Relazione Visione Artificiale

Gruppo 11: De Simone Giuseppe, Giaquinto Gennaro, Marino Christian, Serritiello Simone

{g.desimone39, g.giaquinto9, c.marino34, [s.serritiello3}@studenti.unisa.it](mailto:s.serritiello3%7d@studenti.unisa.it)

# Introduzione

Mentre il riconoscimento di molte variazioni facciali come identità, genere, espressione, sono state ampiamente studiate, la tematica dell’age estimation è attualmente ancora poco approfondita, data la difficoltà intrinseca nel trovare un’associazione tra caratteristiche facciali e l’età del soggetto. Infatti, in contrasto alle altre, le variazioni dovute all’età sono per loro natura un task molto più arduo da far apprendere ad un sistema di machine learning. Inoltre, la difficoltà per un sistema a discriminare i differenti valori in un certo range di età non è la stessa per ogni intervallo, bensì è facilmente intuibile e verificabile come per alcuni range sia più evidente la variazione del tempo, rispetto che per altri.

# Descrizione della soluzione

L’approccio utilizzato per la risoluzione del problema in questione non prevede il design, l’implementazione e il training della rete “*from scratch*”, dato che tale approccio richiederebbe uno sforzo maggiore in quanto le reti che si sono dimostrate efficaci per questo tipo di problema sono caratterizzate da un’elevata profondità, quindi portando la difficoltà intrinseca sia nella progettazione, che nel loro addestramento. Per questo motivo si è scelto di adottare un metodo basato o su “*fine tuning*” o su “*transfer learning*”, per semplificare i processi di design della rete e di training. Utilizzando il transfer learning, si vanno a fissare i pesi di una rete già pre-addestrata, andando a modificare solo gli ultimi livelli aggiuntivi, che andranno a adempiere allo specifico task in questione. Il fine tuning, invece, prevede l’ottimizzazione della rete pre-addestrata, fissando solo parte dei pesi di essa, e specializzando gli ultimi livelli per adempiere al task richiesto, al più modificando l’ultimo livello. Nel nostro caso si è deciso di esaminare diversi approcci al problema, sia basandoci su fine tuning, sia su transfer learning e con diverse architetture di rete, che, fissata una backbone di feature extraction (VGG-Face), assolvono al compito di classificare tramite diversi approcci, basati sia su singolo classificatore che su un sistema a multi-classificatore sulle diverse fasce d’età. In questo paragrafo sarà presentata la soluzione che ha portato ad avere performance migliori tra tutte e, nel paragrafo 3, saranno argomentate le motivazioni che hanno portato a tale scelta.

## Convolutional Neural Network

Utilizzare una CNN pre-addestrata sicuramente è vantaggioso in merito a tempistiche e risorse necessarie per il training, per questo motivo, la scelta comune, per il problema dell’age estimation, ricade nella VGG-Face2, modello proposto da [1] e che è specializzata nell’estrarre feature da immagini bidimensionali raffiguranti volti a bassa risoluzione. La VGG-Face2 consiste di undici livelli, otto convoluzionali e tre fully-connected, come riporta la figura 1; per ogni blocco convoluzionale, è presente un rectification layer, che applica l’operazione di average pooling. La scelta di questa rete risiede nel fatto che essa raggiunge i risultati dello stato dell’arte e che è stata addestrata su un dataset contenente 3.31 milioni di immagini di 9131 soggetti, con una media di 362,6 immagini per soggetto, le quali hanno una larga variazione in posa, età, illuminazione ed etnia.

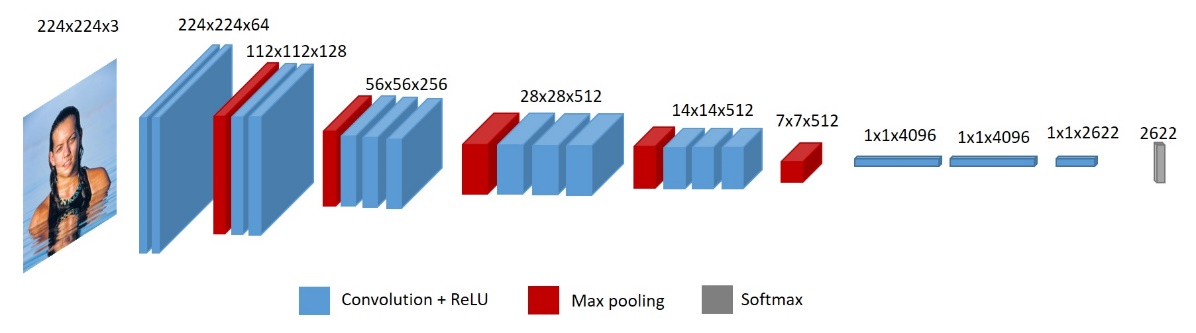


Figura : VGG-Face2 Architecture

Fissata la scelta di utilizzare la VGG-Face2, basata su ResNet, come rete di backbone per la feature extraction, quello che è stato fatto è aggiungere ulteriori due livelli densi addestrabili, il primo da 1024 neuroni, mentre l’ultimo da 101 neuroni, infatti, si è scelto di utilizzare un approccio basato su classificazione. L’ultimo livello ha come funzione di attivazione la “*softmax*” dato il problema a multi-classificazione con classi mutualmente esclusive, mentre la funzione di costo utilizzata è la “*crossentropy*”, che è largamente impiegata per il task multi-classification. In keras viene utilizzata la “*categorical crossentropy*” perché le label seguono una codifica one-hot.

## Procedure di addestramento

### Dataset

Il dataset fornito per la fase di training contiene circa 3 milioni di immagini relative a volti di persone con differenti pose, illuminazioni, ecc. La scelta delle dimensioni del validation set è stata vincolata dalle risorse messe a disposizione dal tool per l’esecuzione del training. Ci si è posta la scelta di aumentare la dimensione del validation set proporzionalmente al training set, a discapito della dimensione dell’intero dataset e quindi prevedendo un training meno ricco, per cui la suddivisione in training e validation è stata effettuata con una percentuale 95%-5%.

L’approccio per valutare le performance durante il training non prevede tecniche di cross-validation, dato l’elevato numero di campioni a disposizione, bensì semplice validazione.

Al fine di effettuare i primi train rapidamente, il dataset è stato suddiviso in pacchetti contenenti immagini raffiguranti soggetti nello stesso range di età, e, data la disparità nella cardinalità dei diversi range, è stato selezionato un sottoinsieme che contenesse quasi lo stesso numero di campioni per ogni categoria. Infine, il training è proseguito in maniera incrementale fino allo scadere del tempo massimo messo a disposizione, con pacchetti di training set da 130.000 campioni.

### Face detection

L’estrazione del volto dall’immagine è stata effettuata in fase di di pre-processing tramite la rete MTCNN presentata da [2]. Lo studio effettuato su tale rete fa emergere il fatto che essa presenta performance superiori allo stato dell’arte soprattutto per quanto riguarda volti con risoluzione bassa e ripresi da diverse angolazioni. La MTCNN consiste in tre stage di livelli convoluzionali specializzati nella rilevazione di facce e di facial landmarks (occhi, naso, bocca, …).

### Face pre-processing

Data la natura variegata delle pose e dell’orientamento dei volti presenti nel dataset, per facilitare un processo di training e massimizzare l’efficacia della rete che si occupa dell’attività di feature extraction, è stato impiegato un tool presente nel framework [3] per il task *Pose Normalization*, ovvero l’allineamento del viso. Basandosi sulla posizione dei facial landmarks, il tool fornisce in output una immagine della dimensione desiderata contenente il volto perfettamente allineato. La scelta di un *FaceAligner* basato sull’estrazione di facial landmarks è motivata dal fatto che essi presentano performance migliori rispetto ai metodi basati sul *best score*, ovvero sul calcolo dello score del volto ruotato a diversi angoli, com’è dimostrato in [4].

Inoltre, è stato effettuato un preprocessing dell’immagine, che prevede una normalizzazione dei colori, coerentemente con l’input richiesto dalla rete backbone in questione.



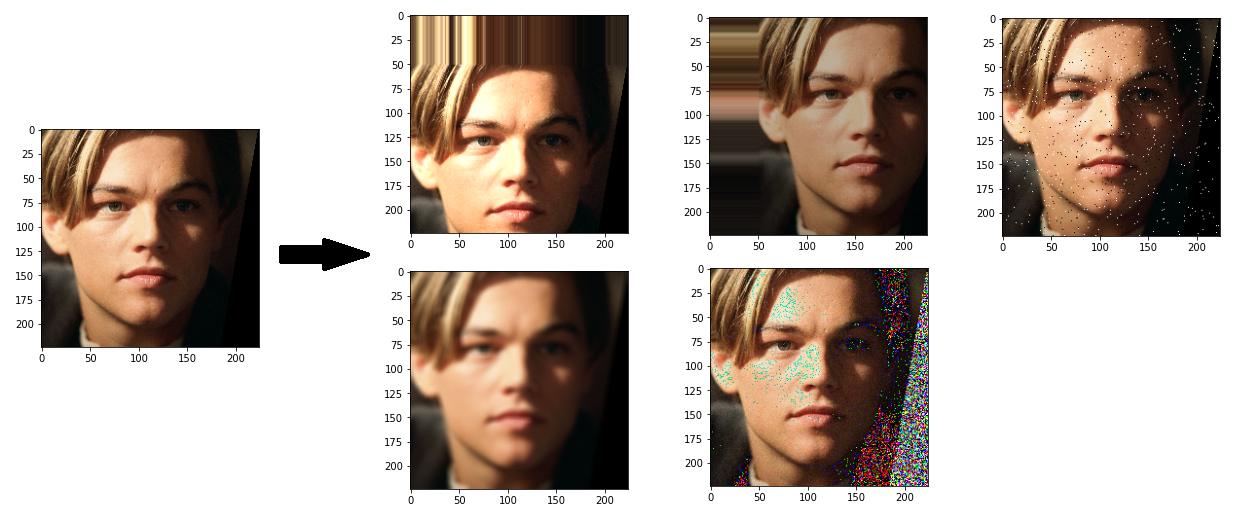
Figura 2: Volto di Leonardo Dicaprio ritagliato ed allineato

### Data augmentation

Per gestire lo sbilanciamento delle classi e per garantire una opportuna cardinalità dei diversi range di età, è stato effettuato una data augmentation in senso lato per le classi che disponevano di pochi campioni (es. range di età agli estremi – anziani o bambini), invece per le altre è stata effettuata una augmentation strettamente legata alla rappresentatività dei volti.

In particolare, è stato aumentato il dataset con immagini propriamente elaborate con tool per la Image Data Augmentation; più nello specifico, si è provveduto ad aggiungere alle immagini rumore gaussiano, sale e pepe o effetto blur. Per aumentare invece la rappresentatività delle immagini, infine, è stata applicata randomicamente a tutte le fasce d’età una alterazione che riguardasse le occlusioni, in particolare è stato effettuato uno scorrimento orizzontale o verticale delle immagini, alterazioni di tipo blur e di luminosità.

Nell’immagine sono visibili le alterazioni effettuate, ovvero: random shift verticale con sovraesposizione della luminosità, random shift orizzontale con sottoesposizione, rumore sale e pepe, random blur e rumore gaussiano.



### Transfer learning

Com’è stato anticipato, la rete è stata costruita con la tecnica del transfer learning, utilizzando tutti i pesi pre-addestrati della rete VGG-Face, dato che è meno costoso in termini di risorse computazionali e tempo necessari all’addestramento. Per cui, sono stati fissati tutti i livelli della VGG-Face e preso l’output della rete. Ad esso è stata concatenata una MLP da due livelli, uno con 1024 neuroni e uno con numero di neuroni uguale alla dimensione del range di età per il problema in questione (101 neuroni).

### Procedura di training

Per la procedura di training è stato utilizzato l’ottimizzatore “*adam*” (Adaptive Momentum Estimation), che si basa sul metodo del gradiente stocastico effettuando una stima adattiva del *momentum* del primo e secondo ordine nella discesa del gradiente.

Per il fit del modello è stata utilizzata la tecnica dello scheduling del learning rate, in particolare è stata utilizzata l’API di keras ReduceLROnPlateau, che gestisce l’addestramento del modello andando a ridurre il learning rate quando la metrica monitorata non migliora per un numero di epoche specificate (patience number). In particolare, è stato impostato un fattore di decremento pari al 0.1, avendo un learning rate iniziale pari a 0.001 e patience number pari a 2 epoche.

Infine, è stato impostato un batch\_size da 60 campioni e 5 epoche di addestramento, con un numero di steps\_for\_epoch pari a . Infine, il training è stato sviluppato in maniera incrementale.

# Risultati sperimentali

Si riportano dunque tutti gli esperimenti effettuati e le reti addestrate nel corso dello studio:

1. CNN presentata in questo documento, addestrata su circa 800.000 immagini, con data augmentation per range di età;
2. CNN simile alla (1) ma con un ulteriore livello dropout (dropout pari a 0.2), addestrata allo stesso modo, ma con meno immagini;
3. Modello multi-classificatore per 5 range di età, addestrato con data augmentation su ogni fascia di età selezionata + 5 piccoli regressori che, data una fascia di età, restituiscono un numero (età) all’interno di essa; multi-classificatore addestrato su 42.500 immagini aumentate, regressori per fascia di età addestrati singolarmente con 40.000 immagini ciascuno;
4. Modello multi-classificatore per 5 range di età sulla base della (3), utilizzando la tecnica del *fine-tuning* su tutti i livelli della rete VGG-Face2 e addestrato su 40.000 immagini aumentate. È stato utilizzato come ottimizzatore il gradiente stocastico, con learning rate=0.01 e momentum=0.9. Rete con tre livelli densi con dimensione pari a 4096 ciascuno e dropout 0.5, basandoci su [5] e [6]. (Inoltre, è stato effettuato un ulteriore test con learning rate=0.1 e questo ha causato una divergenza del gradiente);
5. CNN simile alla (1), utilizzando la tecnica del *fine-tuning* su tutti i livelli della rete VGG-Face2, ma con ottimizzatore SGD, learning rate=0.01, momentum=0.9 e tre livelli densi con dropout 0.5. Addestrata con 150.000 campioni.

I risultati riportati su MSE e MAE sono estratti da una fase finale di valutazione dei singoli modelli, in modo tale da definire esattamente le performance di ciascuno di essi. Per la validazione finale sono stati utilizzati 16.000 campioni prelevati da una parte del training set non utilizzata in precedenza per l’addestramento di nessuno dei modelli. Le reti 1 e 2 differiscono soltanto per il livello dropout aggiuntivo nel secondo modello, ma dai risultati sperimentali si può evincere che, nonostante siano stati utilizzati più campioni di training per una delle due reti, l’indice MAE riporta un errore quasi simile per entrambe. Questo è anche dovuto al fatto che è stata utilizzata una rete backbone pre-addestrata, per cui, per i livelli finali aggiuntivi, non è necessario un training molto ampio per apprendere informazione. Mentre, per quanto riguarda le reti (4) e (5), che sono state addestrate con una minore data augmentation rispetto alle altre 3 reti, è possibile evincere che esse hanno appreso effettivamente meno informazione, in quanto le immagini erano meno rappresentative delle variazioni ambientali e i vari rumori che possono succedere in un ambiente non controllato.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rete | Tipologia | # train samples | Learning rate | Validation MSE | MAE |
| 1 | Classificatore | 800.000 | 0.001 + (scheduling) | 48,14 | 5,64 |
| 2 | Classificatore | 136.000 | 0.001 + (scheduling) | 55,83 | 6,03 |
| 3 | Classif. + Regress. | 242.500 | 0.001 + (scheduling) | 121,00 | 8,25 |
| 4 | Classif. + Regress. | 240.000 | 0.01 + (scheduling) | 161,00 | 9,50 |
| 5 | Classificatore | 150.000 | 0.01 + (scheduling) | 133,00 | 9,20 |

# Conclusioni

Risultati alla mano, ci sentiamo di concludere che la tecnica del transfer learning - o grossomodo del fine-tuning - permette di ottenere risultati simili con ambedue le architetture proposte, ovvero quella basata su regressore e quella su classificatore. Dato l’utilizzo di una rete backbone molto ben addestrata sul problema del rilevamento delle feature dei volti (che è la VGG-Face), è emerso che le soluzioni più semplice e canoniche per affrontare il problema, si sono rivelate anche le più performanti e meno onerose in termini di risorse necessarie al loro addestramento. Tuttavia, siamo consapevoli del fatto che il dataset riguardante i volti e le relative età presentava un evidente sbilanciamento verso taluni range di età, non garantendo, quindi, un’adeguata rappresentatività dei range di età agli estremi, rendendo sempre più necessarie tecniche di data augmentation.

# Riferimenti

|  |  |
| --- | --- |
| [1] | A. V. a. A. Z. O. M. Parkhi, Deep face recognition, British Machine Vision Conference, 2015. |
| [2] | Z. Z. Z. L. S. M. I. a. Y. Q. S. M. I. Kaipeng Zhang, «Joint Face Detection and Alignment Using Multitask,» *IEEE,* 2016. |
| [3] | jrosebr1. [Online]. Available: https://github.com/jrosebr1/imutils. |
| [4] | A. G. ,. G. P. ,. a. M. V. Vincenzo Carletti, «Age from Faces in the Deep Learning Revolution,» *IEEE,* 2020. |
| [5] | «Wolfram,» 2018. [Online]. Available: https://resources.wolframcloud.com/NeuralNetRepository/resources/Age-Estimation-VGG-16-Trained-on-IMDB-WIKI-and-Looking-at-People-Data. |
| [6] | A. A. M. B. D. B. Zakariya Qawaqneh, «Deep Convolutional Neural Network for Age Estimation based on VGG-Face Model,» 2017. |