Addestramento di una rete neurale per la detection di 68 keypoint facciali per pazienti con disturbi neurologici affetti da disartria

Christopher Buratti and Massimiliano Piccinini

Università Politecnica delle Marche, Ingegneria Informatica e dell'Automazione, Via Brecce Bianche 12, Ancona, 60131, Italia.

Contributing authors: chris.bura00@gmail.com;
massimilianopiccinini.10@gmail.com;

Abstract

Introduzione: La disartria è un disturbo neurologico del linguaggio che causa anomalie nei movimenti vocali. Le cause principali sono malattie neurodegenerative, condizioni infiammatorie e patologie vascolari. È importante valutare l'evoluzione della disartria per prescrivere strategie di compensazione, monitorare la malattia, partecipare a studi clinici e valutare correlazioni con altri segni facciali.

Stato dell'arte: La valutazione standard della disartria si basa principalmente su valutazioni soggettive da parte di professionisti, ma presenta limitazioni nell'individuare cambiamenti sottili. Sistemi di supporto basati sull'apprendimento automatico sono stati proposti per analizzare le prestazioni vocali. Applicazioni basate sul telemonitoraggio sono state sviluppate per valutare l'intelligibilità della parola e le abilità motorie orali, ma la maggior parte di esse è stata testata solo in scenari controllati. Sono stati compiuti sforzi limitati per monitorare i muscoli oro-facciali in modo automatico.

Materiali e metodi: L'apprendimento di rappresentazioni ad alta definizione svolge un ruolo fondamentale in molti problemi di Computer Vision. L' HRNET sviluppata di recente mantiene rappresentazioni dettagliate lungo l'intero processo, collegando convoluzioni ad alta e bassa risoluzione in modo parallelo e generando rappresentazioni ad alta definizione attraverso fusioni ripetute tra le convoluzioni parallele.

Risultati e discussioni: Per determinare la bontà dell'algoritmo si è valutato l'errore di NME (Normalized Mean Error), che in fase di testing assume il valore di 0.2%.

Conclusioni: Il nostro sistema di Facial Landmark Detection può migliorare sensibilmente la soluzione telemetrica che aiuta la diagnosi di malattie.

Keywords: Paziente, SLA, Ictus, Disartria, Landmark, Detection, HRNet

1 Introduzione

La disartria rappresenta un insieme di disturbi neurologici del linguaggio che causa anomalie nella forza, nella velocità, nell'ampiezza, nella costanza, nel tono o nella precisione dei movimenti tipici degli aspetti respiratori, risonatori o articolari della produzione vocale. I disturbi neurofisiologici responsabili del controllo o dell'esecuzione, sono dovuti ad anomalie che, il più delle volte, includono debolezza, spasticità, assenza di coordinazione, movimenti involontari, etc. . Le malattie neurodegenerative (come la sclerosi laterale amiotrofica, detta SLA), le condizioni infiammatorie (come la sclerosi multipla) e le patologie vascolari (come l'ictus) sono le principali cause di insorgenza della disartria. Offrire continuità di cura ai pazienti che soffrono di disartria è importante per evitare l'insorgere di condizioni di disagio sociale che potrebbero colpire i pazienti, le loro famiglie e chi li assiste. Infatti, la compromissione della capacità di comunicazione, causata dall'insorgenza e dall'evoluzione della disartria, riduce la possibilità dell'individuo di mantenere ed estendere i contatti sociali, con un forte impatto sull'equilibrio psichico e sul benessere generale.

Detto ciò, in letteratura viene riconosciuta l'importanza di un'assidua valutazione dell'evoluzione della disartria per: identificare prontamente il momento giusto per prescrivere strategie di compensazione alla disabilità comunicativa; monitorare l'evoluzione delle malattie per gestire e programmare al meglio la cura del paziente; convocare, qualora necessario, i pazienti in studi clinici; valutare la correlazione con altri segni facciali.

Per cercare nuove misure quantitative per valutare i progressi della disartria, sono state proposte numerose metodologie convalidate clinicamente. Questi approcci riguardano principalmente il monitoraggio delle note vocali dei pazienti affetti da disartria, sia a casa che in ospedale. Tuttavia, come affermato in anche le caratteristiche dei muscoli oro-facciali dovrebbero essere prese in considerazione per: identificare sottili cambiamenti nelle prestazioni del paziente, accelerare l'implementazione di strategie correttive (ad esempio, la riabilitazione del linguaggio), valutare i progressi dei trattamenti farmacologici e non farmacologici.

Il monitoraggio della muscolatura oro-facciale si basa principalmente su ispezione visiva da parte dei medici, combinata con la stesura di scale di valutazione, come il profilo di disartria di Robertson. Questa procedura ha gli svantaggi di essere qualitativa e risultare, dunque, soggetta alla variabilità dei giudizi intra/inter-clinici (cioè attuabile solo durante la valutazione ambulatoriale). Inoltre, la valutazione è altamente influenzata dallo stato emotivo e fisico del paziente al momento della visita (ad esempio, la stanchezza causata dal viaggio verso la struttura). I risultati di queste scale vengono spesso raccolti in formato cartaceo in modo non strutturato, ostacolando la consultazione e condivisione dei dati tra i diversi centri clinici.

Una possibile soluzione, per attenuare i problemi causati dalle valutazioni sui muscoli oro-facciali impiega sensori elettromiografici posizionati sulla superficie del viso e all'interno della cavità orale, così da monitorare la disartria. Tuttavia, questo esame, oltre alla complessità del sistema di acquisizione, può essere profondamente invasivo per il paziente. Per valutare in modo oggettivo, e non invasivo, i disturbi oro-facciali nei pazienti affetti da malattie neurologiche, esiste una metodologia di deep learning (DL) per la valutazione dell'allineamento facciale, partendo da dei video RGB di pazienti affetti da SLA e ictus. È stato rilasciato, poi, il dataset "Toronto NeuroFace", che è una raccolta di fotogrammi RGB, ognuno con la propria annotazione associata dei 68 punti di riferimento facciali. Il "Toronto NeuroFace" è il primo dataset annotato in questo campo, rilasciato per supportare la comunità scientifica a proporre metodologie innovative per la valutazione dei muscoli oro-facciali nei pazienti con SLA e ictus.

A seguito del lavoro svolto e delle applicazioni di rilevamento di punti facciali basati su video in campi molto simili (ad esempio, per valutare i sintomi della depressione e la presenza di paralisi cerebrale, la presente ricerca propone un sistema di rilevazione automatica della posizione dei 68 punti di riferimento facciali, a partire da immagini RGB acquisite ovunque si voglia.

I contenuti del nostro lavoro sono riassunti qui di seguito:

- Una rete neurale convoluzionale end-to-end (CNN) per l'individuazione di punti facciali in pazienti con SLA e ictus utilizzando il dataset Toronto NeuroFace. La CNN è stata ispirata da (He, Gkioxari, Dollár e Girshick, 2017) e l'abbiamo definita maschera di landmark facciale RCNN.
- Un'indagine qualitativa delle performance del sistema testato su immagini acquisite direttamente da noi.

2 Stato dell'Arte

2.1 Valutazione della disartria

Nel campo della logopedia, il metodo standard per valutare la disartria consiste nella valutazione soggettiva da parte dei professionisti. Queste valutazioni vengono utilizzate per differenziare le condizioni cliniche, determinare la gravità della condizione e monitorare il progresso del trattamento. Queste valutazioni coinvolgono tipicamente compiti che valutano sia le capacità linguistiche (come linguaggio spontaneo, vocalizzazione e lettura) sia compiti non linguistici o motori (come prove di prassia oro-facciale e diadochocinesi).

Tuttavia, l'uso di scale cliniche in queste valutazioni presenta limitazioni. Possono raggiungere un effetto soffitto, rendendo difficile rilevare cambiamenti sottili nelle prestazioni del paziente, soprattutto nelle prime fasi della malattia. Ciò ostacola gli sforzi di ricerca, come i trial clinici, che si basano su valutazioni qualitative. Per superare queste limitazioni, i ricercatori hanno proposto vari sistemi di supporto che analizzano le prestazioni vocali utilizzando metodi di apprendimento automatico. Questi sistemi mirano a valutare la disartria mediante l'analisi dei dati audio raccolti in contesti ospedalieri e domestici. Sebbene questi sistemi siano preziosi per i clinici

nel rilevare e monitorare la disartria, spesso si concentrano su aspetti specifici della malattia legati alla parola parlata e alle prestazioni vocali.

Sono stati compiuti anche sforzi per sviluppare applicazioni basate sulla telemonitoraggio per valutare l'intelligibilità della parola e le abilità motorie orali dei pazienti disartrici. Queste applicazioni mirano a valutare automaticamente l'evoluzione della disartria e fornire ai clinici informazioni preziose. Tuttavia, la maggior parte di questi approcci è stata testata in scenari controllati e non cattura appieno le caratteristiche complete della disartria. Inoltre, molti di questi approcci si basano su metodologie di elaborazione del segnale, che potrebbero non essere sufficientemente robuste per gestire la variabilità dei dati acquisiti in ambienti domestici.

Sono stati compiuti sforzi limitati in letteratura per monitorare automaticamente i muscoli oro-facciali, che sono cruciali per caratterizzare completamente la progressione della disartria. Uno studio ha utilizzato sensori elettromiografici per monitorare i muscoli della regione oro-facciale, ma questo esame richiede personale specializzato e può essere eseguito solo in strutture sanitarie.

Un'altra prospettiva discussa in una rassegna è l'uso di applicazioni basate sulla telemedicina per i pazienti con SLA, evidenziando i benefici di tali sistemi sulla base del feedback dei pazienti. Tuttavia, la maggior parte dei sistemi presentati si concentra sulla telemedicina in tempo reale, con un uso limitato di approcci asincroni per il monitoraggio dei pazienti in contesti di riabilitazione fisica.

2.2 Rilevamento dei punti di riferimento facciali: dai metodi alle sfide del telemonitoraggio self-service

La rilevazione dei landmark facciali basata su video è un campo di ricerca con numerose applicazioni cliniche, che vanno dal riconoscimento delle espressioni umane alla rilevazione della fatica e all'identificazione della paralisi cerebrale.

È stata proposta una metodologia basata sull'apprendimento automatico (ML) per valutare la posizione dei landmark facciali al fine di rilevare la paralisi facciale. Nonostante i risultati promettenti, il Machine Learning tradizionale, rispetto al Deep Learning (DL), potrebbe (i) fallire nella generalizzazione dei video acquisiti in condizioni non ottimali, ad esempio in ambienti con illuminazione e sfondo variabili, e (ii) richiedere una fase di estrazione manuale delle caratteristiche che rende l'approccio inadatto per la pratica clinica effettiva.

Per risolvere questi problemi, esiste una metodologia a cascata basata su reti neurali convoluzionali (CNN). Il loro framework, in primo luogo, localizza approssimativamente la posizione dei landmark e poi la raffina tramite una sotto-rete di regressione. Questa metodologia presenta difficoltà nell'affrontare: (i) condizioni di scarsa illuminazione che possono verificarsi nell'ambiente domestico, (ii) parti del viso occluse che possono essere dovute alla presenza di abiti che coprono alcune porzioni del viso e (iii) posizioni difficili dei landmark, che nel caso dei pazienti neurologici possono essere dovute a gravi deficit oro-facciali. Inoltre, l'uso di due reti successive potrebbe essere inefficiente dal punto di vista del deployment su cloud e aumentare i costi.

Contributi recenti della letteratura affrontano le sfide poste da scenari reali, beneficiando di framework basati su reti neurali convoluzionali end-to-end originariamente progettati per stimare la postura delle persone. Ispirandosi a queste ipotesi di ricerca,

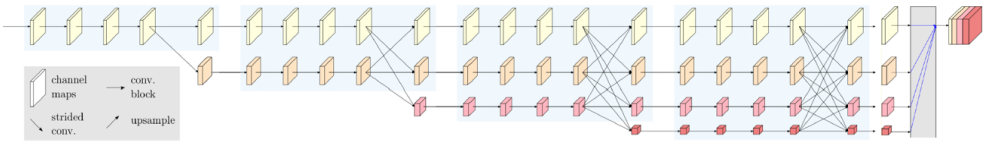

Fig. 1 Esempio di rete ad alta risoluzione. Ci sono quattro fasi. La prima è costituita da convoluzioni ad alta risoluzione. La seconda (terza, quarta) fase ripete due blocchi di risoluzioni (tre blocchi, quattro blocchi).

la nostra metodologia individua i landmark facciali nei pazienti affetti da SLA e ictus utilizzando il dataset Toronto Neuroface.

3 Materiali e metodi

3.1 Rete utilizzata

Per lo svolgimento di questo lavoro è stata adottata una rete della famiglia delle *High-Resolution Net (HRNet)*, ovvero l'*HRNetV2*. La scelta è stata supportata dal task in esame (keypoint detection); infatti, esistono due tipi di rappresentazione: quelle "low-resolution", adatte maggiormente a task di classificazione, e quelle "high-resolution", adatte, invece, a task di segmentazione, detection, stima della posa, etc.

Nello specifico, l'HRNet è stata sviluppata inizialmente per la stima della posa umana. Tale rete mantiene rappresentazioni ad alta risoluzione collegando, in parallelo, convoluzioni ad alta risoluzione a quelle a bassa risoluzione e conducendo ripetutamente fusioni multiscala attraverso convoluzioni parallele. Le risultanti rappresentazioni ad alta risoluzione non sono solo forti ma, anche, precise dal punto di vista spaziale.

A partire da questa, è stata apportata una semplice modifica considerando le rappresentazioni di tutte le convoluzioni parallele ad alta-bassa risoluzione diverse dalle sole rappresentazioni ad alta risoluzione. Questa modifica aggiunge un piccolo overhead, ma ci porta a rappresentazioni ad alta risoluzione più forti. La rete risultante è denominata HRNetV2.

In Figura 1 viene illustrata l'architettura dell'HRNet. Vengono riportate quattro fasi; la seconda, la terza e la quarta sono formate dalla ripetizione di blocchi multi-risoluzione modularizzati. Nel dettaglio, un blocco multi-risoluzione è costituito da un gruppo convoluzionale multi-risoluzione e una convoluzione multi-risoluzione (come illustrato in Figura 2 (a) e (b)). Il gruppo convoluzionale multi-risoluzione è una semplice estensione della convoluzione di gruppo, dividendo i canali di input in diversi sottoinsiemi di canali ed eseguendo una convoluzione regolare su ciascun sottoinsieme, a partire da diverse risoluzioni spaziali considerate separatamente. La convoluzione multi-risoluzione è illustrata nella Figura 2 (b). Ricorda la connessione completa multi-branch della convoluzione regolare, illustrata in Figura 2 (c). Una convoluzione regolare può essere divisa in più piccole convoluzioni. I canali di input sono divisi in diversi sottoinsiemi, così come quelli di output. Questi sottoinsiemi sono collegati in modo completamente connesso, e ciascuna connessione è una convoluzione regolare. Ogni

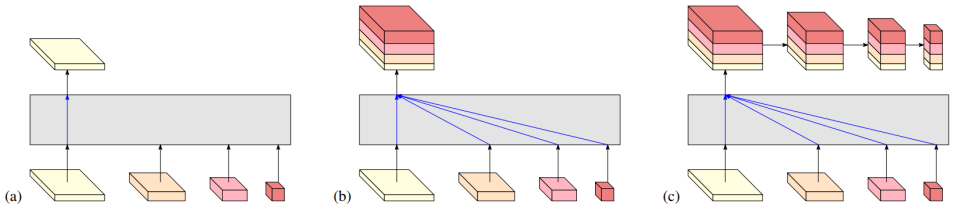

Fig. 2 (a) Rappresentazione ad alta risoluzione (HRNetV1); (b) Concatenazione delle rappresentazioni (sovracampionate) che provengono da tutte le risoluzioni per la segmentazione semantica e il rilevamento dei punti di riferimento facciali (HRNetV2); (c) Una piramide di feature formata come in (b) per il rilevamento di oggetti (HRNetV2p). Le rappresentazioni a quattro risoluzioni nella parte inferiore di ciascuna sotto-Figura vengono emesse in output dalla rete in Figura 1 e il box grigio indica come la rappresentazione dell'output è ottenuta dalle rappresentazioni a quattro risoluzioni dell'input.

sottoinsieme di canali di output è una sommatoria degli output delle convoluzioni su ciascun sottoinsieme di canali di ingresso.

Ci sono due differenze. La prima è che in una convoluzione a risoluzione multipla ogni sottoinsieme di canali è su una risoluzione diversa. La seconda è che la connessione tra i canali di ingresso e quelli di uscita deve gestire la diminuzione della risoluzione, utilizzando, dunque, diverse convoluzioni 3×3 con stride 2. L'aumento della risoluzione è semplicemente implementato mediante sovracampionamento bilineare (vicino più prossimo).

Ora, si illustra in che modo avviene l'istanziazione della rete. Questa, parte da una conFigurazione che consiste in convoluzioni con stride pari a 2, abbassando la risoluzione a 1/4 della stessa. Il primo stage contiene 4 unità residue formate da un collo di bottiglia con larghezza 64, ed è seguito da una convoluzione 3×3 che riduce la larghezza delle feature map a C. Il secondo, terzo e quarto stage contengono rispettivamente 1, 4 e 3 blocchi multi-risoluzione. Le larghezze (numero di canali) delle convoluzioni delle quattro risoluzioni sono rispettivamente C, 2C, 4C e 8C. Ogni ramo del gruppo convoluzionale multirisoluzione contiene 4 unità residue, ed ognuna contiene due convoluzioni 3×3 in ogni risoluzione. Si mischiano le rappresentazioni di output (Figura 2 (b)), da tutte e quattro le risoluzioni attraverso una convoluzione 1×1, producendo una rappresentazione con dimensione 15C. Poi, si passa la rappresentazione ottenuta, in ogni posizione a un classificatore/regressore lineare con la softmax/MSE loss per prevedere heatmap dei punti di riferimento facciali.

3.2 Imagenet Dataset

Per addestrare il nostro modello, si è scelto di utilizzare la tecnica del fine-tuning. A questo scopo, si è scelta una rete preaddestrata su un sottoinsieme del dataset *ImageNet*. Nel dettaglio, ImageNet è un dataset di immagini organizzato secondo la gerarchia di WordNet; ogni concetto significativo in WordNet, descritto da parole o frasi, è chiamato "synset". Ci sono più di 100.000 synset in WordNet. In ImageNet, vengono fornite in media 1000 immagini per illustrare ogni synset. Le immagini di ogni

concetto sono di qualità e annotate da umani. Per il nostro task, di rilevante interesse è la parte di dataset rafFigurante volti di persone.

3.3 Toronto NeuroFace Dataset

Il fine-tuning è stato condotto sul *Toronto NeuroFace* dataset, ovvero una raccolta di video RGB di pazienti affetti da SLA (11 soggetti: 4 maschi, 7 femmine), da ictus (14 soggetti: 10 maschi, 4 femmine) e altri sani (11 soggetti: 7 maschi, 4 femmine) di pari età.

I video sono stati registrati in un ambiente controllato con condizioni ottimali di luce, tramite la telecamera Intel RealSense piazzata a 30-60 cm dal viso del soggetto. Ad ogni soggetto è stato richiesto di eseguire alcuni specifici task vocali e motori, come ad esempio: mantenere la massima apertura della bocca e muovendo le labbra per ripetere le sillabe /pa-ta-ka/ o /pa/.

I fotogrammi sono stati estratti da ciascun video per massimizzare la variabilità intra-soggetto, ad esempio, per attività motorie ripetitive sono stati considerati 3 fotogrammi per ripetizione: l'inizio del gesto; il suo apice; il punto medio tra i precedenti. Dopo l'estrazione dei fotogrammi, è stata eseguita, manualmente, l'annotazione dei 68 punti di riferimento facciali e del bounding box del volto per 3306 fotogrammi dei quali: 1015 sono dei soggetti sani, 920 dei pazienti con la SLA e 1371 dei pazienti con l'ictus.

Per lo scopo del nostro lavoro, il dataset è stato suddiviso in una parte di testing, una di validation e una di training, prendendo 32 soggetti per il training e la validazione, e 4 soggetti (2 maschi e 2 femmine; 2 affetti da SLA e 2 da ictus) per il testing.

3.4 Impostazioni di training

Il modello utilizzato è stato preaddestrato, come precedentemente detto, su una porzione del dataset ImageNet. I pesi della rete risultanti sono stati utilizzati come punto di partenza per effettuare il fine tuning sul dataset Toronto Neuroface.

Il learning rate iniziale è stato impostato 0.0001 e viene diminuito di un fattore 10 alla trentesima epoca e di un ulteriore fattore 10 alla cinquantesima epoca. I modelli sono stati addestrati per 60 epoche con un batch size pari a 16 su una GPU (GeForce RTX 2080Ti). Inoltre, è stata effettuata, anche, la data augmentation, ruotando le immagini sul piano, di $\pm$ 30 gradi, impostando dei valori di scala compresi tra 0.75 e 1.25 e capovolgendole casualmente. La rete adottata è la HRNetV2-W18 per la detection dei punti facciali, la quale ha 9.3M parametri e costo computazionale (GFLOPs) pari a 4.3G (entrambi i parametri risultano essere minori ad altre due reti di confronto, quali: ResNet-50, con numero di parametri = 25.0M e GFLOPs = 3.8G; Hourglass, con numero di parametri = 25.1M, GFLOPs = 19.1G).

3.5 Metriche di valutazione

Per valutare le performance del modello testato, l'errore medio normalizzato (NME_k) è stato calcolato su tutte le immagini del test come segue:

$$NME_k = \left(\frac{1}{N_L} \sum_{i=1}^{N_L} \frac{\sqrt{(x_i - xp_i)^2 + (y_i - yp_i)^2}}{Diag_{bbox}} \right) \cdot 100 \quad (1)$$

dove $Diag_{bbox}$ è la lunghezza della diagonale del bounding box dell'immagine in esame, N_L rappresenta il numero di landmark della parte della faccia in esame (k di NME_k), (x_i, y_i) sono le coordinate di ground-truth del landmark in esame e (xp_i, yp_i) sono le coordinate predette in riferimento allo stesso landmark. L'NME è stato calcolato per la totalità dei 68 landmark facciali (NME_{68}), per i 17 landmark del mento (NME_{chin}), per i 10 landmark delle sopracciglia ($NME_{eyebrows}$), per i 9 landmark del naso (NME_{nose}), per i 12 landmark degli occhi (NME_{eyes}) e per i 20 landmark della bocca (NME_{mouth}).

4 Risultati e discussioni

Nella Tabella 1 vengono riportati i risultati, delle metriche adottate, comparandoli con altre CNN.

N.B. I valori riportati sulla colonna "Nostra CNN" sono i valori degli NME di test.

Table 1 Risultati in termini di NME (%) del nostro lavoro comparato con altre CNN.

	Nostra CNN	RCNN	N-FLMask	300VW-FLMask	N-Mask
NME_{68}	0.20	1.79	2.70	3.88	13.55
NME_{chin}	0.61	2.62	4.81	4.39	15.31
$NME_{eyebrows}$	0.53	0.02	0.03	0.05	0.12
NME_{nose}	0.35	1.55	2.08	3.60	5.61
NME_{eyes}	0.23	1.03	0.94	3.04	5.23
NME_{mouth}	0.26	1.49	2.19	3.70	21.17

Nelle figure 3, 4, 5 e 6, vengono illustrate 4 immagini dal set di testing con, in verde, i 68 landmark di ground truth, e, in rosso, i 68 punti predetti dal nostro modello.

Nelle figure 7, 8, 9, 10, 11, 12 e 13, vengono illustrati i grafici dell'andamento delle rispettive metriche in fase di training e di validation.

5 Conclusioni

5.1 Carbon Footprint

Ai fini del lavoro, è stata tenuta traccia, anche, del carbon footprint, ovvero sono state stimate le emissioni in atmosfera di gas serra causate dal lavoro svolto dalla gpu in fase di training, testing e di utilizzo finale. I risultati sono riportati in Tabella 2. In particolare, le emissioni di training fanno riferimento alle 60 epoche di training, appunto, e quelle di utilizzo finale fanno riferimento all'utilizzo del nostro modello su due immagini. Se volessimo rappresentare le emissioni totali causate dal nostro lavoro passato e dal suo eventuale utilizzo futuro con una formula matematica, questa è osservabile nella formula 2. Nota: la N che compare in tale formula rappresenta il

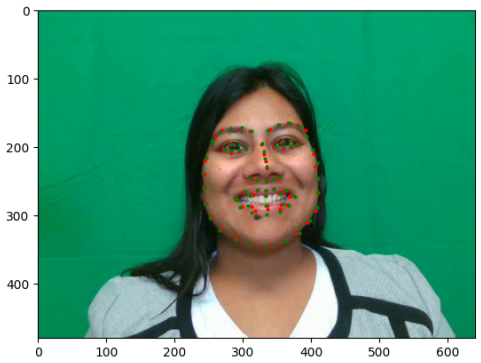

Fig. 3 Paziente 1

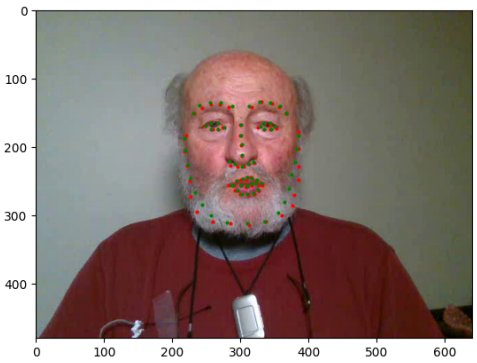

Fig. 4 Paziente 2

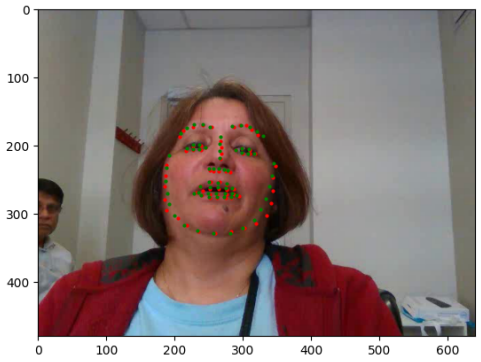

Fig. 5 Paziente 3

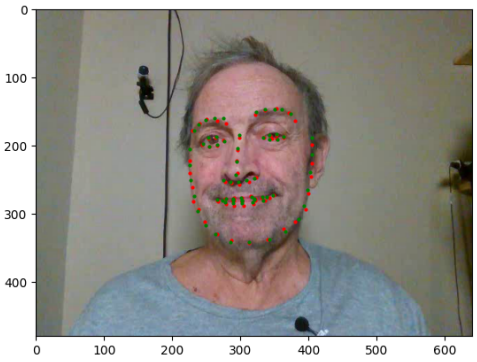

Fig. 6 Paziente 4

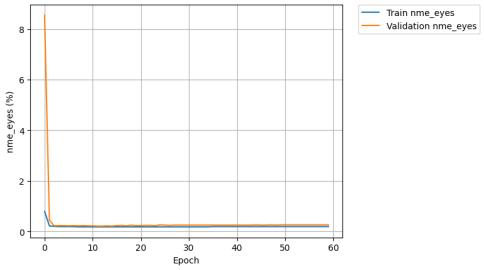

Fig. 7 NME occhi

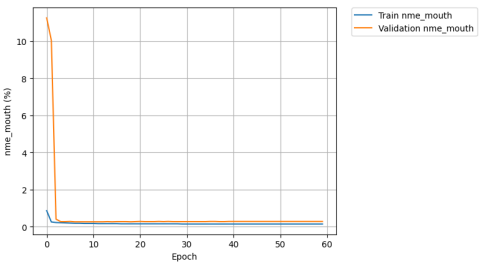

Fig. 8 NME bocca

numero di immagini di cui si vogliono far predire i 68 landmark facciali dal nostro modello.

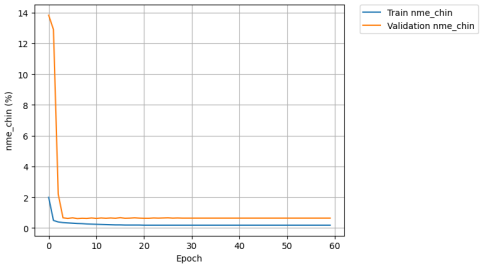

Fig. 9 NME mento

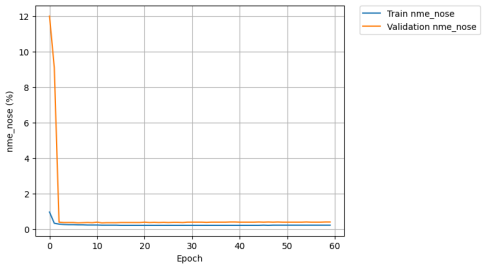

Fig. 10 NME naso

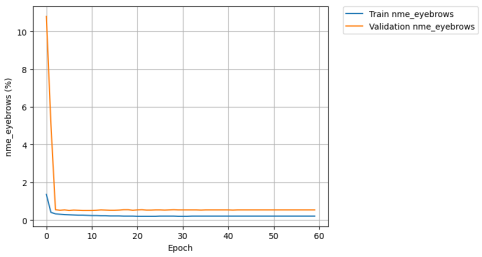

Fig. 11 NME sopracciglia

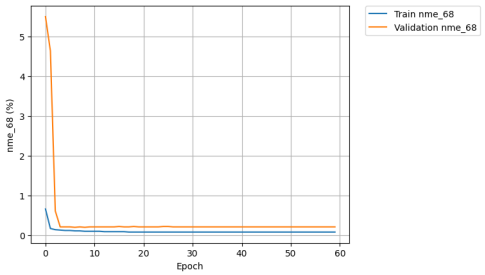

Fig. 12 NME 68 landmark totali

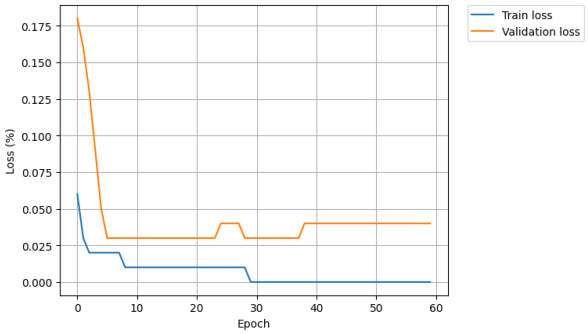

Fig. 13 Loss

$$Emissioni_{tot} = (0.051138648458796336 + 0.0002191036015030425) + \frac{2.1715678174173568e-05}{2} \cdot N = \\ = (0.05135775206 + 0.00001085783 \cdot N)[kg] \tag{2}$$

Table 2 Emissioni di gas serra causate dal nostro lavoro.

	Emissioni (kg)
Training	0.051138648458796336
Testing	0.0002191036015030425
Utilizzo finale	2.1715678174173568e-05

5.2 Avvio del Progetto

Per avviare il progetto GitHub, scarica il repository dal link fornito utilizzando il comando:

```
$ git clone https://github.com/christopherburatti/CV-DeepLearning
```

Questo importerà l'intero progetto nella directory locale. Successivamente, bisogna scaricare il file model_best_FINAL_3.pth dal link. Una volta ottenuto il file model_best_FINAL_3.pth, va posizionato nella directory di default del progetto. Ci si assicuri che il file sia correttamente posizionato affinché il programma possa accedervi correttamente. Per installare le librerie, si usi il comando:

```
$ pip install -r requirements.txt
```

Una volta completato, si può procedere con l'esecuzione delle funzioni principali del progetto.

Per avviare il processo di training, utilizzare il comando:

```
$ python tools/train.py --cfg hrnetv2_w18_imagenet_pretrained.pth
```

Questo inizierà l'addestramento del modello utilizzando i dati disponibili.

Per effettuare il test, eseguire il comando:

```
$ python tools/test.py --cfg hrnetv2_w18_imagenet_pretrained.pth
--model model_best_FINAL_3.pth
```

Questo valuterà le prestazioni del modello utilizzando un set di dati di test.

Infine, se si desidera eseguire una demo, si utilizzi il comando:

```
$ python tools/demo.py --cfg hrnetv2_w18_imagenet_pretrained.pth
--model model_best_FINAL_3.pth
```

Questo avvierà la demo del progetto, in cui abbiamo inserito due nostre immagini di prova, sulle quali verranno plottati i facial landmarks predetti.

In Figura 14 riportiamo il QR code per accedere direttamente al nostro repository github.

5.3 Demo

Ecco un esempio di descrizione formale in LaTeX per la demo del progetto che restituisce i 68 landmark facciali date due foto:

1. Per avviare la demo del progetto, eseguire il seguente comando:

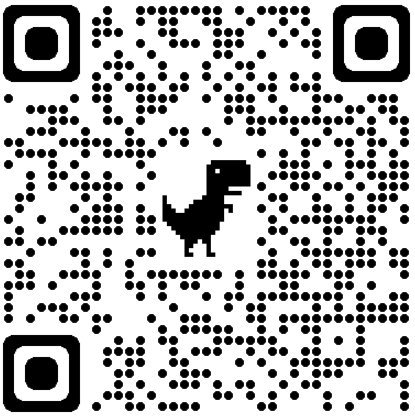

Fig. 14 QR code per il repository github contenente il codice. In alternativa, il rispettivo link è: https://github.com/christopherburatti/CV-DeepLearning

```
$ python tools/demo.py --cfg hrnetv2_w18_imagenet_pretrained.pth
--model model_best_FINAL_3.pth
```

2. Una volta eseguito il comando, verrà aperta una nuova finestra contenente le foto da analizzare.
3. Le due foto saranno visualizzate una alla volta. Ogni foto mostrerà il volto di una persona.
4. Utilizzando un modello di machine learning preaddestrato, i 68 landmark facciali verranno predetti per ciascun volto.
5. I punti predetti saranno evidenziati in rosso sulla foto, indicando la posizione approssimativa dei landmark facciali.
6. Sarà possibile visualizzare i punti predetti per entrambe le foto e osservare le differenze tra i volti analizzati.

Questi passaggi consentono di eseguire una demo che mostra i punti predetti dei 68 landmark facciali per due foto, offrendo una rappresentazione visuale dei risultati ottenuti.

5.4 Ambienti di Sviluppo

Nel progetto sono state utilizzate le seguenti versioni: Python 3.8.5 e PyTorch 1.9.0.

Per la fase di training, è stato impiegato un server. Il server era equipaggiato con il sistema operativo Linux Ubuntu 20.04.3 LTS (Focal Fossa). Tali specifiche hanno consentito di eseguire il training del modello in modo efficiente, sfruttando le capacità computazionali della GPU per accelerare il processo di apprendimento. La GPU usata è GeForce RTX 2080Ti.

Fig. 15 Predizione 1

Fig. 16 Predizione 2

5.5 Sviluppi futuri

Al termine del nostro lavoro, sono stati considerati degli eventuali sviluppi futuri per migliorare e/o ampliare l'utilizzabilità del nostro operato.

Per primo, è stato pensato di accorpare tale modello all'interno del progetto "The Homely Care cloud-based store-and-forward telemonitoring system" sviluppato da Lucia Migliorellia, Daniele Berardini, Kevin Cela, Michela Coccia, Laura Villani, Emanuele Frontoni e Sara Moccia.

Per secondo, è stato pensato di poter affiancare un software di detection facciale per acquisire le coordinate (x, y) del centro del viso dall'immagine proposta, così da aumentare l'utilizzabilità del prodotto finale ad un utente qualsiasi.

References