurali ricorrenti)**: Le RNN sono progettate esplicitamente per le strutture dati sequenziali, come serie temporali, sequenze di eventi e linguaggio naturale. La loro architettura offre cicli di feedback interni e quindi consente l'apprendimento sequenziale di pattern per modellare le dipendenze temporali formando una memoria. Le semplici architetture RNN sono problematiche poiché soffrono del problema della scomparsa del gradiente, con conseguente scarsa o nessuna influenza dei primi ricordi. Architetture più sofisticate, come le **reti di memoria a lungo termine (LSTM, Long Short-Term Memory)** si occupano di questo problema. Le RNN sono in genere applicate per la previsione di serie temporali, previsione del comportamento di un processo e attività di elaborazione del linguaggio naturale (NLP, Natural Language Processing).

- **Autoencoder**: Gli autocodificatori forniscono una rappresentazione densa delle caratteristiche dei dati di input. Tuttavia, non sono limitati ai dati in linguaggio naturale, ma possono essere applicati a qualsiasi tipo di input. Tali architetture di solito consistono di una fase di codifica in cui l'input è compresso in una rappresentazione a bassa dimensione e una fase di decodifica in cui la rete cerca di ricostruire l'input originale dalle caratteristiche apprese. In questo modo, la rete è costretta a mantenere informazioni significative nella rappresentazione latente ignorando il rumore irrilevante. Gli autocodificatori vengono comunemente applicati per l'apprendimento di funzionalità senza supervisione e per la riduzione della dimensionalità, in combinazione con altri sistemi di apprendimento successivi. Tuttavia, grazie alla loro capacità di quantificare gli errori di ricostruzione, che si presume siano significativamente più elevati per i campioni anomali che per i casi regolari, possono anche essere applicati per rilevare anomalie, come attività fraudolente.

- **Generative adversarial neural networks (GAN, reti neurali generative avversarie)**: Le GAN appartengono alla famiglia dei modelli generativi, che mirano all'apprendimento di una distribuzione di probabilità su un insieme di dati di addestramento, in modo che la rete possa generare casualmente nuovi campioni di dati con qualche variazione. A tale scopo, le reti GAN sono costituite da due sottoreti concorrenti. La prima rete è una rete generatore che cattura la distribuzione dell'input e genera nuovi esempi. La seconda rete è una rete discriminatore che cerca di distinguere esempi reali da quelli generati artificialmente. Entrambe le reti sono addestrate insieme in maniera competitiva nel contesto di un gioco a somma zero, in cui il guadagno di una rete è la perdita di un'altra, fino a quando il discriminatore non è più in grado di distinguere tra entrambi i tipi di campioni. Su questa base, è probabile che i GAN rivoluzionino i domini in cui vengono creati continuamente nuovi contenuti o nuove configurazioni di prodotto

(ad esempio, la composizione di arte e musica), o dove il contenuto viene convertito da una rappresentazione all'altra (ad esempio, testo in immagine per le descrizioni dei prodotti).

Ma quando parliamo di deepfake, quali architetture di reti neurali sopra citate, o quali architetture non citate, sono di nostro interesse? Possono i deepfake venire creati da qualunque tipo di rete neurale? O ve ne sono alcune specifiche allo scopo? E per quanto riguarda il campo del deepfake detection? Se vi sono reti neurali che generano questi media, vi saranno reti neurali in grado di rilevarli? Ebbene, le architetture sopra brevemente introdotte, sono più che sufficienti per i problemi proposti. In particolare, autoencoders e reti generative avversarie (GAN) sono tra le tipologie più utilizzate nella creazione di deepfake. Mentre per quanto riguarda la deepfake detection, la soluzione al problema si fa più complicata, in quanto, come vedremo successivamente in un capitolo dedicato, rilevare dei falsi, considerato il livello di realismo e credibilità oggi raggiunto da essi, può risultare complicato anche per algoritmi di deep learning. In ogni caso, le reti neurali convoluzionali (CNN) sono tra le più adatte ed utilizzate nel campo della computer vision, quindi fondamentali per qualunque algoritmo di deepfake detection si voglia sviluppare. È da notare che, un algoritmo di deep learning non per forza è limitato all'utilizzo di una sola architettura, infatti molte delle soluzioni proposte al problema, fanno un uso combinato di più tipologie di algoritmi.

Procediamo quindi nell'approfondire queste architetture di rete per comprenderne il funzionamento.

3.3.2.1 Reti neurali convoluzionali

Le CNN erano precedentemente note come LeNet. LeNet prende il nome dal suo creatore Yann LeCun, il quale creò nel 1989[4] una rete per l'identificazione delle cifre scritte a mano, basandosi sul lavoro precedentemente svolto da Kunihiko Fukushima, uno scienziato giapponese che aveva progettato il Neocognitron [5], una rete neurale essenziale per il riconoscimento delle immagini. LeNet-5, che descrisse i componenti primitivi delle reti convoluzionali, potrebbe essere considerata come l'inizio di queste. In quegli anni, a causa della scarsità di apparecchiature hardware, in particolare GPU (unità di elaborazione grafica), LeNet-5 non era particolarmente nota, di conseguenza, anche le ricerche riguardanti le CNN tra il 1990 e il 2000 non furono molte. Ad aprire le porte alle applicazioni di visione artificiale e a molte diverse varianti di reti convoluzionali fu il successo di AlexNet[6] nel 2012, una versione di rete CNN che prevedeva il raggruppamento di più strati convoluzionali tra loro per adattare il modello a sistemi con due GPUs.

Durante l'ideazione di questa architettura, i ricercatori si sono ispirati alle funzioni e all'organizzazione della corteccia visiva umana, infatti, è stata progettata per assomigliare alle connessioni tra i neuroni nel cervello umano.

Più nello specifico, una CNN comprende tre tipi principali di strati neurali:

1. Strati convoluzionali;
2. Strati di sottocampionamento (pooling);

[4] [LeC+89]
[5] [Fuk80]
[6] [KSH12]

3. Strati compleatamente connessi (fully-connected).

Ogni tipo di strato svolge un ruolo diverso, trasformando il volume di dati che riceve in input in un volume di output di attivazione neuronale, arrivando infine agli strati finali completamente connessi, i quali eseguiranno una mappatura dei dati ricevuti su un vettore di dimensione 1.

Strati convoluzionali

Questo è il primo passo nel processo di estrazione di caratteristiche da un'immagine. Ogni immagine è considerata come una matrice dei valori dei pixel.

Uno strato di convoluzione ha diversi filtri che eseguono l'operazione di convoluzione sulla matrice.

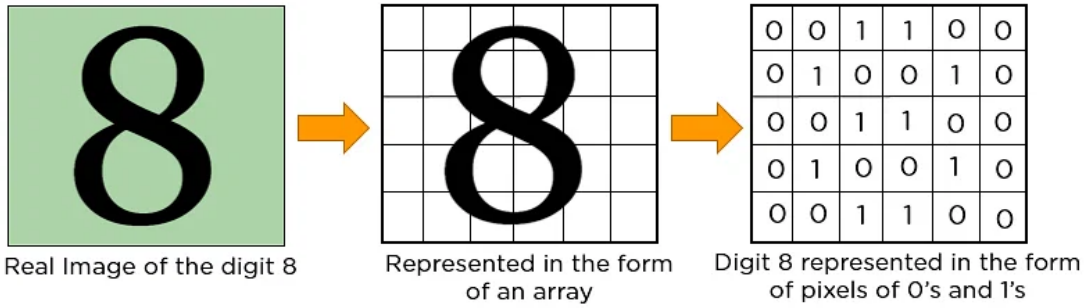

Figure 3.6: Esempio di rappresentazione di un'immagine sotto forma di matrice[Cnnb]

La convoluzione è un'operazione matematica che descrive una regola su come combinare due funzioni.

Nel caso in oggetto delle CNN, la mappa delle caratteristiche (o dati di input) e il kernel di convoluzione vengono mescolati insieme per formare una mappa delle caratteristiche trasformata. La convoluzione è spesso interpretata come un filtro, dove il kernel filtra la mappa delle caratteristiche per informazioni di un certo tipo (ad esempio un kernel potrebbe filtrare per i bordi e scartare altre informazioni).

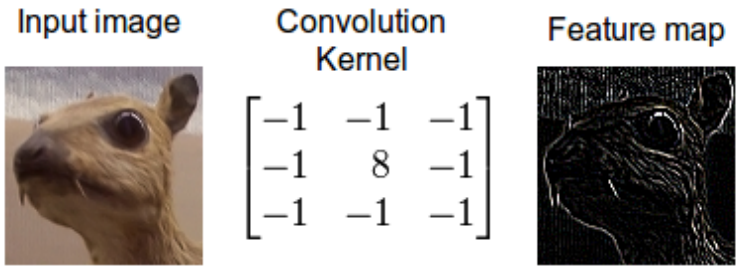

Figure 3.7: Convoluzione di un'immagine con un kernel rilevatore di bordi[Con]

Il filtraggio svolto dal kernel sulla matrice contenente i valori dell'immagine in input non è altro che un semplice prodotto scalare tra le due matrici.

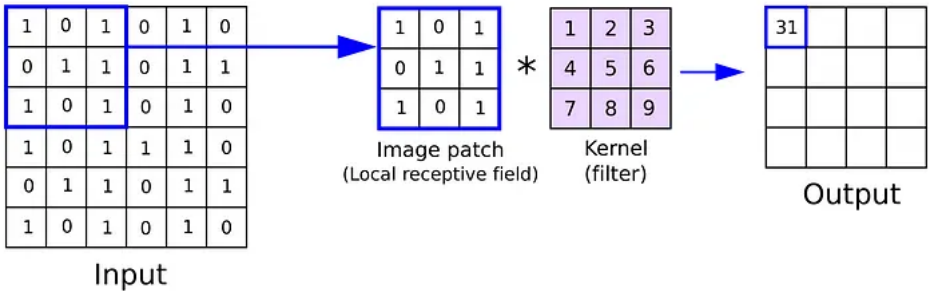

Figure 3.8: Prodotto scalare tra la matrice input e la matrice filtro (Kernel)[Cnnc]

Il kernel scorrerà lungo l'altezza e la larghezza dell'immagine, producendo la rappresentazione di essa per la specifica regione ricettiva in cui si trova. Questo produce una rappresentazione bidimensionale dell'immagine nota come mappa di attivazione che fornisce la risposta del kernel in ogni posizione spaziale dell'immagine. La dimensione secondo la quale il kernel si sposterà nella matrice di input è chiamata *stride (passo)*.

Ora, un fenomeno che si verifica dopo ogni operazione di convoluzione è il restringimento dell'immagine di input. In sostanza, se la matrice originale è

$$(N * N)$$

e il filtro è

$$(F * F),$$

la matrice risultante sarebbe

$$(N - F + 1) * (N - F + 1).$$

Questo avviene perché i pixel sui bordi e sugli angoli, vengono utilizzati meno dei pixel al centro dell'immagine. Ciò causa una duplice conseguenza: si avrà un caso di perdita delle informazioni contenute nei bordi, le quali non vengono conservate così come quelle al centro; ed in più viene posto un limite massimo al numero di volte in cui tale operazione potrebbe essere eseguita prima che l'immagine si riduca a nulla, precludendoci così la costruzione di reti più profonde.

Padding Il **padding (riempimento)** è semplicemente un processo di aggiunta di strati di pixel, con valori settati a zero, alle nostre immagini di input in modo da evitare i problemi sopra menzionati. Ciò impedisce il restringimento poiché, se

$$P = \textit{numero di strati di pixel aggiunti al bordo dell'immagine},$$

la nostra immagine

$$(N * N)$$

diventa un'immagine

$$(N + 2P) * (N + 2P)$$

dopo il riempimento. Quindi, applicando l'operazione di convoluzione,

$$\textit{con filtro } (F * F),$$

si ottengono immagini

$$(N + 2P - F + 1) * (N + 2P - F + 1).$$

Ciò aumenta il contributo dei pixel ai bordi dell'immagine originale portandoli al centro dell'immagine estesa. Pertanto, le informazioni sui bordi vengono preservate così come le informazioni al centro.

Vi sono differenti strategie di padding, le più comuni sono le seguenti:

- Valid padding: implica nessuno strato aggiuntivo. L'immagine di input viene lasciata nella sua forma valida/inalterata.
- Same padding: in questo caso, aggiungiamo strati di riempimento "P" in modo tale che l'immagine di output abbia le stesse dimensioni dell'immagine di input. Quindi, prendendo come esempio l'immagine 3.9 sottostante, se usiamo un filtro (3 x 3), uno strato di zeri deve essere aggiunto ai bordi per ottenere un'immagine delle medesime dimensioni. Allo stesso modo, se si utilizza un filtro (5 x 5), è necessario aggiungere 2 strati di zeri al bordo dell'immagine.

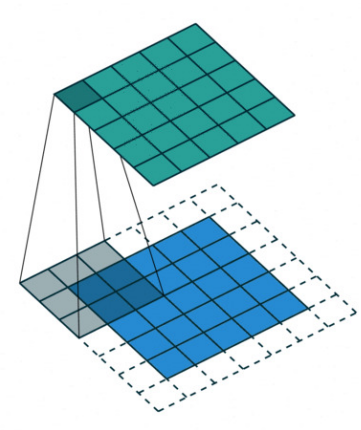

Figure 3.9: Tecnica *same padding* per mantenere le dimensioni dell'immagine originale[Cnna]

Cosa accade se nelle mappe risultanti dal prodotto scalare delle due matrici vi sono valori negativi? Le mappe di attivazione vengono fornite come input ad un'altra funzione matematica, chiamata funzione di attivazione (ogni neurone artificiale ne ha una). Si potrebbe quindi incorrere nel problema della scomparsa del gradiente, man mano che le mappe vengono processate verso gli strati più profondi della rete.

Unità lineare rettificata (ReLU layer) Poiché la convoluzione è un'operazione lineare e le immagini sono tutt'altro che ciò, gli strati di non linearità vengono spesso posizionati direttamente dopo il livello convoluzionale, per introdurre la non linearità nelle mappe di attivazione. **ReLU**, che sta per unità lineare rettificata, è sicuramente la funzione di attivazione più comune utilizzata, tanto che in molti si riferiscono a questa

come **"Strato ReLU"**. In realtà non è una componente separata dello strato convoluzionale. ReLU esegue un'operazione basata sugli elementi delle mappe ed imposta tutti i pixel negativi su 0. Introduce la non linearità nella rete e l'output generato è una mappa delle caratteristiche rettificata.

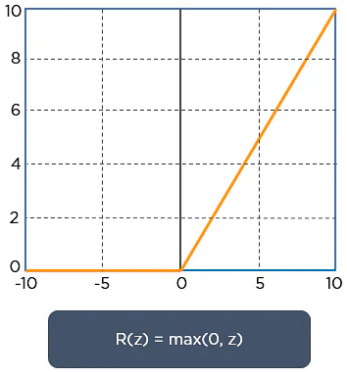

Figure 3.10: Funzione ReLU[Cnnb]

Strati di sottocampionamento

Dopo aver ottenuto le mappe delle caratteristiche rettificate, è necessario aggiungere un livello di pooling (sottocampionamento) nelle CNN, accanto a un livello di convoluzione. Il compito di questo livello è ridurre le dimensioni spaziali delle mappe di features. Come risultato della riduzione della dimensionalità, la potenza del computer richiesta per elaborare i dati viene ridotta. Ciò aiuta anche nell'estrazione delle caratteristiche principali, fornendo l'invarianza di base alle rotazioni e alle traslazioni. Esistono due forme di pooling: il pooling massimo e il pooling medio.

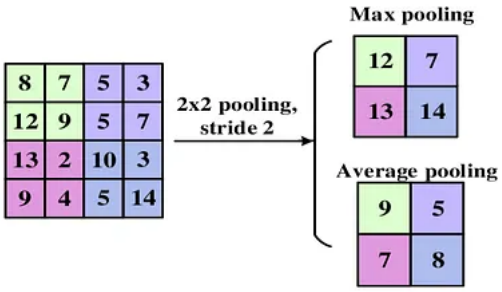

Figure 3.11: Le due forme di sottocampionamento[Cnnc]

Strati completamente connessi

Flattening Il flattening (appiattimento) viene utilizzato per convertire tutte le mappe di features raggruppate, risultanti dagli strati di pooling, in un unico vettore lineare lungo e continuo. La matrice appiattita viene inviata come input allo strato completamente connesso.

Lo strato completamente connesso (fully-connected) è dove avviene la classificazione delle immagini nella CNN in base alle caratteristiche estratte nei livelli precedenti.

Qui, completamente connesso significa che tutti gli input o nodi di un livello sono collegati a ogni unità di attivazione o nodo del livello successivo.

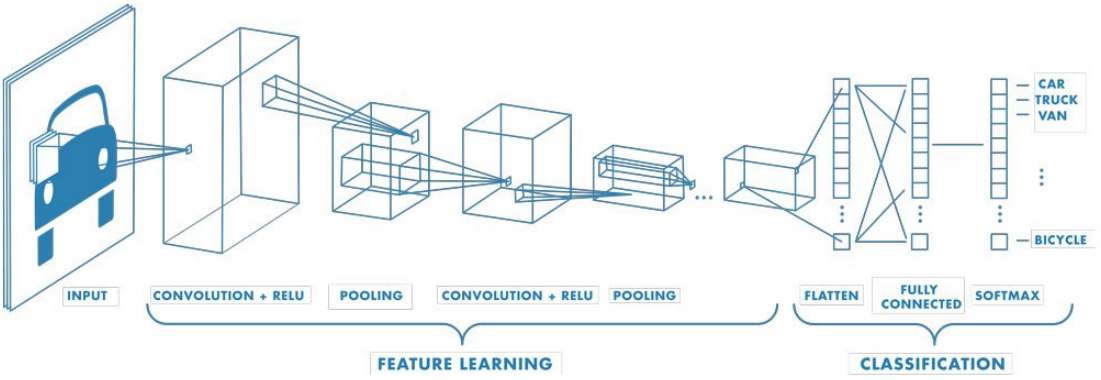

Figure 3.12: Architettura CNN completa[Cnnd]

3.3.2.2 Reti generative avversarie

Cosa hanno in comune tutte le persone sopra mostrate?

Non esistono.

Un'intelligenza artificiale le ha inventate. Più precisamente, le ha generate utilizzando milioni di strutture di pixel simili come esempi.

Ho creato le immagini sopra riportate sul sito this-person-does-not-exist.com[7]. Tali ritratti falsi così realistici sono resi possibili dall'invenzione delle cosiddette "Generative Adversarial Networks" (GAN). Come avevamo già anticipato nel capitolo 2, le GAN sono state introdotte per la prima nell'articolo scientifico[8] pubblicato nel 2014 da Ian Goodfellow e i suoi colleghi.

Goodfellow dimostrò come fosse possibile utilizzare la potenza di calcolo moderna per generare falsi esempi che sembrano immagini reali di numeri, persone, animali o qualsiasi cosa si possa immaginare; fintanto che i dati in input vengano curati. Nell'immagine seguente, estratta dall'articolo di Goodfellow, le colonne gialle sono esempi di immagini generate dalla prima GAN nel 2014.

Già un anno dopo la loro introduzione, la ricerca inizia ad apportare i primi miglioramenti a queste reti. Nel 2015 i ricercatori iniziano a combinare GAN con reti neurali

[7] [Thi]
[8] [Goo+14]

Figure 3.12: Già nel 2014, vi erano indicazioni che le GAN potessero produrre volti credibili[Goo+14]

convoluzionali (CNN) ottimizzate per il riconoscimento delle immagini, le quali offrivano la possibilità di elaborare molti dati in parallelo e funzionare particolarmente bene con le schede grafiche. L'architettura risultante prese il nome di **DCGAN (Deep Convolutional GAN)**[9]. Due anni dopo, nel 2016 invece, i ricercatori combinano due GAN tra di loro in un'unica architettura di rete: lo scopo era quello di far cooperare gli agenti delle due sotto-reti facendogli condividere informazioni tra loro, eseguendo il processo di apprendimento in parallelo. Il nome dato a questa variante fu invece **CoGAN (Coupled GAN)**[10].

I risultati iniziavano a diventare più credibili ma lontano dall'essere realistici, i volti stessi avevano ancora molti difetti di immagine.

Figure 3.13: La struttura più complessa delle reti convoluzionali consentì la generazione di persone più credibili[RMC15]

Figure 3.14: Con le GAN accoppiate, gli esseri umani artificiali potevano anche indossare occhiali da sole o gioielli[LT16]

[9] [RMC15]
[10] [LT16]

Il più grande salto di qualità è avvenuto nel 2017 grazie ad alcuni ricercatori Nvidia (casa produttrice di GPUs), i quali riuscirono a risolvere un grande problema delle precedenti GANs. Gli agenti generatori spesso producevano immagini a bassa risoluzione perché erano più difficili da rilevare come falsi per l'agente discriminatore: più pixel significano potenzialmente più fonti di errore. Quindi aveva senso per l'IA del falsario evitare risoluzioni elevate per superare l'esaminatore. La soluzione di Nvidia fu la seguente: addestrare la rete in più fasi. Innanzitutto, l'intelligenza artificiale del falsario impara a creare immagini a bassa risoluzione. Quindi, la risoluzione aumenta gradualmente. Questa versione di reti GAN è chiamata **ProGAN (Progressive growing of GAN)**[11].

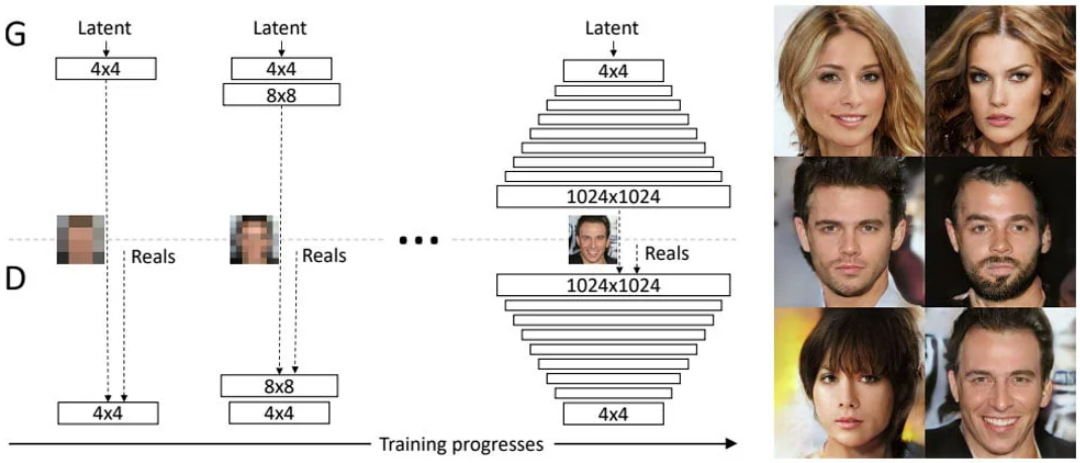

Figure 3.15: La GAN viene introdotta passo dopo passo ad alte risoluzioni[Kar+17]

Le GAN, che in questo modo crescono passo dopo passo, producono ritratti falsi di qualità senza precedenti: le immagini hanno ancora dei difetti, ma possono sicuramente ingannare le persone che non guardano molto da vicino.

Figure 3.16: I volti generati nel 2017 superano i risultati precedenti e alcuni sono appena riconoscibili come prodotti AI[Kar+17]

[11] [Kar+17]

Nel 2018, ancora una volta i ricercatori Nvidia riescono a controllare meglio le loro GAN: possono mirare a singole caratteristiche dell'immagine, ad esempio "capelli scuri" e "sorriso" nei ritratti. In questo modo, le caratteristiche delle immagini di addestramento possono essere trasferite in modo specifico alle immagini generate dall'IA. Il cosiddetto trasferimento di stile applicato alle reti GAN, ha dato forma ad una variante chiamata **StyleGAN** [12].

Figure 3.17: Il trasferimento di stile può essere utilizzato per controllare specificamente l'IA dell'immagine, ad esempio per creare immagini di sole persone sorridenti[KLA18]

Naturalmente, il principio GAN non funziona solo per i ritratti: all'intelligenza artificiale non interessa affatto il tipo di struttura dei pixel che produce. Richiede solo i dati di addestramento corrispondenti. Alla fine del 2018, Deepmind ad esempio, mostra cibo, paesaggi e animali generati dall'IA che sembrano incredibilmente credibili.

Le varianti nominate finora sono solo alcune delle varianti di GAN sviluppate dai ricercatori, ma sicuramente sono tra quelle che hanno contribuito di più al loro avanzamento tecnologico.

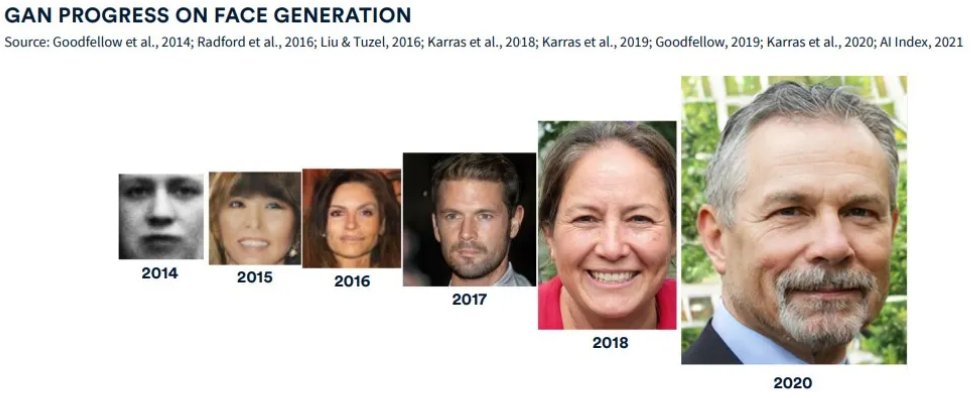

Figure 3.18: Miglioramento delle GAN nella generazione di volti umani dal 2014 al 2020

[12][KLA18]

Architettura di rete

Analizziamo ora la loro struttura di rete.

Le reti generative avversarie, o GAN in breve, sono un approccio alla **modellazione generativa** che utilizza metodi di deep learning. La modellazione generativa è un'attività di apprendimento rientrante nel machine learning non supervisionato (Unsupervised learning) che implica la scoperta e l'apprendimento automatico delle regolarità o dei modelli nei dati di input, in modo tale che il modello possa essere utilizzato per generare o produrre nuovi esempi che plausibilmente avrebbero potuto essere tratti dal set di dati originale.

Più in generale, le GAN sono un'architettura per l'addestramento di un modello generativo, ed è comune utilizzare altri modelli di deep learning in questa architettura, come per esempio nel caso della variante standardizzata DCGAN che fa utilizzo delle reti CNN.

L'architettura del modello GAN prevede due sottomodelli: il modello generatore che addestriamo per generare nuovi esempi e il modello discriminatore che cerca di classificare gli esempi come reali (da il dominio) o falso (generato). I due modelli vengono addestrati insieme in un contesto di gioco a somma zero, competitivo, fino a quando il modello discriminatore viene ingannato circa la metà delle volte, il che significa che il modello generatore sta generando esempi plausibili.

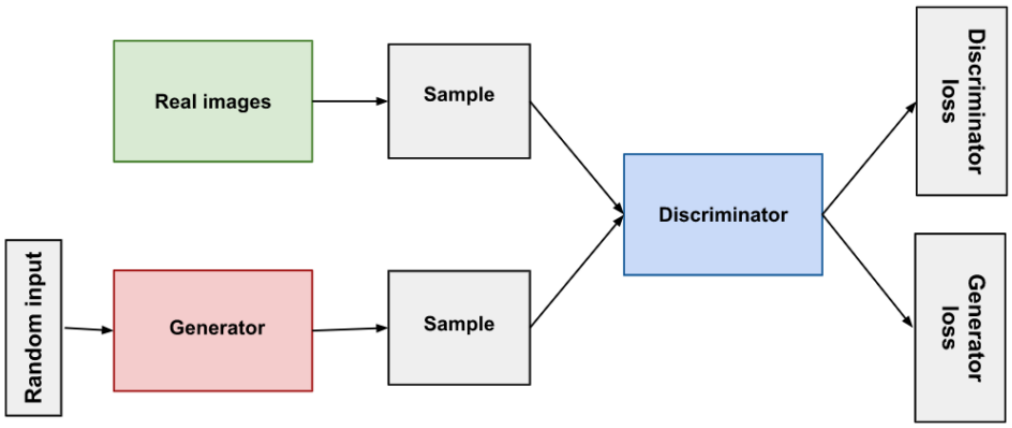

Figure 3.19: Intera architettura GAN[Gan]

Addestramento di rete

L'addestramento di una rete neurale comporta molte iterazioni del seguente ciclo a due passaggi:

1. Durante il passaggio in avanti (**forward pass**), il sistema elabora un batch di esempi per produrre previsioni. Il sistema confronta ogni previsione con ogni valore di etichetta. La differenza tra la previsione e il valore dell'etichetta è la **perdita (loss)** per quell'esempio. Il sistema aggrega le perdite per tutti gli esempi per calcolare la perdita totale per il batch corrente. La funzione utilizzata per calcolare il valore di perdita totale, viene chiamata **loss function** (o **cost function**, o ancora **error function**).

2. Durante il passaggio all'indietro (**backpropagation**), il sistema cerca di ridurre la perdita totale regolando il valore dei pesi e bias di tutti i neuroni, in tutti gli strati nascosti. Le reti neurali spesso contengono molti neuroni su molti strati nascosti. Ciascuno di questi neuroni contribuisce alla perdita complessiva in modi diversi. La retropropagazione determina se aumentare o diminuire i pesi applicati a particolari neuroni.

Il **tasso di apprendimento (learning rate)** è un iperparametro del modello e funge da moltiplicatore che controlla il grado in cui ogni passaggio all'indietro aumenta o diminuisce ogni peso. Un learning rate elevato aumenterà o diminuirà ogni peso in misura maggiore rispetto ad un tasso di apprendimento ridotto.

In termini di calcolo, la backpropagation calcola la derivata parziale della **funzione d'errore (loss function)** rispetto a ciascun parametro di ogni neurone.

Dopo aver riassunto il procedimento di training, universale per qualunque rete neurale di tipo **feed forward** (reti neurali con flusso in avanti le cui connessioni tra i nodi non formano cicli), si pone un particolare problema nella casistica trattata attualmente. Le GAN, come già abbiamo detto più volte, sono reti neurali formate a loro volta da due sottoreti (rete generatore e discriminatore). Queste, devono svolgere il loro processo di addestramento in maniera indipendente, senza influenzare su quello dell'altra.

Modello discriminatore

Il discriminatore in una GAN è semplicemente un classificatore. Cerca di distinguere i dati reali dai dati creati dal generatore. Potrebbe utilizzare qualsiasi architettura di rete appropriata al tipo di dati che sta classificando. L'utilizzo di reti convoluzionali tuttavia è molto frequente.

I dati di addestramento del discriminatore provengono da due fonti:

- Istanze di dati reali, come immagini reali di persone. Il discriminatore utilizza questi casi come esempi positivi durante l'allenamento.
- Istanze di dati false create dal generatore. Il discriminatore utilizza questi casi come esempi negativi durante l'addestramento.

Durante l'addestramento del discriminatore il generatore non si addestra. I suoi pesi rimangono costanti mentre produce esempi su cui il discriminatore può allenarsi.

Addestramento discriminatore Il discriminatore si collega a due funzioni di perdita. Durante il proprio addestramento, il discriminatore ignora il valore di perdita del generatore e utilizza solamente il proprio.

Durante l'addestramento, il discriminatore:

1. classifica sia i dati reali che i dati falsi che gli vengono forniti dal generatore;
2. viene eventualmente penalizzato dalla sua funzione di perdita per aver classificato erroneamente un'istanza reale come falsa o un'istanza falsa come reale;
3. aggiorna i suoi pesi attraverso la retropropagazione, dalla funzione di perdita attraverso tutta la propria struttura di rete.

Modello generatore

La parte generatore di un GAN impara a creare dati falsi incorporando il feedback del discriminatore. Il suo obiettivo è fare in modo che quest'ultimo classifichi il suo output come reale.

L'output del generatore è talvolta indicato come spazio latente o vettore latente, ovvero uno spazio multidimensionale astratto contenente valori caratteristici che non si possono interpretare direttamente, ma che sono codificati in una rappresentazione interna significativa. Non è altro che una rappresentazione compressa dell'informazione.

Per ottimizzare il generatore, bisogna prima passare l'output proveniente da esso attraverso il discriminatore. Successivamente, è possibile eseguire la retropropagazione (backpropagation) e calcolare gli errori.

L'addestramento del generatore richiede un'integrazione più stretta tra il generatore e il discriminatore rispetto a quanto richiesto dall'addestramento del discriminatore. Questo semplicemente perchè il generatore si trova all'inizio dell'architettura di una GAN, e affinchè venga modificato il valore dei pesi contenuti nella sua sottorete, l'algoritmo di retropropagazione dovrà percorrere a ritroso anche la sottorete del discriminatore.

Addestramento generatore Quindi, per riassumere, la retropropagazione inizia dalla funzione di perdita del generatore che si trova nello strato di output della rete GAN e dovrà, attraversando il discriminatore, ritornare all'input del generatore. Allo stesso tempo, non vogliamo che il discriminatore cambi durante l'addestramento del generatore. Cercare di raggiungere un obiettivo in movimento renderebbe solamente più difficile per esso convergere su una soluzione.

Quindi, addestriamo il generatore con la seguente procedura:

1. dei dati in input casuali (anche chiamati rumore) vengono forniti al generatore;
2. viene prodotto un output significativo dal rumore dato in input;
3. si ottiene la classificazione "reale" o "fake" dal discriminatore per l'output del generatore;
4. si calcola il valore loss del generatore parteno dalla classificazione del discriminatore;
5. backpropagation attraverso il discriminatore e il generatore, per ottenere i gradienti;
6. utilizzo dei gradienti calcolati per modificare solamente il valore dei pesi del generatore.

Dopo aver quindi compreso i processi di addestramento individuali di generatore e discriminatore, come formiamo la GAN nel suo insieme?

La formazione di una rete GAN procede a periodi alterni:

1. il discriminatore si allena per una o più epoche;
2. il generatore si allena per una o più epoche.

I passaggi 1 e 2 vengono continuamente ripetuti per proseguire nell'addestramento delle due sottoreti.

Man mano che il generatore migliora con il training, le prestazioni del discriminatore peggiorano perché non può più distinguere facilmente tra reale e falso. Se il generatore funziona perfettamente, il discriminatore avrà una precisione del 50%. Questa progressione pone un problema per la convergenza della rete nel suo insieme, poiché il feedback del discriminatore diventa via via meno significativo nel tempo e se la gan dovesse continuare ad allenarsi oltre il punto in cui il discriminatore fornisce un feedback completamente casuale, il generatore farebbe uso di un feedback spazzatura e la sua stessa qualità potrebbe crollare.

Per una GAN quindi, la convergenza è spesso uno stato temporaneo, piuttosto che stabile.

È questo avanti e indietro che consente alle GAN di affrontare problemi generativi altrimenti intrattabili. Si può quindi affermare che questa architettura ha trovato una punto d'appoggio nella risoluzione di un difficile problema generativo (unsupervised learning), partendo dal risolvere prima un problema di classificazione (supervised learning) molto più semplice.

3.3.2.3 Autoencoders

Così come le GAN appena trattate, un autocodificatore è un tipo di rete neurale artificiale non supervisionato.

Gli autoencoder sono un tipo specifico di reti neurali feedforward in cui l'output ottenuto sarà uguale all'input fornito, il quale viene compresso in un codice di dimensioni inferiori e successivamente ricostruito da questa rappresentazione. Il codice è una compressione dell'input, chiamato anche rappresentazione dello spazio latente.

La differenza o la perdita di qualità tra i dati di output e di input è chiamata **perdita di ricostruzione (reconstruction loss)**.

L'obiettivo principale per gli autocodificatori è rappresentare dati complessi utilizzando il minor codice possibile con una ricostruzione minima o nessuna perdita di "compressione". Per fare ciò, l'autoencoder deve esaminare i dati e costruire una funzione in grado di trasformare una particolare istanza di dati in un codice significativo. Possiamo pensare a questo come a una rimappatura dei dati originali utilizzando meno dimensioni. Possiamo anche tenere presente che questo codice deve essere interpretato successivamente da un decodificatore per accedere ai dati.

Architettura di rete

Iniziamo con una rapida panoramica dell'architettura degli autoencoder.

Essi sono composti da 3 parti:

- **codificatore (encoder)**: un modulo che comprime i dati di input in una rappresentazione codificata che in genere è di diversi ordini di grandezza inferiore;
- **bottleneck (o codice)**: un modulo che contiene le rappresentazioni della conoscenza compressa ed è quindi la parte più importante della rete;
- **decodificatore (decoder)**: un modulo che aiuta la rete a "decomprimere" le rappresentazioni dello spazio latente e ricostruisce i dati dalla loro forma codificata.

L'architettura nel suo insieme è simile alla seguente:

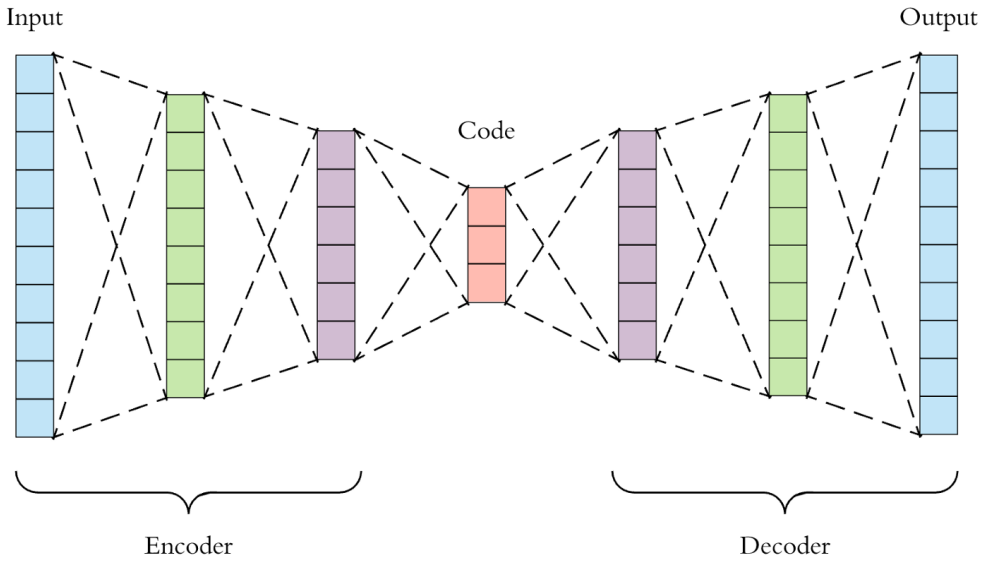

Figure 3.20: Architettura autoencoder[Aut]

Sia il codificatore che il decodificatore sono reti neurali feedforward completamente connesse con il codice. Il codice è un singolo strato di una ANN con la dimensionalità di nostra scelta. Il numero di nodi nel livello di codice (dimensione del codice) è un iperparametro che impostiamo prima di addestrare l'autoencoder.

Ci sono 4 iperparametri che devono essere impostati prima di addestrare un autocodificatore:

- Dimensione del codice: numero di nodi nel livello intermedio. Una dimensione più piccola si traduce in una maggiore compressione.
- Numero di livelli: l'autoencoder può essere profondo quanto vogliamo. Nella figura sopra abbiamo 2 strati sia nel codificatore che nel decodificatore, senza considerare l'ingresso e l'uscita.
- Numero di nodi per layer: il numero di nodi per strato diminuisce con ogni strato successivo del codificatore e aumenta nuovamente nel decodificatore.
- Funzione di perdita: utilizzata per la ricostruzione del codice da parte del decoder.

Esattamente come per le GAN, anche gli autoencoder hanno differenti varianti di architettura. Ad esempio, quando si discute di immagini come tipo di input, le reti neurali convoluzionali (CNN) possono essere utilizzate insieme agli autoencoder formando una variante dell'architettura predefinita, chiamata **autocodificatori convoluzionali (convolutional autoencoders)**.

Autocodificatori convoluzionali

In questo tipo di autoencoders, le due sottoreti di codifica e decodifica adottano la struttura di reti neurali convoluzionali: il codificatore sarà quindi composto da un insieme di blocchi convoluzionali seguiti da moduli di pooling che comprimono l'input nel bottleneck dell'autoencoder; il decoder invece, consiste in una serie di moduli di

sovracampionamento (oversampling) per riportare le caratteristiche compresse sotto forma di immagine. Questi strati di oversampling, utilizzati anche nel caso delle GAN, più nello specifico nel loro componente generatore, sono chiamati **strati deconvoluzionali (deconvolutional layers)** o **strati convoluzionali trasposti (transposed convolutional layers)**. La seconda forma è la più corretta poichè questo passaggio non invertirà il processo dalla convoluzione, almeno non per quanto riguarda i valori numerici (cosa che invece la funzione matematica deconvoluzione fa), ma ricostruirà semplicemente la risoluzione spaziale precedente. I valori numerici saranno invece semplicemente calcolati con una convoluzione, ma questo non sarà un problema per gli autoencoders poichè, come tutte le reti neurali, potranno sempre aggiustarli durante l'addestramento, con l'algoritmo di backpropagation.

Tuttavia, le due forme di autocodificatori finora trattate, ovvero modello base e convoluzionale, non possono essere considerati modelli di deep learning generativi come le GAN invece. Questo perchè il loro compito è solamente ricostruire l'input dato in una forma più possibile identica all'originale.

Ma allora come è possibile utilizzare gli autocodificatori per la creazione di deepfake, come da noi dichiarato in precedenza? Ebbene, esiste una variante generativa chiamata **Autocodificatori variazionali (VAE, Variational Autoencoders)**. Tuttavia vi è una soluzione alternativa, senza ricorrere all'utilizzo di questi ultimi. Si procede ad illustrarla.

Addestramento autoencoder per deepfake

Innanzitutto è importante notare che se addestriamo due autocodificatori separatamente, questi saranno incompatibili tra loro. I loro spazi latenti si basano su caratteristiche specifiche che ogni rete ha ritenuto significative durante il proprio processo di addestramento, quindi se due autocodificatori vengono addestrati separatamente su volti diversi, i loro spazi latenti rappresenteranno caratteristiche diverse.

Fatta questa premessa, consideriamo due autoencoders e trattiamoli separatamente. Il decoder A è addestrato solo con volti di una persona A; il decoder B è addestrato solo con volti di una persona B. Tuttavia, ciò che i due autoencoders hanno in comune è l'essere provvisti dello stesso encoder. Entrambe le rappresentazioni dello spazio latente quindi, saranno prodotte dallo stesso codificatore, il quale dovrà identificare le caratteristiche comuni in entrambe i volti. Poiché tutte le facce condividono una struttura simile, non è irragionevole aspettarsi che il codificatore apprenda il concetto stesso di "faccia".

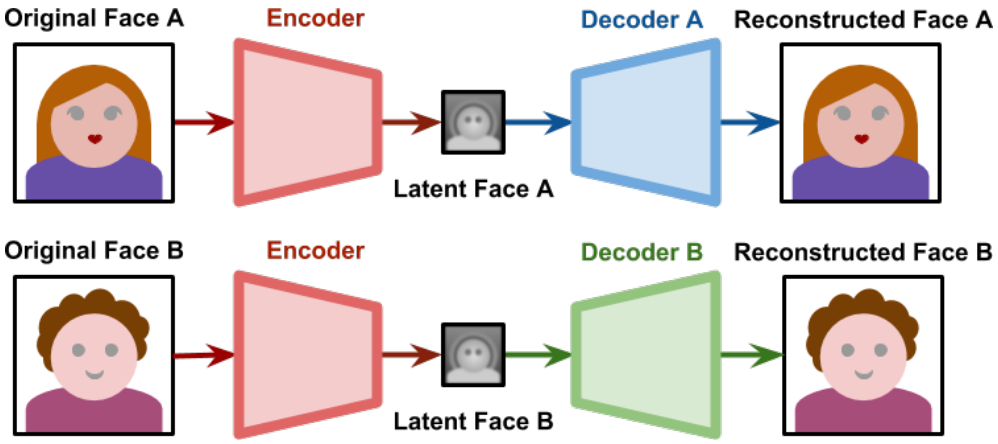

Figure 3.21: Architettura di rete per la generazione di deepfakes tramite autoencoder [Dfa]

Quando il processo di addestramento sarà completato, possiamo passare un viso latente generata dal soggetto A al decoder B. Come si vede nell'immagine sottostante, il decoder B proverà a ricostruire il soggetto B, dalle informazioni ricevute relative al soggetto A.

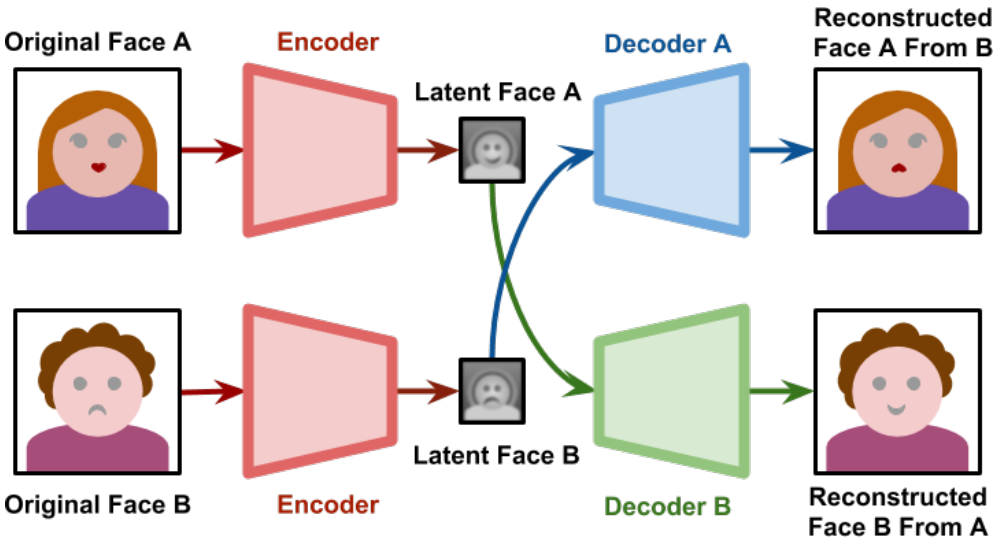

Figure 3.22: [Dfa]

Se la rete ha generalizzato abbastanza bene ciò da cui è composto un volto, lo spazio latente rappresenterà le espressioni facciali e gli orientamenti. Ciò significa generare un volto per il soggetto B con le stesse espressioni e orientamento del soggetto A.

4. Aspetti etico-sociali

La ricerca e lo sviluppo delle tecnologie trattate finora e la loro successiva e immediata disponibilità verso il grande pubblico, non poteva che significare un aumento del fenomeno dei deepfake, specialmente a seguire i primi casi pubblici accaduti nel 2017, già introdotti nel capitolo 2.

Al fine di affrontare i deepfake da un punto di vista etico-sociale e normativo, dobbiamo prima distinguere i loro casi d'uso, ovvero in quale ambiti sociali questi vengono utilizzati.

Per iniziare, possiamo fornire un elenco dei diversi potenziali utilizzi, alcuni dei quali positivi, altri negativi e preoccupanti.

Casi d'uso	**Breve descrizione**
Accessibilità e assistenza sanitaria	L'intelligenza artificiale può costruire strumenti per ascoltare, vedere e presto, ragionare con crescente precisione. I media sintetici (deepfake) generati dall'intelligenza artificiale possono aiutare a rendere gli strumenti di accessibilità più intelligenti e, in alcuni casi, anche convenienti e personalizzabili, il che può aiutare le persone ad aumentare la loro capacità di agire autonomamente ed ottenere indipendenza.
Intrattenimento	I deepfake possono essere utilizzati nel settore dell'intrattenimento sia da un punto di vista amatoriale, per creare meme, video divertenti, come video parodia o video che sovrappongono i volti delle persone a scene di film famosi; sia da un punto di vista professionale nella produzione di film, pubblicità e videogiochi, permettendo di tagliare costi nella loro fabbricazione.
Educazione	I deepfake possono facilitare numerose possibilità nel settore dell'istruzione. Scuole e insegnanti utilizzano media, audio e video in classe già da un po' di tempo ormai. I Deepfake possono quindi aiutare un educatore a fornire lezioni innovative molto più coinvolgenti rispetto ai tradizionali formati visivi e multimediali. Ne consegue quindi, un possibile utilizzo nell'ambito delle simulazioni di formazione per forze dell'ordine o professionisti medici.
Autonomia e libertà di espressione	I media sintetici possono aiutare attivisti per i diritti umani e giornalisti a rimanere anonimi in regimi dittatoriali e oppressivi. L'uso della tecnologia per denunciare le atrocità, su media tradizionali o social media, può dare molto potere a giornalisti cittadini e attivisti. I deepfake possono essere utilizzati per anonimizzare voce e volti e proteggere quindi la loro privacy.

Tabella 4.1: Possibili utilizzi positivi dei deepfake

Casi d'uso	Breve descrizione
Minacce per gli individui: pornografia e revenge porn	Il primissimo caso d'uso di natura dannosa dei deepfake è stato visto nella pornografia, infliggendo violenza emotiva, reputazionale e, in alcuni casi, nei confronti dell'individuo, principalmente donne.
Minacce per la società: disinformazione	I deepfake possono causare danni sociali a breve e lungo termine se utilizzati per diffondere disinformazione, erodendo la fiducia già in declino nei media e nelle istituzioni.
Minacce per la democrazia: influenza politica	Il deepfake di un candidato politico può sabotare l'immagine e la reputazione di un candidato politico e può anche alterare il corso di un'elezione.
Minacce per le imprese: truffe e frodi	I deepfake possono essere utilizzati per impersonare le identità di leader aziendali e dirigenti e facilitare le frodi o per la manipolazione dei mercati.

Tabella 4.2: Possibili utilizzi negativi dei deepfake

Affrontiamo ora questi casi d'uso, iniziando dai negativi per poi passare ai positivi, discutendo alcuni esempi del mondo reale, evidenziando i loro principali vantaggi, importanti preoccupazioni e conseguenze indesiderate. Forniamo anche una breve valutazione da un aspetto etico e giuridico.

4.1 Aspetti negativi

4.1.1 Minacce per gli individui: pornografia e revenge porn

Il primissimo caso d'uso delle GAN è stato quello di creare video di sesso deepfake; in particolare, prendendo di mira delle celebrità o casistiche di **revenge porn**. Con il termine revenge porn si fa riferimento a materiale sessualmente esplicito che viene creato e ampiamente diffuso per umiliare, minacciare o arrecare altro danno ad una persona con la quale si ha interrotto un rapporto. Ultimamente, il termine si è esteso anche a video simili dove i deepfake pornografici non consensuali sono distribuiti da hacker o da chiunque cerchi un guadagno in termini di denaro o notorietà piuttosto che per una vendetta per le relazioni perdute.

Secondo un rapporto[1] di Deeptrace (società di sicurezza informatica con sede ad Amsterdam che fornisce tecnologie di deep learning e computer vision per la rilevazione e il monitoraggio online dei media sintetici, ora chiamata sensity.ai[2]) sui deepfake, il 96% dei deepfake sono video pornografici, con oltre 135 milioni di visualizzazioni solo su siti web di quel tipo.

Figure 4.1: Video pornografici rappresentano la maggioranza significativa dei deepfake online[Ajd+19]

Figure 4.2: Misurazione risalente al settembre 2019 (+100% rispetto a dicembre 2018)[Ajd+19]

[1] [Ajd+19]
[2] [Sen]

Per quanto concerne le possibili azioni da intraprendere in ambito legale, se nessuna delle persone coinvolte nel deepfake ha dato il proprio consenso, entrambe potrebbero sollevare pretese in relazione alla violazione del proprio diritto all'immagine. Inoltre, nelle giurisdizioni in cui la pornografia è protetta dal diritto d'autore, gli autori del film possono sollevare reclami relativi alla modifica del video.

Il revenge porn è senza dubbio la peggiore variante di deepfake in quanto coinvolge contenuto esplicito senza il consenso della vittima e crea il risultato più umiliante per quest'ultima. Tuttavia la reazione sociale a questo genere di contenuti non è stata così ostile. Più in generale, sebbene i deepfake pornografici siano invadenti per la privacy e potrebbero essere considerati contrari all'ordine pubblico e immorali, un gran numero di comunità online sono piuttosto indifferenti. Questo può essere spiegato da diverse intuizioni comportamentali. Primo, fintanto che un individuo ottiene un certo piacere e non vi è alcuna minaccia per i propri diritti personali, proprietà o reputazione, l'individuo non è contro ciò di cui sta usufruendo. In secondo luogo, nel caso dei deepfake concernenti celebrità, le persone vedono ciò che vogliono sia vero, cioè la celebrità piuttosto che la persona sulla quale il volto è sovrapposto (fenomeno chiamato bias di desiderabilità[3]).

4.1.2 Minacce per la società: disinformazione

La diffusione della disinformazione è qualcosa che sta accadendo con crescente regolarità man mano che le persone scelgono le notizie che si adattano alla loro visione del mondo. I social media lo rafforzano poiché le persone interagiscono naturalmente con coloro che condividono convinzioni e prospettive simili, ma ora questa **guerra dell'informazione** viene utilizzata come arma anche dalle nazioni/Stati e i deepfakes contribuiranno solamente alla confusione che già esisteva.

I deepfake rappresentano un punto di svolta nella guerra dell'informazione. Aumenteranno la portata delle notizie false e ridurranno la nostra connessione a una comprensione condivisa dei fatti. Se le persone non possono fidarsi di ciò che vedono e sentono con i propri occhi e le proprie orecchie online, allora sceglieranno ciò in cui vogliono credere.

4.1.3 Minacce per la democrazia: influenza politica dei deepfake

Strettamente collegato a quanto ciò appena detto vi è la casistica in cui i deepfake vengano utilizzati allo scopo di influenzare la democrazie e la politica di uno Stato. Le false informazioni sulle istituzioni, la politica e i leader pubblici alimentate da un deepfake possono essere sfruttate per diffondere informazioni e manipolare le convinzioni. I deepfake renderanno difficile per le istituzioni, pubbliche e private, respingere gli attacchi reputazionali e sfatare la disinformazione già diffusa.

Il 17 aprile 2018, BuzzFeed[4] pubblicò un video deepfake del presidente Obama per dimostrare quanto fosse facile mettere parole in bocca a qualcun altro. In quel deepfake, il presidente Obama parlava con la propria voce e "imitava" le parole del creatore del video, alcune delle quali difficilmente sarebbero state pronunciate dal vero Obama.

I deepfake possono avere profonde conseguenze negative per le democrazie: notizie deepfake potrebbero mirare a prendere di mira la reputazione di determinati individui,

[3] [Ras19]
[4] [Buza]

Figure 4.3: Frame estratto dal video deepfake di Obama realizzato dal regista Jordan Peele[Buzb]

ritrarre eventi falsi (ad esempio, un falso attacco terroristico) o avere un impatto su processi democratici come campagne elettorali o altri eventi socialmente significativi. I deepfake possono essere usati come catalizzatori per erodere la fiducia nelle istituzioni politiche ed incrementare la divisione tra i gruppi sociali. Se utilizzati da governi ostili, potrebbero persino rappresentare una minaccia per la sicurezza nazionale o compromettere relazioni internazionali.

Prima delle elezioni presidenziali del 2016 è stata pubblicata una storia falsa che accusava Clinton e il suo presidente della campagna, John Podesta, di gestire un giro di abusi sui minori da un ristorante chiamato Comet Ping Pong. Mentre la storia, PizzaGate[5], si diffondeva, Comet Ping Pong ha ricevuto centinaia di minacce dai credenti della teoria. La polizia arrestò un uomo della Carolina del Nord dopo che questo sarebbe entrato in Comet Pizza con un fucile semiautomatico per "auto-indagare" sulla teoria, puntando l'arma contro un dipendente e sparando almeno un colpo. Sebbene questa storia falsa non usò alcun deepfake, quali sarebbero state le conseguenze se invece ci fosse stato un video a sostenere la cospirazione?

Presto detto, nel 2018 invece, la gente del Gabon sospettava che il loro presidente, Ali Bongo, fosse gravemente malato o fosse morto poiché assente dalla vita pubblica da diversi mesi. Per sfatare la speculazione, il governo Gabonese annunciò che il presidente ebbe avuto un ictus ma si trovava in buona salute, pubblicando un video di Bongo che pronunciava il discorso di Capodanno alla popolazione del Gabon. Entro una settimana, i militari lanciarono un colpo di stato senza successo. Tutto ciò accadde perchè il video venne considerato un deepfake da molti, sui social media. Successivamente, non fu mai stabilito se lo fosse in realtà, ma avrebbe cambiato il corso del governo in Gabon se il colpo di stato fosse riuscito. L'idea dei deepfake è sufficiente per accelerare il disfacimento di una situazione già precaria.

Oppure ancora, un'altra vicenda politica scaturita dall'utilizzo di deepfake, è stato lo scandalo emerso nel giugno 2019 che circondava il Ministro degli Affari Economici Malese Azmin Ali e il segretario di un ministro rivale, i quali presumibilmente appari-

[5][Piz]

vano insieme in un video sex tape.

Attività sessuali tra persone dello stesso sesso sono illegali in Malesia, con politici che sono stati in precedenza arrestati controversamente.

Mentre l'altro uomo affermò che il video era reale e creato senza consenso, Ali e i suoi sostenitori, incluso il Primo Ministro Malese, hanno sostenuto che il video fosse un deepfake realistico realizzato per sabotare la sua carriera politica. Successivamente, il segretario fu arrestato sotto accuse di sodomia. Il CEO di Deeptrace, , affermò che la qualità della risoluzione video era troppo bassa per eseguire un'analisi conclusiva. Nessuno degli analisti poté confermare che il signor Azmin fosse l'uomo nel video, l'unica conclusione raggiunta fu che il video non sembrò essere stato alterato digitalmente. [6]

Dal punto di vista normativo, trovare una risposta efficiente ai deepfake utilizzati per influenzare i processi politici è particolarmente impegnativo. Alcune sanzioni per la diffusione di informazioni false potrebbero essere imposte nelle leggi penali. Tuttavia, la concessione di diritti di "censura" eccessivi alle agenzie amministrative potrebbe essere impugnata come una restrizione incostituzionale della libertà di parola. Probabilmente, i politici le cui immagini sono utilizzate per creare deepfake diffamatori o falsi potrebbero cercare rimedi radicati nelle leggi sulla responsabilità civile o sul diritto d'autore. Il problema di questa argomentazione è che non fornisce rimedi efficaci per ripulire le conseguenze causate dai video diventati virali.

4.1.4 Minacce per le imprese: frodi e truffe

Uno studio[7] ha stimato che le aziende perdono circa 78 miliardi di dollari ogni anno a causa della disinformazione che le riguarda. La cifra include 9 miliardi di dollari per riparare i danni alla reputazione e altri 17 miliardi di dollari persi a causa della disinformazione finanziaria.

I deepfake nel campo aziendale, vengono utilizzati per impersonare le identità di leader aziendali e dirigenti, facilitando così le frodi.

Nel marzo 2019, l'amministratore delegato di un'azienda energetica con sede nel Regno Unito, ha ascoltato al telefono mentre il suo capo, il leader della società madre tedesca dell'azienda, ordinava il trasferimento di 220.000 euro a un fornitore in Ungheria. I notiziari avrebbero poi dettagliato che l'amministratore delegato ha riconosciuto il "lieve accento tedesco e la melodia[8]" della voce del suo capo e ha seguito l'ordine di trasferire il denaro (equivalente a circa 243.000$) entro un'ora. Il chiamante ha provato diverse altre volte a ottenere un secondo giro di denaro, ma a quel punto l'esecutivo inglese si è insospettito e non ha effettuato più trasferimenti. I 220.000 euro sono stati trasferiti in Messico e incanalati su altri conti, e la società energetica - che non è stata identificata - ha denunciato l'incidente alla sua compagnia di assicurazioni. Un funzionario di quest'ultima ha affermato che i ladri hanno utilizzato l'intelligenza artificiale per creare un deepfake della voce del dirigente tedesco.

Normativamente, c'è un consenso legale limitato sul chi sia il possessore di una voce, ovvero chi possa esercitare un diritto di proprietà su di essa. Molti degli Stati americani ad esempio, proteggono solamente nomi e sembianze/immagine, solo pochi proteggono la voce. La maggior parte di questi per di più protegge questi diritti solo per i vivi, con una minoranza di Stati che estendono la protezione dei diritti postumo.

[6] [Bla19]
[7] [Cas19]
[8] [Stu19]

Quindi, come si può tutelare la propria voce? Ebbene la voce di per sé, non può essere soggetta a copyright, ma bensì solamente le opere realizzate con essa (registrazioni, recitazioni, canzoni, o qualunque altra opera dell'ingegno). Ciò è stato dimostrato dal tentativo di un cantante[9] hip-hop americano nel far rimuovere deepfake audio di se stesso presenti su Youtube per violazione di copyright. Due[10] di questi video, dopo essere stati inizialmente rimossi, sono ritornati disponibili.

"Dopo aver esaminato le richieste di rimozione DMCA per i video in questione, abbiamo stabilito che erano incomplete", ha detto a The Verge[11] un portavoce di Google. *"In attesa di ulteriori informazioni da parte del ricorrente, abbiamo temporaneamente ripristinato i video."*

Abbiamo dunque appurato come l'imitazione vocale possa essere lecita nel caso in cui fosse espressamente dichiarata o palese, come accade nel caso dello spettacolo di un comico imitatore o di una parodia. Ma cosa accade invece se l'intento dell'imitatore fosse quello di generare confusione e far ritenere che la voce identifichi la persona imitata come nel caso delle frodi discusse sopra?

Un esempio statunitense: nel 1988, la Ford voleva utilizzare la voce della cantante Bette Midler per lo spot pubblicitario di una nuova autovettura. Non avendo ottenuto il consenso, la Ford interpellò un'altra cantante che riusciva ad imitarne perfettamente la voce. Il caso **Midler v. Ford Motor Co.**[12] proclamò che

"Una voce è distintiva e personale come un volto. La voce umana è uno dei modi più palpabili in cui si manifesta l'identità"[13].

Il tribunale Californiano ha ritenuto che una persona riceve protezione, in caso di appropriazione della sua voce, ai sensi della legge attraverso i suoi diritti della persona (*"right of publicity"* negli Stati Uniti).

Tuttavia, come già affermato sopra, non tutti gli stati garantiscono le stesse protezioni. A dimostrazione di ciò, possiamo affermare che nel nostro paese, Italia, la giurisprudenza si è espressa come segue a riguardo di alcuni casi concernenti la voce:

"mentre il ritratto di una persona ***permette di identificare senza alcuna difficoltà la persona ritratta, è assai difficile identificare una persona attraverso la voce"***[14]***.***

Di conseguenza, la giurisprudenza ha negato che le norme poste a protezione della immagine possano estendersi per analogia alla protezione della voce. In Italia, *"alla voce è riconosciuta una protezione* ***solamente*** *nei limiti della tutela degli artisti interpreti ed esecutori"*[15].

Se si comparano le due sentenze, è alquanto buffo notare che sono l'una l'opposto dell'altra. Eppure i deepfake possono venire utilizzati in entrambi i paesi, e soprattutto possono avere come bersaglio anche persone comuni e non necessariamente artisti. È quindi forse giunta l'ora che le leggi vengano aggiornate? O sarebbe sufficiente cambiare il punto di vista dal quale queste vengono interpretate?

[9] [Sta20]
[10] [Dft; Dfw]
[11] [Sta20]
[12] [Mid]
[13] *ibidem.*
[14] [Tut]
[15] *ibidem.*

4.2 Aspetti positivi

Procediamo ora ad un'analisi dei possibili utilizzi positivi dei deepfake.

4.2.1 Accessibilità e assistenza sanitaria

Con gli aggiornamenti del regolamento generale sulla protezione dei dati (GDPR) nell'UE, il libero flusso di dati è stato limitato per garantire il consenso e l'anonimato del paziente. Anche dati anonimizzati non devono essere condivisi tra gruppi di ricerca in paesi diversi, perché combinando poche variabili in un set di dati, si potrebbe risalire all'identità dell'individuo. Tutti i trasferimenti di dati sanitari richiedono che ogni paziente riceva il consenso informato. Tuttavia, per la medicina personalizzata sono necessari set di dati medici ad accesso aperto su larga scala e pubblicamente disponibili per migliorare le soluzioni machine learning in medicina.

È qui che i deepfake possono arrivare in soccorso. Con essi è possibile creare "pazienti artificiali" partendo dai dati di quelli reali. In questo modo sarà possibile alleviare il problema della privacy ma allo stesso tempo condividere dati (artificiali) con altri gruppi di ricerca.

Un altro possibile utilizzo generativo di deepfake in ambito medicina, è stato pubblicato in un recente studio[16], il quale ha utilizzato la tecnologia deepfake per creare dati sintetici di elettrocardiogrammi realistici che possono essere utilizzati a fini di ricerca; evitando, anche in questo caso, tutti i problemi relativi alla privacy.

Oltre a utilizzare i deepfake per la generazione di dati sintetici, ci sono altri possibili usi.

Per molti anni è stato possibile far parlare un computer digitando del testo in un'applicazione. Ora esiste la tecnologia deepfake per farlo con la voce di una determinata persona anche se non ha precedentemente registrato le parole in questione. Questa sta diventando una tecnologia che cambierà la vita alle persone che hanno perso la capacità di parlare in modo intelligibile, come persone affette da ictus o con una malattia progressiva come la sclerosi laterale amiotrofica (SLA). Il linguaggio sintetico può aiutare queste persone a parlare con la propria voce ai propri cari, anche dopo aver perso la possibilità di parlare, come si vede nel documentario del 2016, Gleason[17].

Oppure, ulteriore possibile caso d'uso, sotto la supervisione di un terapista, le persone possono ora avere conversazioni video realistiche con un deepfake di una persona cara deceduta, il che può dare loro consolazione e conforto come aiuto per superare la perdita.

4.2.2 Intrattenimento

L'industria del gaming è un'altra arena naturale per l'utilizzo dei deepfake.

Con audio sintetici deepfake, è possibile cambiare la tonalità della propria voce, e questo, a volte, ha portato a conseguenze positive impreviste. Ad esempio, delle cosiddette "skin vocali" consentono alle persone LGBT+ di cambiare la propria voce all'interno dei giochi, risultando in un gameplay più piacevole per loro. Una scoperta non sorprendente date le statistiche del 2020 dell'Anti-Defamation League[18] secondo cui più della metà degli utenti vengono molestati durante il gioco nelle chat vocali e il 37% dei giocatori LGBT+ viene molestato sulla base del proprio orientamento sessuale.

[16] [Tha+21]
[17] [Gle]
[18] [Tox]

Inoltre, la combinazione di impressionanti modelli di generazione del linguaggio naturale, come i linguaggi GPT(Generative Pretrained Transformer) sviluppati da OpenAI[19], abbinati a possibili utilizzi di deepfake all'interno di giochi, si tradurrà in NPC[20] (non-playable character, personaggio non giocante) in possesso della capacità illimitata di conversare con il proprio personaggio, con convincenti movimenti sincronizzati del viso e della bocca, senza la necessità di seguire script di codice specifici previsti dagli sviluppatori.

Figure 4.4: Lo sviluppatore di Modbox ha collegato il riconoscimento vocale di Windows, l'intelligenza artificiale GPT-3 di OpenAI e la sintesi vocale naturale di Replica per una demo unica: probabilmente il primo NPC AI[Gpta]

Un altro possibile utilizzo della tecnologia deepfake è per effetti speciali e editing facciale avanzato in post-produzione nell'ambito del cinema. In questo caso, può essere utilizzato per invertire l'invecchiamento e creare una versione più giovane di un attore o per poter realizzare comunque scene in cui un attore non è in grado di partecipare. Questa non è una novità poiché molti film hanno già utilizzato filmati generati da computer per scene in cui i personaggi originali erano interpretati da attori deceduti.

Rogue One è un film del 2016, il primo della serie Star Wars Anthology. In questo film, fu necessario far "ritornare in vita" due attori tramite gli effetti di post-produzione. Per il personaggio di Tarkin, originariamente interpretato da Peter Cushing, venne scelto l'attore britannico Guy Henry come sostituto. Uno dei motivi della scelta fu la somiglianza con l'attore scomparso. Per la Principessa Leia invece, la cui attrice originale anch'essa deceduta era Carrie Fisher, fu scelta l'attrice norvegese Ingvild Deila. Le tecnologia utilizzata fu la **CGI (computer-generated imagery)** abbinata a filmati d'archivio degli attori originali, modificati per ricrearne le fattezze.

[19][Ope]
[20][Gptb]

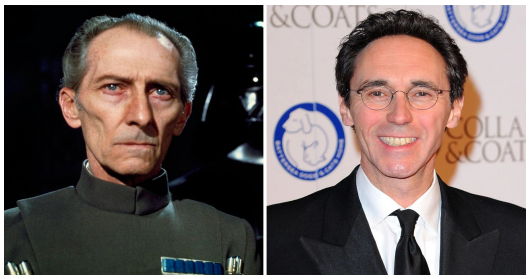

Figure 4.5: La CGI di Grand Moff Tarkin (a sinistra) e Guy Henry, l'attore che ha interpretato Tarkin sul set[Gia16]

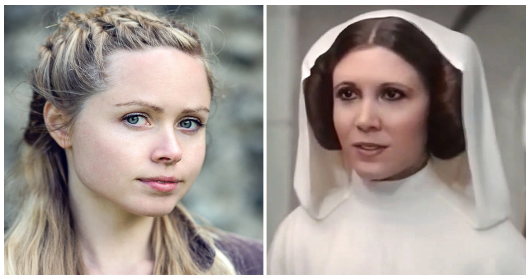

Figure 4.6: La CGI della Principessa Leia (a destra) e Ingvild Deila, l'attrice che ha interpretato Leia sul set[Cou17]

Figure 4.7: Confronto tra il Grand Moff Tarkin originale (sinistra) e la realizzazione con CGI[Lin18]

Figure 4.8: Confronto tra la principessa Leia originale (sinistra) e la realizzazione con CGI[Bon16]

Gli effetti speciali erano all'avanguardia all'epoca, ma già 4 anni dopo, sono iniziati a sembrare un po datati paragonati con la tecnologia deepfake fiorita negli anni successivi all'uscita del film. In particolare, uno youtuber di nome Shamook ha notevolmente migliorato l'aspetto di Grand Moff Tarkin e della Principessa Leia Organa nel film. Peter Cushing e Carrie Fisher sembrano più naturali nel video deepfake, che potrebbe persino far credere ad alcuni spettatori che sia il video reale.

Figure 4.9: Confronto tra la realizzazione con CGI (sinistra) e la realizzazione deepfake[Tar]

Figure 4.10: Confronto tra la realizzazione deepfake (sinistra) e la realizzazione con CGI[Tar]

"Usando un software deep fake sono riuscito a migliorare la versione CGI di Hollywood della Principessa Leia in Rogue One. Il processo ha richiesto solo 24 ore su un PC da 800$ e 500 immagini di Carrie Fisher nei film originali di Star Wars. Come puoi vedere, è inevitabile che Hollywood inizi a utilizzare questo metodo di VFX, quando e come è ancora la domanda."

Ad affermarlo è l'autore dei deepfake sopra mostrati, nella descrizione del video[21] sul suo canale Youtube.

Per decenni, Hollywood ha utilizzato tecnologie CGI, VFX e SFX di fascia alta per creare mondi artificiali ma credibili per una narrazione avvincente. I deepfakes possono democratizzare queste costose tecnologie come un potente strumento per narratori indipendenti a una frazione del costo.

Ulteriore caso d'uso: sebbene i sottotitoli consentano agli spettatori di vivere appieno la performance di un attore, possono distrarre gli spettatori e può essere difficile leggere e seguire l'azione allo stesso tempo. Il doppiaggio, d'altra parte, consente agli spettatori di concentrarsi meglio sul film ma può portare alla perdita di sfumature (es. accenti locali), ed è molto dipendente dalla qualità recitativa delle persone che fanno la voce fuori campo. Inoltre, la sincronizzazione labiale, ovvero l'abbinamento dei movimenti delle labbra di chi parla con l'audio preregistrato, è difficile e la corrispondenza non è sempre ottimale (i movimenti della bocca misti sono comuni). La tecnologia deepfake potrebbe essere un punto di svolta in questo senso. Normalmente pensiamo ai deepfake come alla manipolazione dell'intera immagine di una persona o di una scena, ma la tecnologia utilizzata da Flawless AI, si concentra su un solo elemento: la bocca[22]. I modelli di machine learning di Flawless studiano come gli attori muovono la bocca e quindi modificano i movimenti per sincronizzarsi perfettamente con le parole doppiate in diverse lingue. Grazie a questa tecnologia quindi quando il doppiaggio in lingua straniera è pronto, la rete neurale può modificare il volto dell'attore originale per sincronizzarsi perfettamente con il dialogo straniero. Tuttavia, vi è un ulteriore possibilità!

*"**Respeecher** consente ai produttori di film e ai creatori di contenuti di far sembrare*

[21] [Pri]
[22] [Vin21]

chiunque come se fosse qualcun altro[23] " (riferendosi sempre in ambito vocale). È quindi possibile clonare la voce dell'attore originale e modificare quelle dei doppiatori per farle sembrare come la prima.

Combinando entrambe le tecnologie, i produttori possono ottenere una qualità di doppiaggio eccezionale. L'animazione facciale modificata consente di trasferire la voce originale dell'attore in un'altra lingua. Pertanto, la voce del doppiaggio corrisponde alle espressioni facciali dell'attore originale. In più, il doppiaggio stesso è prodotto per dare l'impressione che sia l'attore stesso a parlare cinese o giapponese, per esempio. Ciò significa che gli spettatori non saranno in grado di affermare con certezza se l'attore sta parlando o meno nella sua lingua madre.

4.2.3 Educazione

I media sintetici generati dall'intelligenza artificiale possono riportare in vita figure storiche per un'aula più coinvolgente e interattiva. Un video sintetico di rievocazioni storiche o voce e video di un personaggio storico può avere più impatto, coinvolgimento e sarà uno strumento di apprendimento migliore.

Ad esempio, la risoluzione di John Kennedy per porre fine alla guerra fredda, un discorso che non fu mai pronunciato, è stato ricreato utilizzando tecniche IA con la sua voce sintetica e il suo stile di parola[24]. Ciò potrebbe portare gli studenti a conoscere il problema in modo creativo.

Anatomia umana sintetica, macchinari industriali sofisticati e qualunque altra cosa possa venire in mente, possono tutti essere modellati e simulati in un mondo di realtà mista per insegnare agli studenti, ai futuri medici, forze dell'ordine, o chiunque debba svolgere simulazioni in un corso di formazione, utilizzando un visore per la realtà virtuale come Microsoft Hololens[25].

L'uso creativo di voce e video sintetici può aumentare il successo complessivo e i risultati dell'apprendimento con dimensioni e costi limitati.

4.2.4 Autonomia e libertà di espressione

I deepfake, anonimizzando voci e volti, possono aiutare difensori dei diritti umani, attivisti e giornalisti che operano sotto regimi oppressivi per esprimere le proprie opinioni, diffondere notizie e rimanere anonimi. Tramite i deepfake, queste persone possono creare avatar digitali per l'espressione online che fornirebbero autonomia e privacy contribuendo così ad estendere i loro scopi, idee e convinzioni e consentire la libera espressione.

In conclusione, l'insieme degli esempi trattati lungo tutto il capitolo, evidenziano come l'IA sia una tecnologia abilitante, né intrinsecamente buona né cattiva. Una tecnologia dipende dal contesto in cui la si crea ed in cui la si utilizza.

[23] [Dfd]
[24] [Joh]
[25] [Hol]

5. Software per creazione deepfake

Il codice di machine learning alla base dei deepfake del 2017 che hanno scioccato il mondo fu rilasciato nel subreddit r/deepfakes, un forum che fu bloccato quasi non appena scoppiarono le controversie. Tuttavia, un utente riuscì a copiare il codice su GitHub prima del divieto, e da allora quel codice è stato "biforcato" (ovvero, altre persone hanno adottato o adattato il progetto) più di mille volte[1].

Tuttavia, gli output di solo due di quei fork, **DeepFaceLab** e **FaceSwap**, sono arrivati a dominare ciò che il pubblico ancora attualmente concepisce come "deepfake".

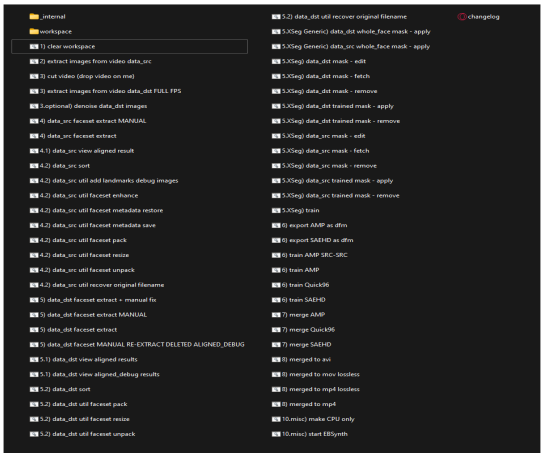

Figure 5.1: DeepFaceLab è composto da un insieme di file batch da eseguire in uno specifico ordine

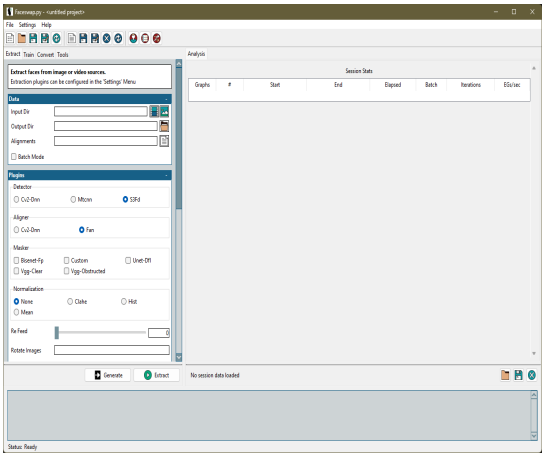

Figure 5.2: A differenza di DeepFaceLab, basato su riga di comando, FaceSwap ha invece una GUI intuitiva

Questi pacchetti software sono mantenuti su base volontaria da sviluppatori entusiasti open source e, al momento in cui si sta scrivendo, alimentano praticamente tutti i video deepfake virali che si possano trovare in rete.

I deepfaker SFW (safe for work, ovvero non contengono scene di nudo) più dedicati che utilizzano questi pacchetti, sono diventati celebrità di YouTube e TikTok. Diversi, come Shamook (di cui abbiamo parlato nei capitoli precedenti), l'indotto di ILM (Industrial Light and Magic, divisione effetti di LucasFilm)[2], e Ctrl-Shift-Face[3], sono passati alla produzione VFX professionale.

Come vedremo, alcune aziende VFX hanno incorporato questo codice open source

[1] [Dff]
[2] [Sha]
[3] [Ctr]

in sistemi chiusi e proprietari o hanno decostruito l'approccio in nuove architetture. Sebbene le società di produzione professionali possano avvalersi di costose risorse GPU per l'addestramento dei modelli, e di personale per curare e perfezionare i set di dati facciali che alimentano quest'ultimi, nonché risorse di sviluppo per migliorare i pacchetti originali, il codice in lenta evoluzione dietro DFL e FaceSwap rimane disponibile per chiunque sia interessato a dedicare le ore considerevoli necessarie per padroneggiare la curva di apprendimento.

DeepFaceLab (DFL) e Faceswap, usano entrambi un'architettura incentrata sugli autocodificatori, di cui abbiamo già spiegato il funzionamento nel capitolo 3.3.2.3.

Poiché gli autoencoder cercano dati "essenziali" da un'immagine, sono molto bravi a eliminare il rumore visivo. Nel caso di software per deepfake, ciò significa che un'architettura basata su questi, può apprendere caratteristiche e tratti fondamentali da un'immagine del volto, ignorando in gran parte fattori estranei come grana, ombre e altri elementi "non facciali" che possono essere presenti, risultando in un'immagine versatile e modelli ben generalizzati.

Dal 2017 ad oggi, sono emersi molti altri sistemi, applicazioni e progetti di ricerca che eseguono funzionalità uguali o simili. Molti dei quali piuttosto semplificati ed esclusivi per i dispositivi mobile, come l'app **Reface**[4] per iOS e Android.

La crescente ottimizzazione dei modelli deep learning e la maggiore potenza di calcolo dei dispositivi mobile, stanno iniziando a offrire capacità di deepfake più sostanziali anche ad utenti mobile, non tecnicamente al livello di utilizzare soluzioni come DFL, Faceswap o framework eseguibili esclusivamente su PC. Nell'aprile del 2022, una collaborazione accademica dalla Cina ha proposto MobileFSGAN[5], un sistema di autoencoder completo che pesa poco più di 10 MB, che può eseguire scambi di volti direttamente su telefoni iOS e Android.

In questo capitolo quindi, ci poniamo l'obiettivo di ricercare e confrontare tra di loro, le possibili soluzioni software disponibili alla creazione di deepfake, comprendendo nell'insieme anche applicazioni mobile. Il perché di quest'ultima scelta è semplicemente dato da quanto già detto in precedenza ad inizio capitolo: in ambito deepfake video, soltanto due dei migliaia di fork eseguiti sul framework originale del 2017 sono sopravvissuti fino ad oggi e vengono tutt'ora mantenuti aggiornati. Non avrebbe quindi alcun senso mettere a paragone software regolarmente aggiornati con altri che non lo sono, e i quali risultati risultano ovviamente più scandenti poiché datati.

[4] [Ref]
[5] [Yu+22]

5.1 Soluzioni software PC

Per quanto concerne le soluzioni software proposte per una piattaforma PC, la comparazione principale avviene tra quei due programmi che si trovano, ovviamente, sulla vetta della classifica in riferimento alla qualità dell'output nell'ambito deepfake video: DeepFaceLab e Faceswap.

Di seguito vengono riportate le loro principali differenze.

Caratteristiche	**DeepFaceLab**	**Faceswap**
Formato grafico	A linea di comando	GUI (Graphical user interface)
Massima risoluzione di training supportata	640px [6]	1024px [7]
Modelli di training disponibili	3 modelli tra cui scegliere	9 modelli tra cui scegliere
Pre-training	Consente l'utilizzo di modelli pre trainati da altri utenti per velocizzare l'inizio del proprio processo di training, in quanto la rete, in questo modo, sarà già a conoscenza del concetto di faccia	Non consiglia di svolgere pre training, ma implementa una funzione analoga che consente di caricare nella rete neurale i valori dei pesi da un modello già precedentemente addestrato, per "dare il via" all'inizializzazione del nuovo. La rete dovrà imparare dall'inizio il concetto di faccia potendo richiedere più tempo necessario per il processo di training
Opzioni	Quasi nessun opzione da modificare se non quelle relative al training, durante il suo svolgimento	Molto personalizzabile, con opzioni modificabili per ogni tipologia di modello

Tabella 5.1: Principali differenze tra i due software

Riassumendo quindi, DeepFaceLab è un programma caratterizzato dall'esecuzione in un specifico ordine di file batch, prevede quindi un'interazione con l'utente a linea di comando, a differenza di Faceswap che è dotato invece di una GUI.

Tuttavia esiste un'alternativa software che consente a DeepFaceLab di ottenere una interfaccia grafica anch'esso. Si tratta di un programma chiamato **Machine Video Editor**, il quale implementa al suo interno tutto il workflow di DeepFaceLab, con l'aggiunta di una semplificazione anche nel processo di creazione delle maschere dei volti.

[6] [Dflb]
[7] [Fac]

Figure 5.3: Editor XSEG originale di DeepFaceLab: l'utente dovrà segmentare il perimetro delle maschere in un numero limitato di volti, il programma proverà ad inferire automaticamente le altre durante il training XSEG

Figure 5.4: Machine Video Editor facilita il procedimento creando in automatico delle maschere generiche man mano che si scorrono le immagini e permette di modificarle semplicemente con uno strumento "pennello" anziché mediante segmentazione

È quindi possibile stabilire quale dei due software sia migliore dell'altro? Probabilmente no. Entrambi hanno comunità di utilizzatori affermate, e sono tutt'oggi mantenuti aggiornati.

Se si mira ad ottenere video deepfake della migliore qualità possibile, uno di questi due programmi dovrà sicuramente essere la scelta, a patto di avere il tempo necessario da investire per apprenderne il funzionamento e curare i dataset di volti da fornirgli in input.

Tuttavia possiamo ricorrere a delle alternative, qualora fossimo disposti a rinunciare alla qualità del risultato, risparmiando però sul tempo o qualora non avessimo un hardware sufficientemente buono per eseguire i comandi sopra citati.

Vi sono infatti una serie di soluzioni web che permettono la creazione di semplici deepfake, a partire da una sola foto della persona da sovrapporre ed un video della persona da sostituire. È ovviamente importante notare che queste soluzioni non si avvicinano minimamente alla complessità dei risultati che è possibile ottenere con DFL o Faceswap, in aggiunta al fatto che la maggior parte di questi servizi web richiedano un pagamento in base al tempo di utilizzo, il sottoscritto si trova pertanto a sconsigliare l'utilizzo di questi ultimi soprattutto dato che, ricercando abbastanza in rete, è possibile trovare alternative gratuite dello stesso genere.

Nella tabella a pagina seguente vengono riportati alcuni esempi.

Caratteristiche	**Deepswap**	**Deepfake Web**	**Faceswap Akool**
Quota di utilizzo	Richiede l'acquisto di crediti per eseguire l'upload di file (quota condivisa tra immagini e video) in aggiunta ad un abbonamento mensile se si vuol eseguire più di 2 swap gratuiti concessi.	Richiede l'acquisto dei minuti di utilizzo per le risorse (GPU) condivise tramite cloud. Il minimo obbligato sono 5 ore (300 minuti) a 15$, con una stima di consumo di 240-360 minuti per la creazione di un singolo video senza abbonamento premium.	Richiede anch'esso l'acquisto di una quota in minuti, che vengono però scalati all'upload dei video in base alla durata, in aggiunta all'acquisto di una quota anche per l'upload di immagini (quote separate tra immagini e video). A ciò si aggiunge il consueto abbonamento premium mensile, opzionale, per benefits come rimozione del watermark.
Prova gratuita	Non fornisce crediti iniziali e quindi, alcuna prova gratuita.	Non fornisce minuti iniziali e quindi, alcuna prova gratuita.	Fornisce una quota di upload per prova gratuita di 20 secondi video e 4 immagini.
Durata massima video	2 minuti per utenti non premium, 10 per utenti premium.	150 secondi per utenti non premium (max 50Mb), 300 secondi per utenti premium (max 100Mb).	3 minuti per utenti premium.

Tabella 5.2: Tre servizi web di faceswapping

È quindi evidente come nessuno dei tre sopra citati sia uno strumento economicamente vantaggioso, costringendoci a pagare una quota sproporzionata per dei brevi video di qualche minuto e di una qualità anche dubbia. Ma come precedentemente anticipato, esistono alternative gratuite che svolgono praticamente le stesse funzioni dei servizi sopra citati. Una di queste, è un software che si chiama **Swapface**[8], il quale, seppur restituendo risultati di uno stesso livello qualitativo delle sue controparti web, non ci costringe perlomeno al pagamento di una somma esagerata. Di seguito vengono confrontati due frame dei video realizzati con Swapface e il servizio web Face Swap Akool usufruendo dei secondi di prova gratuita.

[8] [Swa]

Figure 5.5: Video originale [Ren]

Figure 5.6: Video originale [Drb]

In questa occasione, si è provato ad dare in input al software una clip video di Matteo Renzi, al quale si è richiesto di sovrapporre il viso di Mr.Bean. Come è possibile notare dai risultati mostrati di seguito, il risultato non è dei migliori. I due nevi (o nei) presenti sulla guancia sinistra di Mr.Bean non sono individuabili in nessun momento del video ricevuto in output, tuttavia invece, i nei presenti sul viso di Matteo Renzi (due sopra l'occhio sinistro ed uno sotto il labbro inferiore) scompaiono e ricompaiono durante il video a piacimento. Il fenomeno risulta piu evidente nel video ricevuto in output dal servizio web Face Swap Akool rispetto a quello ricevuto dall'applicazione desktop Swapface.

Figure 5.7: Frame video di output Swapface

Figure 5.8: Frame video di output Face Swap Akool

Punto a favore da tener presente tuttavia, è l'aver ricevuto i video in output nel lasso di tempo di qualche minuto. Tenendo quindi a mente questo particolare, è giusto supporre che le reti neurali nel backend di questi software, non avrebbero potuto imparare in nessun modo i particolari di un viso in così poco tempo. Per paragone, i volti durante il processo di training in DeepFaceLab iniziano ad assumere definizione solamente dopo ore. Non è quindi giusto affermare che questi strumenti non valgano nulla. Il problema relativo alle soluzioni web sopra trattate, secondo il sottoscritto, sorge nel momento in cui viene chiesto un prezzo non rapportato a ciò che viene offerto. Nelle due immagini appena sopra presenti per esempio, confrontare la parte in alto a destra (più evidente) o sinistra dei volti: nel caso dell'immagine 5.8 (servizio a pagamento) risulta molto ben visibile un taglio netto nella colorazione del viso, non presente invece nel video originale fornito in input al servizio né tanto meno nell'output fornito dalla controparte software

gratuita. Questi strumenti, trovano quindi un utilizzo piuttosto limitato, poiché le condizioni richieste per poter restituire un risultato ottimale sono molto ristrette. Basti pensare che, se una cosa normalissima come dei nei sul viso abbiano già creato delle difficoltà, un paio di occhiali o dei capelli lunghi che potrebbero coprire una parte di viso probabilmente peggiorerebbero soltanto il risultato.

Un'altra tipologia di software invece, è quella dei programmi che consentono di eseguire un faceswap in tempo reale, sull'input video fornito da una webcam o qualsiasi dispositivo analogo. Anche in questo ambito, la soluzione software migliore è probabilmente quella presentata dagli stessi sviluppatori di DeepFaceLab, ovvero **DeepFaceLive**. Il programma offre un elenco di modelli di volti pubblici pronti all'uso, oppure la possibilità di utilizzare i propri modelli trainati tramite DeepFaceLab.

5.2 Soluzioni mobile

Passando invece a parlare di applicazioni per dispositivi mobile, le alternative che ci si presentano sono già svariate. Inutile precisare come queste non rappresentino una soluzione alla creazione di deepfakes professionali, così come i servizi web. Ma, da questi ultimi, si distinguono per una differenza principale: la loro accessibilità a tutti senza dover pagare quote di utilizzo. Risultano quindi opzioni già più ragionevoli, anche se solo per la creazione di qualche video divertente da condividere con amici o di meme.

Presentiamo di seguito alcuni esempi.

Nome	**Breve descrizione**
Reface	Reface è una delle app deepfake più conosciute. Sovrappone il proprio viso a immagini, meme e GIF utilizzando l'intelligenza artificiale per lo scambio di volti. Inoltre, è possibile importare nell'applicazione anche le proprie gif in modo da utilizzarle per lo scambio volto. Contiene pubblicità, per via della sua natura gratuita.
Wombo	Un'app mobile di sincronizzazione labiale. Wombo trasforma i selfie in deepfake con sincronizzazione labiale; è sufficiente caricare un selfie, scegliere una canzone e l'app si prenderà cura del resto. In realtà, alcuni risultati, trovabili in rete, realizzati da quest'app, mostrano risultati riusciti, abbastanza divertenti, anche fornendo in input all'applicazione foto di persone non reali, come personaggi di videogiochi o cartoni animati.
FaceApp	È un'app che consente agli utenti di modificare delle foto mediante l'IA, offrendo una varietà di sfondi, effetti, filtri, che possono essere utilizzati per modificare la propria apparenza. La maggior parte dei filtri è disponibile solamente per chi ha un abbonamento attivo, ma c'è da ammettere che i risultati sembrano alquanto realistici.

Tabella 5.3: Applicazioni mobile utilizzanti l'IA per la creazione di media sintetici

Ovviamente, esistono molte altre alternative alle tre proposte sopra mostrate. Si è semplicemente deciso di limitare il numero degli esempi alle applicazioni più conosciute nell'ambito, appunto per l'infinità di opzioni possibili, che svolgono infine le stesse funzioni.

Figure 5.9: Wombo applicato a Goro Majima, personaggio della serie videoludica Yakuza, nel videogioco Yakuza 0 del 2015.[Woma]

Figure 5.10: Wombo applicato a Mr.House, personaggio nel videogioco Fallout New Vegas del 2010.[Womb]

Un'applicazione sopra non citata, ma comunque degna di nota, è **Zao**, un'app deepfake di origine cinese gratuita sia per iOS che Android che svolge anch'essa una funziona di faceswapping su volti di attori o persone famose in modo abbastanza convincente. Unico motivo per cui quest'ultima non è stata citata insieme alle precedenti, è dato dal fatto che tutte le app di cui sopra sono scaricabili dagli store ufficiali (Play Store e Apple Store), mentre Zao si ottiene installando un file .apk (android package) ottenibile dal suo sito ufficiale[9]. È richiesto per cui, un livello di fiducia maggiore in ciò che si sta scaricando.

Figure 5.11: Faceswap eseguito con Zao su una scena di Game Of Thrones[Zaob]

[9] [Zaoa]

6. Creazione di un deepfake video

In questo capitolo, si analizzerà il processo di creazione di un deepfake professionale fornendo una panoramica generale del procedimento, indipendentemente dall'utilizzo di DeepFaceLab o Faceswap come soluzione software. Inoltre, si esplorerà il problema della gestione della batch size durante il training dei deepfake e del bottleneck presentato dalle GPU e il loro quantitativo di VRAM. L'obiettivo di questo capitolo è di fornire una comprensione completa del processo di creazione di un deepfake e dei problemi pratici che possono sorgere durante la loro generazione, utilizzando software di questo tipo.

6.1 Processo di creazione

6.1.1 Raccolta di dati

In primo luogo, è essenziale creare un set facciale: una raccolta di immagini di volti da cui l'autoencoder trarrà informazioni essenziali su quella particolare identità.

Pertanto dobbiamo ottenere molte immagini del soggetto di origine o sorgente (volto che vogliamo sovrapporre) e del soggetto di destinazione (volto che verrà sovrapposto).

La maggior parte degli sviluppatori e dei professionisti del deepfake consigliano di estrarre e curare da 5 a 10.000 immagini per ogni soggetto, idealmente da videoclip di alta qualità con illuminazione varia e con pose ed espressioni diverse.

Set facciali dedicati o universali

Poiché possono essere necessarie 1-2 settimane per addestrare un modello di alta qualità, anche su una GPU con specifiche adeguate, sarebbe fantastico se quel modello potesse essere utilizzato per inserire il soggetto di origine in qualsiasi clip con il soggetto di destinazione.

Tuttavia, un set di volti dedicato, che comprenda esclusivamente fotogrammi del soggetto origine in una particolare clip, produrrà un modello più accurato, anche se è probabile che quel modello funzioni male su qualsiasi altra clip.

Questo perché, durante l'addestramento, la rete neurale dedicherà una parte molto più ampia delle sue risorse alle caratteristiche esatte della clip sorgente, di solito risultando in un eccellente successivo scambio di volti, ma producendo un modello costoso e dispendioso in termini di tempo che non può essere utilizzato per qualsiasi altra attività.

Al contrario, l'utilizzo di immagini del volto non specifiche e diverse di una persona si tradurrà in un modello meglio generalizzato e in grado di eseguire buoni scambi su molti video diversi, ma non con la stessa qualità di un modello addestrato su un dataset di una clip specifica.

Ovviamente, il set di volti sorgente dovrà essere diverso e leggermente casuale, perché non esiste un filmato reale del soggetto origine che interpreta il ruolo del soggetto destinazione; ma le clip dell'identità sorgente dovrebbero essere scelte in base a quanto i volti sono simili alla quelli contenuti nella clip che vogliamo modificare (ovvero, ove possibile, grana, illuminazione, espressioni, ecc.).

6.2 Estrazione, riconoscimento tratti facciali e creazione maschera

Sia che scegliamo di addestrare un modello universale o dedicato (cioè specifico per la clip), la fase successiva consiste nel far analizzare al software tutte le immagini curate finora; per stimare le pose, comprese le espressioni facciali, presenti nelle immagini dei volti; e per creare maschere che delimitano il volto estratto, in modo che solo il materiale del volto (e non sfondi, capelli, orecchini e altri elementi estranei) venga addestrato nella rete neurale.

Il processo di estrazione itera attraverso l'intero set di volti, utilizzando una rete di allineamento facciale (FAN, Facial Alignment Network)[1] per dedurre le pose del viso, comprese le espressioni facciali come sorridere o urlare. Questi allineamenti FAN vengono utilizzati per generare aree di maschere che rivelano solo il contenuto del volto. Se i volti sono ostruiti da elementi vaganti come dita, capelli o persino occhiali, le maschere dovrebbero escludere quel contenuto.

DeepFaceLab affronta il problema con un programma dedicato chiamato XSeg, in cui l'utente disegna manualmente le maschere ogni un certo numero di fotogrammi e XSeg tenta di "riempire" (o "interpolare") le maschere intermedie, sulla base dell'input manuale. Anche FaceSwap dispone di un editor dedicato di maschere e allineamento. In caso di fallimento nella generazione di maschere in qualche frame, le singole maschere potranno essere ridisegnate manualmente o modificate, sia in DeepFaceLab che in FaceSwap.

I volti ora delimitati vengono estratti in immagini ritagliate più piccole, che vengono salvate in una nuova cartella di "volti di addestramento". Questo è ciò che verrà passato come input alla rete dell'autoencoder.

Sia DFL che FaceSwap includono strumenti per accelerare la rimozione manuale di identità "indesiderate" e falsi riconoscimenti, come persone "extra" nell'inquadratura o casi in cui il componente di riconoscimento facciale deduce un volto dove non esiste.

Queste procedure preliminari vengono eseguite sia sull'identità di origine che su quella di destinazione, risultando in una cartella separata di immagini per ciascuna persona.

6.3 Training

A questo punto inizia l'addestramento e dobbiamo scegliere un tipo di modello. Attualmente sono disponibili nove modelli in FaceSwap, con diversi livelli di requisiti hardware, flessibilità, capacità e facilità d'uso. Al contrario, DeepFaceLab offre attualmente solo tre modelli: SAEHD (la scelta standard e da tempo portato su FaceSwap), il quasi altrettanto popolare AMP e un modello "tester" leggero chiamato Quick96.

Le facce estratte vengono inserite nel modello in batch ed elaborate nella GPU. Il modello converte le immagini in informazioni vettoriali e procede all'estrazione dei tratti

[1] [Fan]

fondamentali da ciascuna immagine, costruendo lentamente un database di informazioni relative a ciascuna identità.

Come possiamo vedere nell'immagine sopra, il modello sta imparando a ricreare ogni identità all'interno dello spazio latente del codificatore condiviso, e alla fine produrrà due decodificatori produttivi, uno per ogni identità.

Per questo motivo, i set di dati devono avere il maggior numero possibile di pose in comune. Se c'è un'immagine del soggetto A che guarda verso l'alto e nessuna immagine/posa simile per il soggetto B, il modello non può imparare a creare una buona transizione tra le due identità per quella posa, perché manca metà delle informazioni necessarie. Allo stesso modo, è necessaria un'ampia varietà di espressioni corrispondenti tra i due set di dati. Se il soggetto A non sorride mai e il soggetto B sorride molto, il modello non potrà mai imparare a rappresentare accuratamente il soggetto A sorridente.

Dopo tipicamente 3-14 giorni di training, a seconda delle impostazioni e del volume di dati, il modello raggiunge la convergenza; il punto in cui, ulteriore addestramento diventa ridondante o addirittura dannoso.

6.4 Conversione

Il modello addestrato è ora in grado di ricreare abbastanza bene ogni identità; ma se cambi semplicemente i percorsi del decodificatore, il modello ora può anche imporre l'identità alternativa:

Le immagini alterate vengono automaticamente riassemblate in formato video alla fine del processo.

6.5 I deepfake continueranno davvero a migliorare?

I deepfake, che non sono più riconoscibili come tali, nemmeno dagli algoritmi di rilevamento per essi, potrebbero finalmente cambiare le regole del gioco, sia socialmente che nell'intrattenimento. L'esperto di deepfake Hao Li ritiene che questo sviluppo sia possibile, poiché le immagini in definitiva non sono altro che pixel opportunamente colorati: quindi, una copia perfetta è solo una questione di tempo[2].

Ma i deepfake stanno migliorando perché la tecnologia si sta davvero evolvendo rapidamente? L'obiettivo del continuo miglioramento incontra una serie di ostacoli. Diamo un'occhiata ad alcuni di essi e alle potenziali strade da percorrere.

La maggior parte dei principali miglioramenti alle popolari distribuzioni di software deepfake ha almeno un paio di anni. Alcuni deepfaker hanno imparato ad aggirare la carenza di miglioramenti tecnologi selezionando attentamente i videoclip adatti al processo e mediante un'ampia manipolazione di post-elaborazione, svolta con programmi come Adobe After Effects.

Consideriamo ora alcuni degli ostacoli esistenti al mantenimento della crescita del realismo per i deepfake basati su autoencoder.

[2] [Kni]

6.6 Bottleneck dato dalle GPU

Sebbene la carestia di GPU degli ultimi due anni sembri destinata a diminuire, è improbabile che i prezzi tornino ai livelli pre-pandemici, rendendo il deepfaking di alta qualità un hobby sempre più costoso, quasi alla pari con il mining di criptovalute in termini di costi hardware e consumo di elettricità.

In aggiunta, l'aumento del costo dell'elettricità necessaria per le settimane (o addirittura i mesi) di addestramento necessarie di un singolo modello di machine learning, può alla fine diventare un notevole ostacolo al deepfaking, in particolare qualora non si guadagni nulla dal risultato del processo.

Ma anche supponendo che nessuno dei problemi descritti sopra ci riguardasse, avendo quindi risorse illimitate, i deepfake si troveranno comunque ad affrontare un collo di bottiglia dato dall'architettura; il quale riguarda: il numero di immagini di addestramento che possono passare attraverso la GPU contemporaneamente in un dato momento, quanto possono essere grandi quelle immagini, e se le architetture possono scalare in modo efficace.

Ad esempio, la pagina GitHub di DFL, presenta una galleria/cronologia dell'aumento delle dimensioni delle immagini di addestramento, dall'avvento del progetto ad oggi.

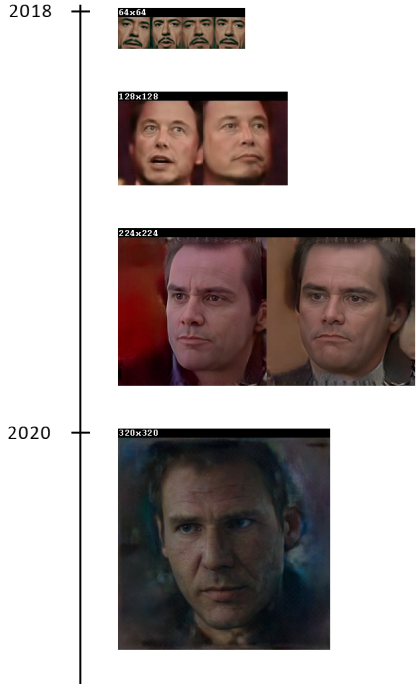

Figure 6.1: Dimensione immagini partendo dall'alto: 64x64, 128x128, 224x224, 320x320, ...

Figure 6.1: ... 448x448, 512x512[Dfla]

Si noti che l'ultimo esempio (Morgan Freeman) richiede 24 GB di RAM video (VRAM), con le schede video adatte a soddisfare questo requisito che vanno da un prezzo di 1200 a 2700€ attualmente, soggette a disponibilità.

Infatti, le dimensioni tutt'oggi ancora utilizzate per la produzione di deepfakes si aggirano intorno ai 512x512 pixels. Ed è per questo motivo, che ci sono stati relativamente

pochi esempi di deepfake nella produzione cinematografica e televisiva professionale.

Quindi, perché non ottenere semplicemente una GPU più grande, utilizzare immagini più grandi e ottenere un output maggiore?

Parte del problema è legato alle dimensioni del batch, ovvero quante immagini il processo di addestramento può esaminare in qualsiasi momento.

6.7 Gestione della batch size

Le immagini vengono addestrate in batch (gruppi). È possibile impostare una dimensione minima del batch pari a 1, mentre non esiste un limite superiore teorico né per le dimensioni del batch né per le dimensioni dell'immagine (che sono invece vincolate dalle limitazioni dell'hardware disponibile).

Tuttavia, più grandi sono le immagini di addestramento, minore è il numero di immagini che possono essere inserite nella GPU in ogni singolo batch.

Quindi, per quanto riguarda la dimensione del batch, più grande non sta ad indicare sempre un risultato migliore.

Maggiore è il numero di immagini in un batch, maggiore è la probabilità che il modello finale si generalizzi bene ma non riesca a catturare i dettagli intrinseci all'identità, risultando in una somiglianza più "vaga" con l'obiettivo. Mentre se si utilizzano batch più piccoli, verrà richiesta un'attività minore della GPU e si consentirà al modello di concentrarsi meglio sui (meno) volti che attualmente passano attraverso la pipeline imparando anche i dettagli minori, al costo però di un aumento del tempo necessario per l'addestramento.

Questo perciò, è il motivo principale per cui né il denaro né la VRAM che si può acquistare con esso, possono risolvere tutti i problemi.

Non che il denaro sia un problema minore: se si è in grado di acquistare una delle GPU Nvidia Tesla A100 da 80 GB di VRAM, si è sicuramente in grado di sostenere alcune immagini 512x512px in un singolo batch, o dimensioni batch più elevate per immagini più piccole; ma il prezzo di 30.000€ è leggermente proibitivo per l'utilizzo amatoriale, ed in generale c'è un salto di costo enorme tra GPU consumer, relativamente convenienti, come la serie NVIDIA GeForce 3000 (30xx) (8-24 GB), o la nuova uscita serie 4000 (40xx), e schede di fascia alta spesso destinate a data center e uso industriale.

Tuttavia, 512x512px non è ancora una risoluzione di produzione; 1024x1024px è il minimo in termini di standard cinematografico.

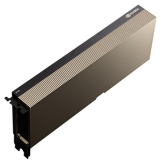

Figure 6.2: Scheda video Nvidia Tesla A100, dotata di 80Gb di VRAM, con un costo che si aggira sui 30 000€[A10]

Figure 6.3: Scheda video Nvidia RTX 4090, dotata di 24Gb di VRAM, con un costo che si aggira sui 2400€[Rtx]

7. Rilevamento deepfake e come riconoscerli

7.1 Stato della tecnologia di rilevamento: un gioco di guardie e ladri

Una serie di ricerche recenti, ha introdotto diversi metodi di rilevamento deepfake video. Alcuni di questi metodi dichiarano un'accuratezza di rilevamento superiore al 99%, ma tali rapporti sull'accuratezza dovrebbero essere interpretati con cautela poichè la difficoltà di rilevare la manipolazione video varia ampiamente in base a diversi fattori come il livello di compressione, la risoluzione dell'immagine ma soprattutto la composizione del dataset di test (ovvero dove si è testato l'algoritmo di rilevamento). Infatti, un'analisi comparativa delle prestazioni di diversi modelli rivelatori basata su cinque set di dati pubblici, che vengono spesso utilizzati nel campo, ha mostrato un'ampia gamma di accuratezza, che va dal 30% al 97%, senza che nessun singolo rivelatore sia significativamente migliore di un altro. In genere, i rilevatori saranno sintonizzati per cercare un certo tipo di manipolazione e spesso, quando questi rilevatori vengono rivolti a nuovi dati, non funzionano bene. Quindi, se è vero che ci sono molti sforzi in corso in questo settore, non è vero che ci sono alcuni rilevatori che sono di gran lunga migliori di altri.

Indipendentemente dalla precisione dei rilevatori attuali, l'ambito del deepfake detecting è un gioco di gatto e topo. I progressi nei metodi di rilevamento si alternano ai progressi nei metodi di generazione di deepfake, come quando, nel 2018, uno studio[1] dell'Università di Albany (nello stato di New York) scoprì che le persone in un deepfake tendevano a sbattere le palpebre molto più o molto meno frequentemente delle persone reali e, un anno dopo, i deepfake stavano già sviluppando soluzioni al problema per far apparire il battito delle palpebre più realistico.

Per tre mesi tra il 2019 e il 2020, Facebook (ora Meta) ha ospitato la Deepfake Detection Challenge, chiedendo ai partecipanti di automatizzare il processo per determinare se una foto è stata manipolata con l'intelligenza artificiale. Il concorso ha assegnato 1 milione di dollari in premi, agli iscritti con i migliori algoritmi di successo. Ma, anche con alcune delle menti più acute dell'intelligenza artificiale, il miglior programma è stato in grado di rilevare i deepfake solo il 65% delle volte.

I partecipanti erano stati incaricati di creare un modello di rilevamento addestrato e convalidato su un set di dati curato di 100.000 video deepfake. Sebbene originariamente il set di dati fosse disponibile solo per i membri della competizione, da allora è stato rilasciato pubblicamente. Degli oltre 35.000 modelli realizzati per l'occasione, quello vincente ha raggiunto una precisione del 65% su un dataset di test di 10.000 video, che

[1] [Dfe]

era stato riservato per il testing dei modelli, e dell'82% sul set di convalida utilizzato durante il processo di addestramento del modello. Il dataset di test non era stato reso disponibile ai partecipanti durante la formazione. La discrepanza nell'accuratezza tra i due dataset indica che c'era una certa quantità di **overfitting** (quando il modello impara le peculiarità del training set ma non riesce ad adattarsi a dati nuovi), ovvero una mancanza di generalizzazione, un problema che tende ad affliggere tutti i modelli di deepfake detection.

7.2 Come riconoscere un deepfake

Dopo aver assodato che il campo del deepfake detection non possa fornire una soluzione definitiva al problema, come ci si può difendere dai deepfake malevoli che è possibile trovare in rete?

Quando l'inventore delle GAN, Ian Goodfellow, presentò il suo lavoro nel 2014, probabilmente non aveva previsto il rapido sviluppo del fenomeno deepfake per come lo conosciamo. Oggi tuttavia, avverte: in futuro, le persone non dovrebbero più credere alle immagini e ai video su Internet come una cosa ovvia[2].

Il suggerimento principale quindi, rimane quello di non fidarsi immediatamente di quello che si sta vedendo, leggendo o ascoltando, ma bensì chiedersi se la sorgente del video, immagine o notizia sia affidabile; controllare le fonti qualora citate e presenti; controllare se la notizia è riportata anche da altri e ciò che questi dicono a riguardo;...

D'altronde, il metodo principale per combattere la disinformazione è informarsi.

Quando parliamo di contenuti video deepfake tuttavia, potrebbero esserci alcuni piccoli elementi, che se presenti, ci potrebbero aiutare nell'identificare un video come tale. Come abbiamo visto nel capitolo precedente, realizzare un deepfake non è un'operazione esente da complicazioni o problemi. Quindi, a meno che non si tratti di un deepfake generato da un professionista con attrezzatura di un certo livello, alcuni artefatti potrebbero essere stati lasciati nel video. Capiamo con precisione cosa guardare, all'interno del video, per capire se ciò che è presente dinanzi ai nostri occhi possa essere un deepfake oppure no.

- **Movimento innaturale dell'occhio**
  Un segnale di avvertimento comune sono i movimenti oculari dall'aspetto innaturale o la mancanza di movimento degli occhi, in particolare l'assenza di battito di ciglia. È difficile imitare il battito delle palpebre in un modo che sembri naturale. È anche difficile riprodurre accuratamente i movimenti oculari perché quando una persona parla con un'altra, i suoi occhi normalmente la seguono.

- **Espressioni facciali innaturali**
  Quando qualcosa non sembra giusto in un volto, potrebbe segnalare il morphing facciale. Ciò si verifica quando un'immagine è stata unita sopra un'altra.

- **Posizionamento innaturale dei tratti del viso**
  Si dovrebbe diffidare dei video che sembrano reali, se il viso e il naso di qualcuno sono puntati in direzioni diverse. Controllare quindi, anche la posizione del naso.

- **Mancanza di emozioni**
  Il morphing facciale si può individuare anche se il viso di qualcuno non sembra

[2] [Ian]

mostrare l'emozione che dovrebbe accompagnare quello che si suppone stia dicendo.

- **Corpo o postura dall'aspetto goffo**
  Un altro segno è se la forma del corpo di una persona non sembra naturale o c'è un posizionamento incoerente della testa rispetto al corpo. Questa potrebbe essere una delle incoerenze più facili da individuare, perché la tecnologia deepfake di solito si concentra sui tratti del viso piuttosto che sull'intero corpo.

- **Movimento o forma del corpo innaturali**
  Se qualcuno appare distorto o strano quando si gira di lato o muove la testa, oppure i suoi movimenti sono a scatti e sconnessi da un fotogramma all'altro, si dovrebbe sospettare che il video sia falso.

- **Discrepanze di colore e illuminazione**
  Tonalità della pelle insolite, macchie, luci e ombre posizionate in modo strano indicano che ciò che si sta vedendo potrebbe essere falso. Se si sta guardando un video sospetto, si prenda nota delle discrepanze nella persona e le si confronti con un riferimento originale. Questo aiuterà a determinare se si tratti di un deepfake o meno.

- **Capelli che non sembrano veri**
  Le immagini false non saranno in grado di generare caratteristiche individuali come capelli crespi o svolazzanti.

- **Denti che non sembrano veri**
  Gli algoritmi potrebbero non essere in grado di generare i singoli denti, quindi l'assenza di contorni di questi potrebbe essere un indizio.

- **Sfocatura o disallineamento**
  Se i bordi delle immagini sono sfocati o degli elementi visivi non sono allineati, ad esempio dove il viso e il collo di qualcuno incontrano il proprio corpo, c'è qualcosa che non va.

- **Suoni o audio incoerenti**
  I creatori di deepfake di solito dedicano più tempo alle immagini video piuttosto che all'audio. Il risultato può essere una scarsa sincronizzazione labiale, voci dal suono robotico, pronuncia di parole strane, rumore di fondo digitale o persino l'assenza di audio.

Vengono proposti alcuni esempi di frame estratti da un primo video deepfake realizzato dal sottoscritto, dove alcuni degli artefatti citati sopra saranno chiaramente individuabili.

Figure 7.1: Viso tutto sommato credibile, privo di artefatti, se non fosse che è possibile notare un colore più chiaro di quello del viso assunto dall'orecchio destro. Notare anche la differenza di colore tra viso e braccio destro.

Figure 7.2: Questo frame invece, risalta la differenza di colore in modo ancora maggiore: notare come il lato destro del visto assume una sfumatura di color giallo, tanto che anche l'occhio destro ha una colorazione differente dal sinistro. Restando in tema di occhi, è anche possibile notare una posizione innaturale di questi ultimi: osservare come l'occhio destro sembra essere rivolto verso l'alto mentre il sinistro sembra guardare il basso.

8. Conclusioni

In conclusione, la tecnologia dei deepfake rappresenta una sfida importante per la società moderna, non solo per le questioni legali ed etiche che suscita, ma anche per la sua capacità di sfidare la nostra percezione della realtà e della verità.

L'elaborazione di algoritmi di deep learning sempre più sofisticati, combinata con l'aumento della disponibilità di dati e il miglioramento dell'hardware informatico, sta rendendo i deepfake sempre più convincenti e difficili da rilevare. Questo solleva preoccupazioni riguardo alla manipolazione dell'informazione e alla possibilità di creare contenuti dannosi o fraudolenti.

Tuttavia, è importante sottolineare che la tecnologia dei deepfake può anche essere utilizzata per scopi positivi, come l'elaborazione di contenuti creativi e l'animazione di personaggi digitali. Inoltre, la ricerca sulla rilevazione dei deepfake sta facendo progressi significativi e potrebbe portare a soluzioni efficaci per mitigare i rischi associati a questa tecnologia.

Per affrontare efficacemente le questioni legate ai deepfake, sarà necessario un approccio olistico che coinvolga non solo i ricercatori informatici, ma anche i professionisti del diritto, i responsabili politici e la società nel suo insieme. Questo approccio potrebbe includere l'implementazione di politiche pubbliche per regolamentare l'uso dei deepfake, la promozione della consapevolezza e dell'alfabetizzazione digitale per il pubblico e la collaborazione tra gli attori coinvolti per identificare e mitigare i rischi associati alla tecnologia dei deepfake.

In definitiva, la tecnologia dei deepfake rappresenta una sfida per la società, ma anche un'opportunità per sperimentare e sviluppare nuove soluzioni tecnologiche e sociali. Il futuro dei deepfake dipenderà dalla nostra capacità di utilizzarli in modo responsabile e creativo, e di affrontare le sfide che essi pongono in modo collaborativo e proattivo.

9. Ringraziamenti

Vorrei innanzitutto ringraziare il prof. Fausto Marcantoni, relatore di questa tesi, per la disponibilità offerta. Successivamente ringrazio tutte quelle persone che mi hanno supportato in questo percorso, familiari e non; in particolare, i miei genitori. Ringrazio infine l'università di Camerino per l'opportunità offerta nello svolgere questo percorso di studi.

Bibliography

[A10] *NVIDIA A100*. URL: https://www.nvidia.com/en-us/data-center/a100/.

[Ajd+19] Henry Ajder et al. *The State of Deepfakes: Landscape, Threats, and Impact*. 2019. URL: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

[Alp] *AlphaGo*. URL: https://www.deepmind.com/research/highlighted-research/alphago.

[Art] *What is an Artificial Neuron?* URL: https://becominghuman.ai/what-is-an-artificial-neuron-8b2e421ce42e.

[Aut] *What is an Autoencoder?* URL: https://blog.roboflow.com/what-is-an-autoencoder-computer-vision/.

[Aws] *AWS Deepracer: il modo piu rapido per partire con il machine learning*. URL: https://aws.amazon.com/it/deepracer/.

[BCS97] Christoph Bregler, Michele Covell, and Malcolm Slaney. "Video Rewrite: Driving Visual Speech with Audio". In: vol. 31. Jan. 1997, pp. 353–360. DOI: 10.1145/258734.258880. URL: http://chris.bregler.com/videorewrite/.

[Bha19] Shripad Bhat. *Feature Engineering for machine learning*. 2019. URL: https://www.izen.ai/blog-posts/feature-engineering-for-machine-learning/#:~:text=Feature\%20engineering\%20refers\%20to\%20creating,converting\%20image\%20to\%20RGB\%20values..

[Bla19] Jarni Blakkarly. *A gay sex tape is threatening to end the political careers of two men in Malaysia*. 2019. URL: https://www.sbs.com.au/news/article/a-gay-sex-tape-is-threatening-to-end-the-political-careers-of-two-men-in-malaysia/ilgqdaqo5.

[Bon16] Josefina Bonnefont. *Así se ve el "resucitado" Tarkin y la joven Leia en "Rogue One: una historia de Star Wars"*. 2016. URL: https://www.upsocl.com/muvi/asi-se-ve-el-resucitado-tarkin-y-la-joven-leia-en-rogue-one-una-historia-de-star-wars/.

[Buza] *BuzzFeed*. URL: https://www.buzzfeed.com.

[Buzb] *You Won't Believe What Obama Says In This Video!* URL: https://www.youtube.com/watch?v=cQ54GDm1eL0.

[Cas19] Michelle Castillo. "Exclusive: Fake News Is Costing the World $78 Billion a Year". In: *Cheddar news* (2019). URL: https://cheddar.com/media/exclusive-fake-news-is-costing-the-world-billion-a-year.

[CET01] T.F. Cootes, G.J. Edwards, and C.J. Taylor. "Active appearance models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.6 (2001), pp. 681–685. DOI: 10.1109/34.927467.

[Cnna] *Come funziona una rete neurale CNN*. URL: https://www.domsoria.com/2019/10/come-funziona-una-rete-neurale-cnn-convolutional-neural-network/.

[Cnnb] *Convolutional Neural Network Tutorial*. URL: https://www.simplilearn.com/tutorials/deep-learning-tutorial/convolutional-neural-network#:~:text=A\%20convolutional\%20neural\%20network\%20is,classify\%20objects\%20in\%20an\%20image..

[Cnnc] *Convolutional Neural Networks (CNN) — Architecture Explained*. URL: https://medium.com/@draj0718/convolutional-neural-networks-cnn-architectures-explained-716fb197b243.

[Cnnd] *Reti neurali convoluzionali - Matlab*. URL: https://it.mathworks.com/discovery/convolutional-neural-network-matlab.html.

[Col17] Samantha Cole. "AI-Assisted Fake Porn Is Here and We're All Fucked". In: *Vice* (2017). URL: https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn.

[Col18] Samantha Cole. "We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now". In: *Vice* (2018). URL: https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley.

[Con] *Matrice di convoluzione - Wikipedia*. URL: https://it.wikipedia.org/wiki/Matrice_di_convoluzione.

[Cou17] Aaron Couch. *'Rogue One': Leia Actress on the Pressures of Re-Creating an Iconic Character*. 2017. URL: https://www.hollywoodreporter.com/movies/movie-news/rogue-one-leia-actress-ingvild-deila-creating-star-wars-character-986775/.

[Ctr] *Ctrl Shift Face - Youtube*. URL: https://www.youtube.com/c/CtrlShiftFace?app=desktop.

[Deea] *Deepfake: dal Garante una scheda informativa sui rischi dell'uso malevolo di questa nuova tecnologia*. 2020. URL: https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9512278#:~:text=I\%20deepfake\%20sono\%20foto\%2C\%20video,imitare\%20fedelmente\%20una\%20determinata\%20voce..

[Deeb] *Deepfake in Vocabolario - Treccani*. 2018. URL: https://www.treccani.it/vocabolario/deepfake_\%28Neologismi\%29/.

[Dfa] *Understanding the Technology Behind DeepFakes - Alan Zucconi*. URL: https://www.alanzucconi.com/2018/03/14/understanding-the-technology-behind-deepfakes/.

[Dfd] *What Is Synthetic Film Dubbing: AI Deepfake Technology Explained*. 2021. URL: https://www.respeecher.com/blog/synthetic-film-dubbing-ai-deepfake-technology-explained.

[Dfe] *Exposing Fake Videos - University at Albany*. URL: https://www.albany.edu/news/87379.php.

[Dff] *Forks - joshua-wu deepfakes_faceswap*. URL: https://github.com/joshua-wu/deepfakes_faceswap/network/members.

[Dfla] *GitHub - iperov DeepFaceLab)*. URL: https://github.com/iperov/DeepFaceLab.

[Dflb] *Using over 640 size image*. URL: https://github.com/iperov/DeepFaceLab/issues/5421.

[Dft] *Jay-Z raps the "To Be, Or Not To Be" soliloquy from Hamlet (Speech Synthesis)*. URL: https://www.youtube.com/watch?v=m7u-y9oqUSw.

[Dfw] *Jay-Z covers "We Didn't Start the Fire" by Billy Joel (Speech Synthesis)*. URL: https://www.youtube.com/watch?v=iyemXtkB-xk.

[Dim] *Dimensions*. URL: https://app.dimensions.ai/discover/publication?search_mode=content&search_text=Deepfake&search_type=kws&search_field=full_search.

[Drb] *DOCTOR Bean Arrives — Mr Bean: The Movie — Funny Clips — Mr Bean Official*. URL: https://www.youtube.com/watch?v=1dW_I_d6oQA&t=211s.

[Fac] *Deepfakes Go High-Res – But Can Deepfakers Handle It?* URL: https://metaphysic.ai/deepfakes-go-high-res-deepfakers-handle/#:~:text=\%27The\%20main\%20difference\%20is\%20that,obtain\%20a\%20more\%20accurate\%20resemblance..

[Fan] *GitHub - 1adrianb face-alignment*. URL: https://github.com/1adrianb/face-alignment.

[Fuk80] Kunihiko Fukushima. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: *Biological Cybernetics* 36.4 (1980), 193–202. DOI: 10.1007/bf00344251.

[Gan] *Overview of GAN Structure - Google Developers*. URL: https://developers.google.com/machine-learning/gan/gan_structure?hl=it.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[Gia16] Carolyn Giardina. *'Rogue One': How Visual Effects Made the Return of Some Iconic 'Star Wars' Characters Possible*. 2016. URL: https://www.hollywoodreporter.com/movies/movie-news/rogue-one-how-grand-moff-tarkin-peter-cushing-returned-957258/.

[Gle] *Gleason*. URL: https://www.imdb.com/title/tt4632316/.

[Goo+14] Ian J. Goodfellow et al. "Generative Adversarial Networks". In: (2014). DOI: 10.48550/ARXIV.1406.2661. URL: https://arxiv.org/abs/1406.2661.

[Gpta] *OpenAI GPT-3 Powered NPCs: A Must-Watch Glimpse Of The Future (Modbox)*. URL: https://www.youtube.com/watch?v=jH-6-ZIgmKY.

[Gptb] *This OpenAI GPT-3 Powered Demo Is A Glimpse Of NPCs In The Future*. URL: https://uploadvr.com/modbox-gpt3-ai-npc-demo/.

[Hol] *Mixed Reality for Education — Microsoft Education*. URL: https://www.microsoft.com/en-us/education/mixed-reality.

[Ian] *Transcript of interview of Ian Goodfellow by Lex Fridman*. URL: https://www.linkedin.com/pulse/transcript-interview-ian-goodfellow-lex-fridman-alfonso-r-reyes.

[Joh] *JFK Unsilenced*. URL: https://www.cereproc.com/it/jfkunsilenced.

[Kam17] Izabella Kaminska. "A lesson in fake news from the info-wars of ancient rome". In: *Financial Times* (Jan. 2017). URL: https://www.ft.com/content/aaf2bb08-dca2-11e6-86ac-f253db7791c6.

[Kar+17] Tero Karras et al. *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. 2017. DOI: 10.48550/ARXIV.1710.10196. URL: https://arxiv.org/abs/1710.10196.

[Ker] *Keras: the Python deep learning API*. URL: https://keras.io.

[KLA18] Tero Karras, Samuli Laine, and Timo Aila. *A Style-Based Generator Architecture for Generative Adversarial Networks*. 2018. DOI: 10.48550/ARXIV.1812.04948. URL: https://arxiv.org/abs/1812.04948.

[Kni] Will Knight. *The world's top deepfake artist is wrestling with the monster he created*. URL: https://www.technologyreview.com/2019/08/16/133686/the-worlds-top-deepfake-artist-is-wrestling-with-the-monster-he-created/.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[LeC+89] Yann LeCun et al. "Handwritten Digit Recognition with a Back-Propagation Network". In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky. Vol. 2. Morgan-Kaufmann, 1989. URL: https://proceedings.neurips.cc/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf.

[Lin18] Lucas Roos Lindgreen. *Blurring the Line Between Real and Fake: the Dangers of DeepFakes*. 2018. URL: https://medium.com/@lucasrooslindgreen/blurring-the-line-between-real-and-fake-the-dangers-of-deepfakes-14894effe295.

[LT16] Ming-Yu Liu and Oncel Tuzel. *Coupled Generative Adversarial Networks*. 2016. DOI: 10.48550/ARXIV.1606.07536. URL: https://arxiv.org/abs/1606.07536.

[Mid] *Midler v. Ford Motor Co. - 849 F.2d 460 (9th Cir. 1988)*. URL: https://www.lexisnexis.com/community/casebrief/p/casebrief-midler-v-ford-motor-co.

[Neu] *Cos'è una rete neurale? — Tibco Software*. URL: https://www.tibco.com/it/reference-center/what-is-a-neural-network.

[Ope] *OpenAI*. URL: https://openai.com.

[Piz] *Teoria della cospirazione del Pizzagate*. URL: https://it.wikipedia.org/wiki/Teoria_della_cospirazione_del_Pizzagate.

[Pri] *Princess Leia Fixed using Deepfakes*. URL: https://www.youtube.com/watch?v=byKy9kGnyvo.

[Ras19] Martijn Rasser. *Why Are Deepfakes So Effective?* 2019. URL: https://blogs.scientificamerican.com/observations/why-are-deepfakes-so-effective/.

[Ref] *Reface: Funny face swap videos*. URL: https://play.google.com/store/apps/details?id=video.reface.app&hl=en&gl=US.

[Ren] *La scena muta di Matteo Renzi in inglese (un video di Alessio Marzilli)*. URL: https://www.youtube.com/watch?v=s90vrzU3zM8.

[RMC15] Alec Radford, Luke Metz, and Soumith Chintala. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. 2015. DOI: 10.48550/ARXIV.1511.06434. URL: https://arxiv.org/abs/1511.06434.

[Rou] *Roundhay Garden Scene*. 1888. URL: https://youtu.be/knD2EhjGwWI.

[Rtx] *ASUS ROG STRIX NVIDIA GeForce RTX4090 O24G GAMING : Amazon.it*. URL: https://www.amazon.it/ASUS-GeForce-RTX4090-Grafica-DisplayPort/dp/B0BHD6N2CK?source=ps-sl-shoppingads-lpcontext&ref_=fplfs&psc=1&smid=A2XWAQBXN7JC29.

[Rus10] Stuart J Russell. *Artificial intelligence: a modern approach*. Pearson Education, Inc., 2010.

[Sen] *Sensity*. URL: https://sensity.ai.

[Sha] *Shamook: Star Wars effects company ILM hires Mandalorian deepfaker*. 2021. URL: https://www.bbc.com/news/entertainment-arts-57996094.

[SSKS17] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. "Synthesizing Obama: Learning Lip Sync from Audio". In: *ACM Trans. Graph.* 36.4 (July 2017). ISSN: 0730-0301. DOI: 10.1145/3072959.3073640. URL: https://grail.cs.washington.edu/projects/AudioToObama/.

[Sta20] Nick Statt. *Jay Z tries to use copyright strikes to remove deepfaked audio of himself from YouTube*. 2020. URL: https://www.theverge.com/2020/4/28/21240488/jay-z-deepfakes-roc-nation-youtube-removed-ai-copyright-impersonation.

[Stu19] Catherine Stupp. "Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case". In: *The Wall Street Journal* (2019). URL: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

[Swa] *Swapface*. URL: https://www.swapface.org/?gclid=Cj0KCQjw2v-gBhC1ARIsAOQdKY1u2u63IR5sSYqtcpFTz-VbziUI6ThheEOwVVEeLvkqhPcF2RnBzPwcB#/home.

[Tar] *Deepfaking Tarkin & Leia in Rogue One: A Star Wars Story [4K]*. URL: https://www.youtube.com/watch?v=_CXMb_MO3aw.

[Ten] *TensorFlow*. URL: https://www.tensorflow.org/?hl=it.

[Tha+21] Vajira Thambawita et al. "Deepfake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine". In: *Scientific Reports* 11.1 (Nov. 2021). DOI: 10.1038/s41598-021-01295-2.

[Thi] *This Person Does Not Exist*. URL: https://this-person-does-not-exist.com.

[Thi+16] J. Thies et al. "Face2Face: Real-time Face Capture and Reenactment of RGB Videos". In: *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*. 2016.

[Tox] *Free to Play? Hate, Harassment and Positive Social Experience in Online Games 2020*. URL: https://www.adl.org/resources/report/free-play-hate-harassment-and-positive-social-experience-online-games-2020.

[Tut] *Tutela giuridica della voce. Diritto d'autore e voce (cantata e recitata)*. URL: https://www.guidelegali.it/approfondimenti-in-propriet-intellettuale-diritto-autore/tutela-giuridica-della-voce-diritto-d-autore-e-voce-cantata-e-recitata-9246.aspx.

[Vin21] James Vincent. *Deepfake dubs could help translate film and TV without losing an actor's original performance*. 2021. URL: https://www.theverge.com/2021/5/18/22430340/deepfake-dubs-dubbing-film-tv-flawless-startup.

[Woma] *joemag on Twitter:"This app is too powerful"*. URL: https://twitter.com/joemag_games/status/1369739853374525441?ref_src=twsrc\%5Etfw\%7Ctwcamp\%5Etweetembed\%7Ctwterm\%5E1369739853374525441%7Ctwgr\%5Ea5b1ba7c10abab9233e4e55b9fd60f69f8db5e66\%7Ctwcon\%5Es1_&ref_url=https\%3A\%2F\%2Fwww.fanbyte.com\%2Fgames\%2Fnews\%2Fvideo-game-deepfakes-are-freaking-me-out\%2F.

[Womb] *tommy on Twitter*. URL: https://www.youtube.com/watch?v=1dW_I_d6oQA&t=211s.

[Yu+22] Haiming Yu et al. "Migrating Face Swap to Mobile Devices: A lightweight Framework and A Supervised Training Solution". In: *in ICME* (2022).

[Zaoa] *Zao*. URL: https://zaodownload.com/.

[Zaob] *Zao is the most downloaded free app in China*. URL: https://zaodownload.com/zao-is-the-most-downloaded-free-app-in-china.