UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento d'Informatica

Corso di Fondamenti di Visione Artificiale e Biometria

PROGETTO

Gender Classification using a Spider Web Method

Anno Accademico 2019/2020

Fabio Grauso
Matr. 0522500811

Sommario

Introduzione al problema

La classificazione del genere umano ha molte applicazioni nel campo della biometria. Il genere infatti è una biometria molto utilizzata, specialmente per dare supporto ad altre biometrie di riconoscimento. Se ad esempio, in ambito forense, si riesce ad identificare il genere di un soggetto dai suoi tratti biometrici, si restringe il campo di ricerca di circa del 50% [1]

Determinare il genere di un volto non chiaro, è un compito impegnativo per i computer a causa della diversità e delle variazioni del volto umano. Sebbene questo compito sia considerato banale per il sistema visivo umano, ci sono ancora possibili errori di classificazione in assenza di segnali esterni (non specifici della regione del viso) come capelli, cosmetici e texture. Burton et al. [2] mostrano che gli esseri umani possono svolgere questa attività con una precisione del 96% in assenza di segnali esterni mentre Bruce et al. [3] sostengono che in media gli umani possono prendere decisioni sul genere dei volti in 613-620 millisecondi. Gli attuali algoritmi di classificazione del genere facciale basati su computer sono molto indietro rispetto all'accuratezza della percezione visiva umana.

Quando analizziamo il genere dal volto stiamo operando con una biometria fisica. La struttura ossea di individui geneticamente maschili o femminili risulta essere differente. Ed è proprio su questo concetto che si basa il seguente progetto, ovvero stimare il genere umano andando principalmente ad analizzare la struttura ossea del soggetto, permettendo così di trascendere da implicazioni di carattere etico legate alla gender recognition.

Modelli utilizzati

Face Detection

Il viso è uno dei tratti che tendono ad essere alterati dall'illuminazione, dalla posa e dall'espressione. Le occlusioni di sciarpe o occhiali da sole e l'auto-occlusione di una parte del viso a causa della posa della testa non frontale sono un'ulteriore fonte di problemi [4]. Queste condizioni possono complicare il riconoscimento del volto fino a farlo fallire.

Quindi anziché utilizzare un rilevatore di volti basato su HOG + SVM, che è passibile a questi tipi di errori, si è deciso di utilizzare in questo progetto il rilevatore MMOD [5] basato su ANN, meno conosciuto ma che offre notevoli performance.

Sebbene questo rilevatore, basato su una rete neurale convoluzionale, è pesante dal punto di vista computazionale esso offre ottimi risultati per il riconoscimento di volti non frontali e in diverse angolazioni come mostrato nelle Figura 2 e Figura 1

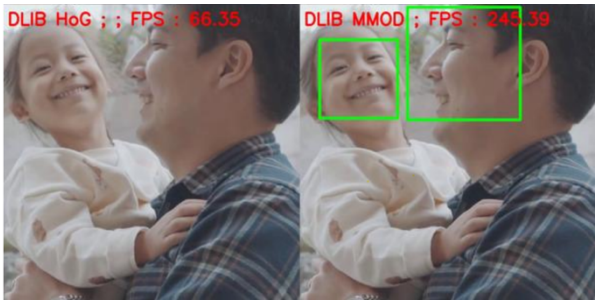


***Figura 2** - HoG vs MMOD, dlib face detection*

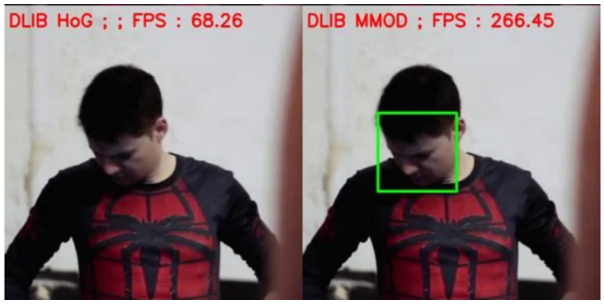


***Figura 1** - HoG vs MMOD, dlib face detection*

Si è adottato questa strategia in quanto per lo svolgimento di questo progetto è stato possibile utilizzare una GPU di ottime prestazioni, e come si evince nel grafico in Figura 3 il rilevatore MMOD, basato su una ANN è molto più veloce su una GPU ed è molto lento su una CPU.

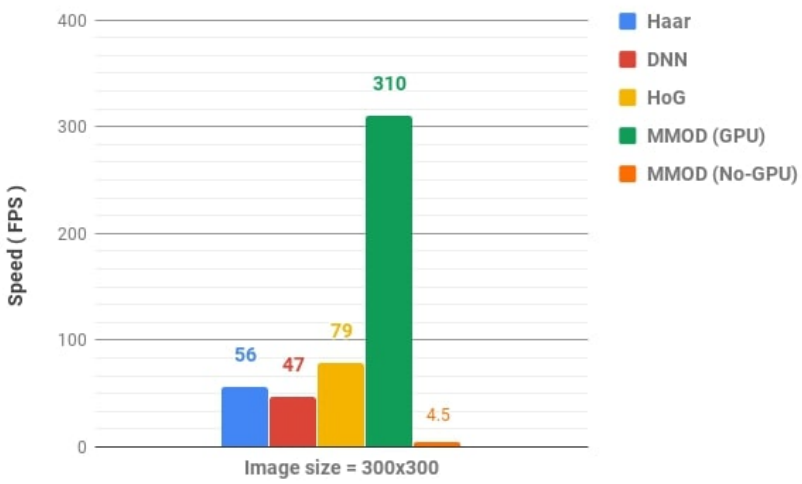


***Figura 3** - Velocità dei metodi di face detection*

Landmark Prediction

Il modello utilizzato per la previsione dei *facial landmarks* è descritto in modo dettagliato in [6]. Sostanzialmente il predittore prende come input un'immagine del volto ed emette le posizioni di 68 *facial landmarks* P_i espresse come coppie di coordinate cartesiane (x_i, y_i) con $i = 1,2, \dots 68$. Le coordinate corrispondono alle posizioni dei pixel dei *landmarks*. Il rilevamento dei punti si basa su un insieme di alberi di regressione. L'addestramento del modello sfrutta una serie di facce che vengono annotate manualmente con le coordinate x e y dei *landmarks* e con la probabilità di distanze tra singole coppie di punti. Il modello ottenuto prevede le posizioni dei punti appartenenti alle seguenti regioni salienti:

- la mascella del viso (i primi 17 punti)
- le sopracciglia (dal 18 ° al 27 ° punto)
- il naso (dal 28 ° al 36 ° punto)
- gli occhi (dal 37 ° al 48 ° punto)
- la bocca (dal 49 ° al 68 ° punto)

Da un lato, la presente proposta non comporta di nuovo il *training* del predittore. Il modello disponibile è stato scelto grazie alla sua robustezza e funziona qualunque sia il set di dati utilizzato per i test o la configurazione del modello. Infine in Figura 4 è possibile osservare i risultati del modello descritto in questo paragrafo che confermano un'alta qualità di previsioni.

Figura 4 *- Risultati ottenuti dal dataset HELEN*

Spider Web Model

Lo Spider Web Model rappresenta una nuova proposta di modello a forma di ragnatela frutto di uno studio scientifico [4] del Dipartimento d'Informatica dell'Università degli Studi di Salerno. Questo algoritmo ha il compito di determinare le posizioni relative al modello specifico dei *landmark* identificati sulla faccia; queste posizioni sono stabilite da un numero di cerchi concentrici e dai loro settori; la procedura assegna ogni *landamark* a un settore specifico e utilizza le informazioni complessive ottenute per costruire un vettore caratteristica.

Il centro e il raggio del modello spider web sono determinati in base ai punti di riferimento (*landmark)* identificati sulla faccia. Il modello a forma di ragnatela è centrato sulla punta del naso (punto numero 33 dal modello precedente) e dimensionato in base alle misure del viso, rendendo quindi indipendente dalla dimensione dell'immagine di avvicinamento. Essendo $O = (x_{33}, y_{33})$ il centro del modello e $P_j = (x_j, y_j)$, $j = 1, \ldots, 68$ tranne $j = 33$, uno degli altri punti di riferimento, il raggio r del modello è uguale alla distanza euclidea d tra O e il punto di riferimento più lontano, ovvero $r = d(O, P_i)$ dove $i = arg \max_j d(O, P_j)$. La Figura 5 raffigura un modello generico insieme ai parametri variabili, i cerchi concentrici appaiono in rosso, delimitano *annuli (o rings)* e sono numerati a partire da quello esterno; un quarto dell'intero modello appare in blu nella figura, il numero di quarti è fissato a 4 nel modello e sono numerati in senso orario a partire da quello positivo-positivo; le sezioni dividono i quarti in un numero di parti uguali di larghezza, i loro contorni appaiono in nero nella figura, dove una fetta è evidenziata in grigio e sono numerate in senso orario a partire dalla prima fetta nel primo quadrante; i settori rappresentano porzioni di sezioni comprese tra due cerchi concentrici vicini, ovvero le intersezioni di anelli con sezioni, sono numerate a partire

dall'intersezione dell'anello esterno con la prima fetta e procedendo in senso orario dall'anello esterno a quello interno, e un settore appare con contorni verdi nella figura.

La configurazione della ragnatela utilizzata in questo progetto è codificata come $4C\_4S\_inv4$, poichè è formata da quattro cerchi e ogni quarto è diviso in quattro sezioni. Il suffisso *inv* indica la distanza tra due cerchi consecutivi e in questa configurazione, il raggio per ogni cerchio è calcolato nel modo seguente:

- 4/10 * R (prima circonferenza)
- 7/10 * R (seconda circonferenza)
- 9/10 * R (terza circonferenza)
- R (quarta circonferenza)

Il numero di elementi nel vettore caratteristico restituito dal modello è uguale al numero di settori e il valore di ciascun *vector element* è il numero di *landmarks* che rientrano nel settore corrispondente, in base all'ordine definito nella parte destra di Figura 5.

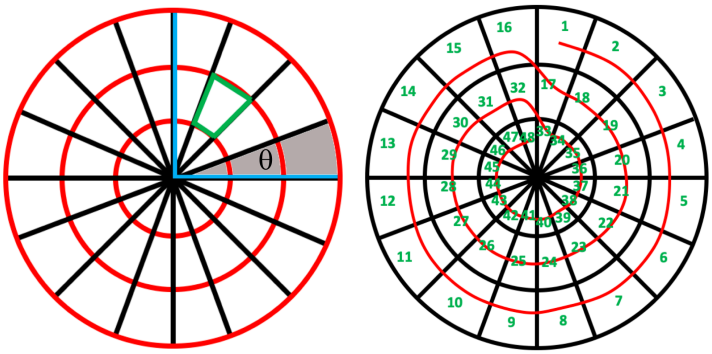


Figura 5 - A sinistra: il modello a forma di nastro applicato sul viso e centrato sul punto di riferimento P_33. A destra: l'ordine in cui i settori vengono analizzati per costruire il vettore caratteristica, a partire dal cerchio esterno del modello

L'applicazione del modello consente di identificare il settore in cui si trova ogni punto di riferimento. Il vettore di caratteristiche viene costruito in base al modello definito nella Figura 5. La dimensione del vettore è uguale al numero di settori. Questo si ottiene come m × 4 × n, ovvero il numero di sezioni moltiplicato per il numero di quarti (sempre 4) per il numero di cerchi (secondo la configurazione scelta, ciò porta a 4 fette × 4 quarti × 4 cerchi = 64 settori). Il vettore è costruito secondo il seguente algoritmo: Il vettore caratteristico ottenuto contiene nella sua posizione i il numero di punti di riferimento situati nel settore i (secondo l'ordine mostrato nell'immagine più a

destra nella Figura 5). Quindi in pratica è come se la ragnatela venisse srotolata come mostrato in Figura 6.

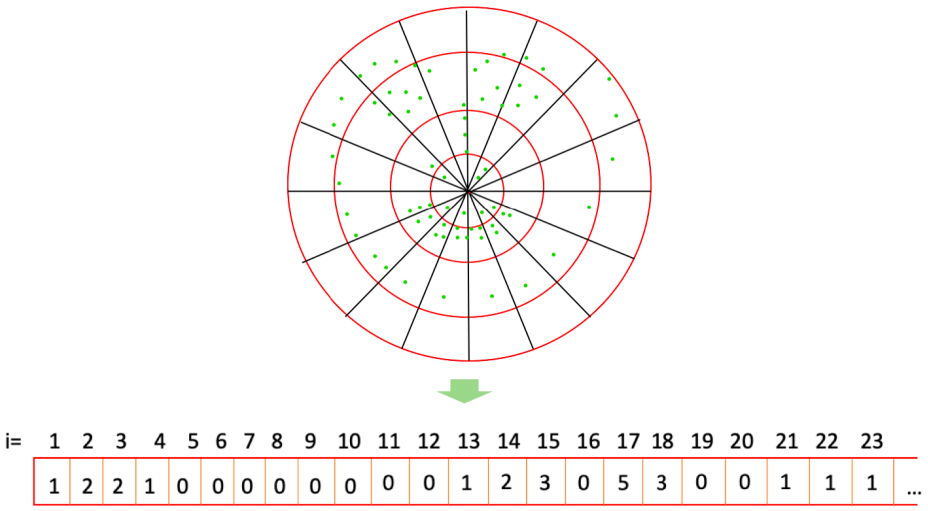


Figura 6 - La figura mostra come ogni landmark è associato a un settore all'interno di un modello con quattro cerchi e quattro sezioni per quarti e la costruzione finale del vettore caratteristico con elementi ordinati secondo la convenzione adottata

Rete Neurale

La stima del genere umano, utilizzando i vettori generati dallo Spider Web Model, è stata una sfida del tutto nuova. Attualmente in letteratura non sono ancora presenti classificazioni di genere, mediante l'utilizzo di questo recente modello a ragnatela, quindi è stata adottata più di una rete neurale per poter raggiungere al meglio buoni risultati.

La prima rete neurale convoluzionale che è stata utilizzata, denotata come ANN1 è descritta nella tabella seguente:

Layer	Output Shape	#Parametri	Funzione Attivazione
Dense1	128	8320	Relu
Dense2	64	8256	Relu
Dense3	32	2080	Relu
Dense4	16	528	Relu
Dense5	8	136	Relu
Dense6	1	9	Sigmoid

Con un numero totale di parametri 19329

La seconda rete neurale convoluzionale, denotata come ANN2 è descritta nella tabella seguente:

Layer	Output Shape	#Parametri	Funzione Attivazione
Dense1	64	4160	Relu
Dense2	32	2080	Relu
Dense3	32	1056	Relu
Dense4	16	528	Relu
Dense5	8	136	Relu
Dense6	1	9	Sigmoid

Con un numero totale di parametri 7969

Dataset utilizzati

CelebA

CelebFaces Attributes Dataset (CelebA) è un set di dati di attributi di facce su larga scala con oltre 200.000 immagini di celebrità, ognuna con 40 annotazioni di attributi. Le immagini in questo set di dati coprono grandi variazioni di posa e disordine dello sfondo. CelebA ha grandi diversità, grandi quantità e ricche annotazioni, tra cui:

- 10177 numero di identità,
- 202.599 numero di immagini di volti e
- 5 punti di riferimento, 40 annotazioni di attributi binari per immagine.

Il set di dati può essere utilizzato come set di addestramento e test per le seguenti attività di visione artificiale: riconoscimento degli attributi del volto, rilevamento del volto, localizzazione del punto di riferimento (o della parte del viso) e modifica e sintesi del volto.

Whe Dataset

Whe Dataset è un set di dati di attributi di volti di piccole dimensioni rispetto a CelebA. Le immagini in questo set di dati sono state accuratamente annotate per genere e colore della pelle (attributi protetti), nonché per fascia d'età, occhiali, posa della testa, fonte dell'immagine e dimensione del viso. I numeri di questo dataset sono:

- 6139 identità uniche
- 152917 immagini di volti

Architettura

L'architettura proposta per un sistema di classificazione del genere umano è mostrata in Figura 7. Questo sistema è formato principalmente da 3 fasi che eseguite in cascata l'una dopo l'altra portano a diversi risultati, che saranno mostrati nel seguito di questo documento. Le fasi principali sono:

- **Pre-processing**
- **Data preparation**
- **Training**

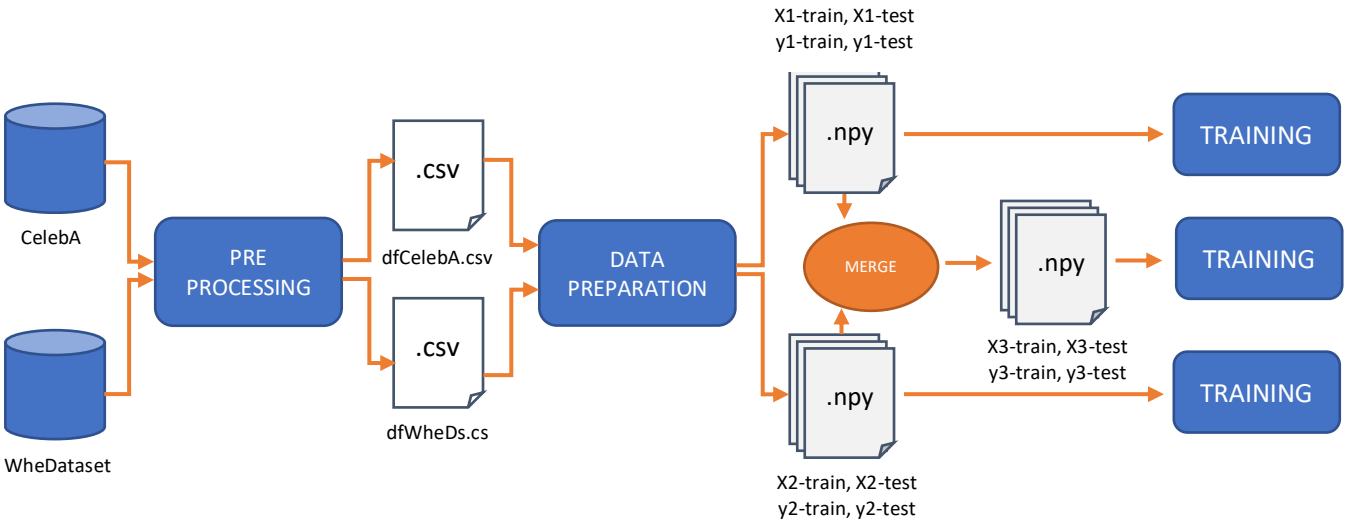


Figura 7 - Architettura proposta per un sistema di gender classification

Pre-processing

L'obiettivo principale di questa fase è quello di: estrarre da ogni immagine presente nei due dataset i volti dei soggetti tramite il modello descritto nella sezione *Face Detection*, procedere alla predizione dei 68 landmark facciali con l'algoritmo mostrato nella sezione *Landmark Prediction* e determinare il vettore caratteristica di ogni immagine tramite il modello descritto accuratamente nel paragrafo *Spider Web Model*. L'intero processo è mostrato in Figura 8.

Una volta effettuati questi passi, viene creato un data frame contenente il vettore caratteristica, l'etichetta del genere umano (0 = donna e 1 = uomo) e un attributo di *valid_face* (0 = volto non rilevato e 1 = volto rilevato) per ogni immagine presente nei dataset e tutti i dati vengono salvati in due file csv a secondo del dataset in questione.

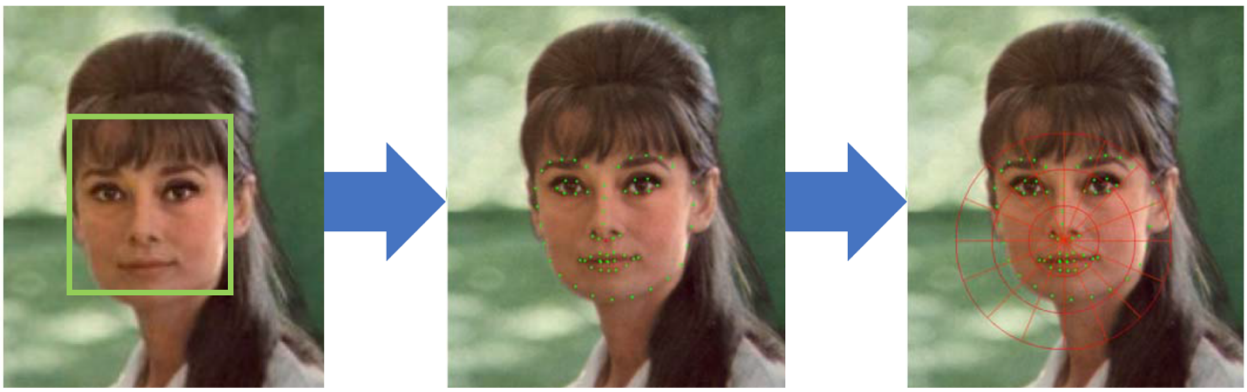


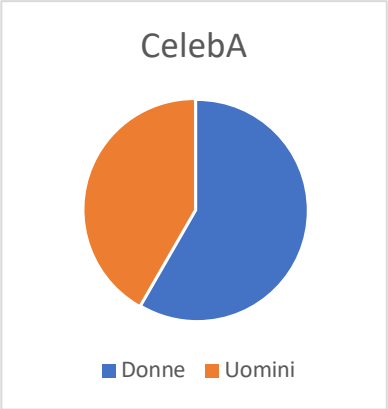
Figura 8 - Cascata di modelli utilizzzati per ricavare il vettore caratteristica

Data preparation

Una volta estrapolati i dati dalle immagini, i due file csv vengono preparati al training effettuando un'operazione di bilanciamento e normalizzazione dei dati.

Dal dataset CelebA sono stati estratti 202599 volti nella fase di pre processing di cui:

- 84.434 uomini
- 118.165 donne



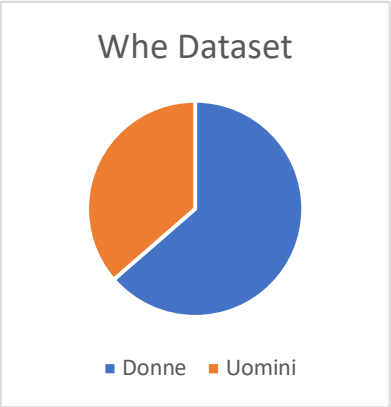
dopo l'operazione di bilanciamento:

- 84.434 uomini
- 84.434 donne

Successivamente il dataset è stato normalizzato e suddiviso al 30% in training set (X1-train, y1-train) e test set (X1-test, y1-test) e salvato in differenti file di array NumPy, in modo da valutare e confrontare i risultati di training sempre sullo stesso set di dati.

Da Whe Dataset sono stati estratti 101187 volti nella fase di pre processing di cui:

- 63.735 uomini
- 36.451 donne



dopo l'operazione di bilanciamento:

- 36.451 uomini
- 36.451 donne

Successivamente il dataset è stato normalizzato e suddiviso al 30% in training set (X2-train, y2-train) e test set (X2-test, y2-test) e salvato in differenti file di array NumPy, in modo da valutare e confrontare i risultati di training sempre sullo stesso set di dati.

Infine sono stati uniti i due dataset (non-bilanciati) e tramite un'operazione di shuffle è stato creato un nuovo dataset con:

- 148.169 uomini
- 154.616 donne

Dopo l'operazione di bilanciamento di quest'ultimo dataset siamo arrivati a:

- 148.169 uomini
- 148.169 donne

Training

Il training dei dati è stato effettuato in modo separato su ogni set di dati con le due reti neurali convoluzionali descritte nel paragrafo Rete Neurale. La fase di addestramento è stata eseguita dando in input al classificatore i file di array NumPy generati nella fase di data preparation e per ogni settaggio di parametri dei modelli e per ogni dataset utilizzato sono stati generati la matrice di confusione e la tabella con i valori di precision, recall e accuracy.

Per riutilizzare in futuro la ANN che ha prodotto maggiori performance, si è pensato di salvare i vari pesi di ogni rete sperimentata, in modo tale da non dover riaddestrare il modello ottimale.

Risultati ottenuti

ANN1

I risultati ottenuti sulla prima ANN non sono stati molto ottimali, ma comunque hanno fatto ben sperare trattandosi della prima rete neurale convoluzionale di questo progetto.

Come possiamo notare sul grafico in Figura 9 i risultati ottenuti sul train set sono stati molto alti infatti ogni dataset ha raggiunto la soglia dell'80% di accuracy, cosa che ha portato subito ad un'ottima speranza che però non è stata confermata andando ad analizzare l'accuracy dei test set sui i tre dataset.

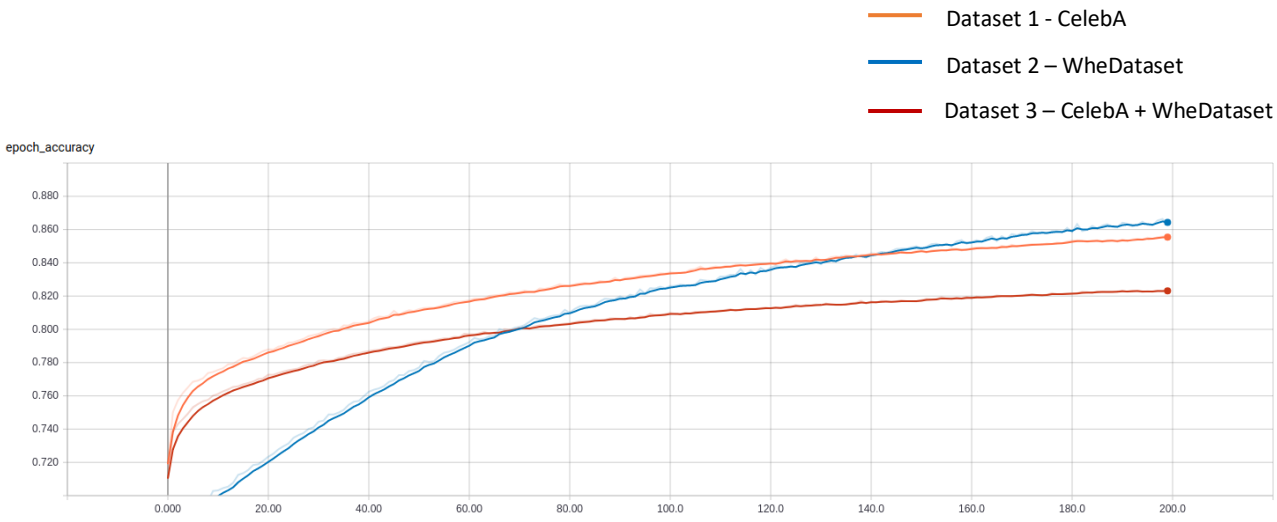


Figura 9 - Grafico di accuracy con ANN1 e 200 epoche

Dataset 1 – CelebA

Sul dataset CelebA con la ANN 1 e con 200 epoche e batch-size 64 sono stati ottenuti i risultati seguenti:

CONFUSION MATRIX		Actual Value	
		positives	negatives
Predicted Value	positives	19161	6117
	negatives	6620	18763

METRIC ACCURACY	precision	recall	f1 - score
0	0.74	0.76	0.75
1	0.75	0.74	0.75
accuracy			0.75

Dataset 2 – Whe Dataset

Sul dataset Whe Dataset con la ANN 1 e con 200 epoche e batch-size 64 sono stati ottenuti i risultati seguenti:

CONFUSION MATRIX		Actual Value	
		positives	negatives
Predicted Value	positives	6435	4474
	negatives	3249	7713

METRIC ACCURACY	precision	recall	f1 - score
0	0.66	0.59	0.62
1	0.63	0.70	0.67
accuracy			0.65

Dataset 3 – CelebA + Whe Dataset

Sul dataset combinato con la ANN 1 e con 200 epoche e batch-size 128 sono stati ottenuti i risultati seguenti:

CONFUSION MATRIX		Actual Value	
		positives	negatives
Predicted Value	positives	**33544**	**10985**
	negatives	**12196**	**32177**

METRIC ACCURACY	precision	recall	f1 - score
0	**0.73**	**0.75**	**0.74**
1	**0.75**	**0.73**	**0.74**
accuracy			**0.74**

ANN2

Nella seconda ANN i risultati ottenuti sono stati molto confortevoli e corrispondono ai risultati finali di questo progetto di sperimentazione.

Come possiamo notare sul grafico in Figura 10 i risultati ottenuti sul train set con 100 epoche sono stati più bassi rispetto alla prima ANN ma hanno portato a valori di accuracy dei test set molto più alti rispetto alla prima rete neurale convoluzionale. Ciò ci fa pensare che nella prima rete c'è stato un overfitting dei dati andando quindi a fornire ottimi risultati sul train set e scarsi risultati sul test set.

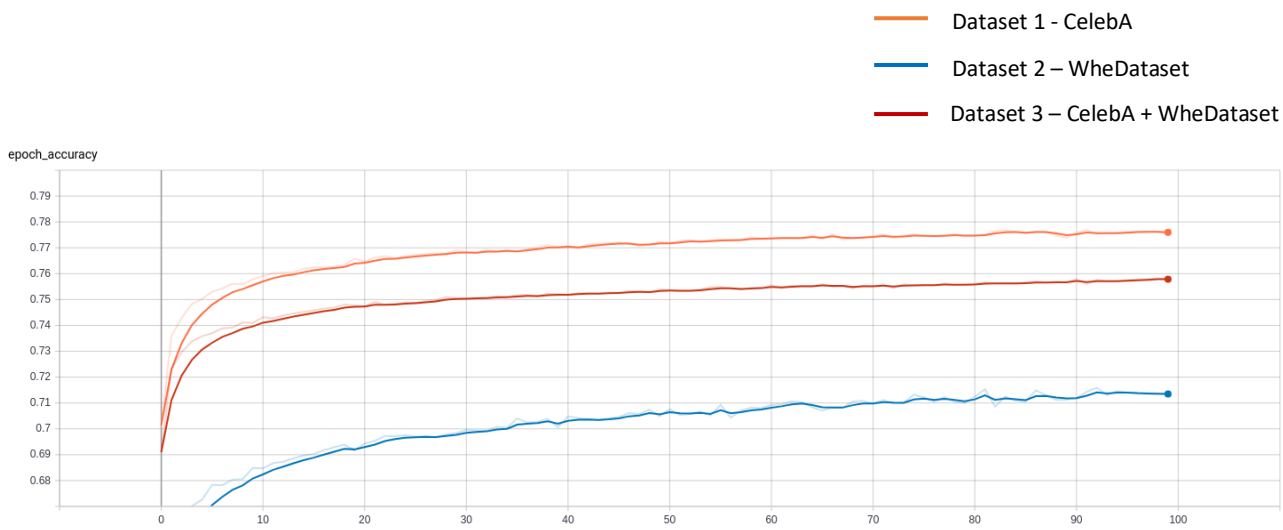


Figura 10 - Grafico di accuracy con ANN2 e 100 epoche

Con l'aumentare dei valori di accuracy sul train e test set su tutti e tre i dataset, si è deciso di portare il numero di epoche a 200 in modo da poter raggiungere risultati ancora migliori. Questi ultimi sono rimasti stazionari come mostrato nel grafico in Figura 11.

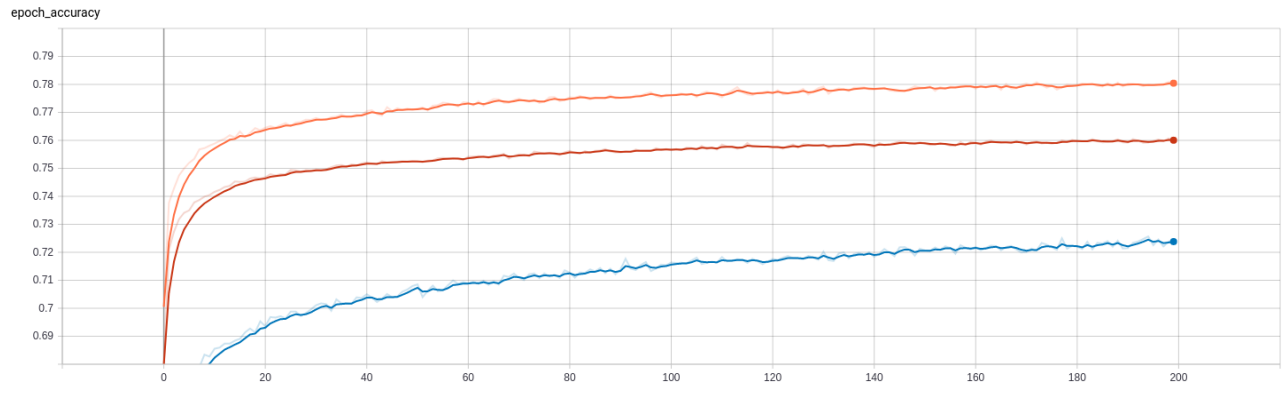


Figura 11 - Grafico di accuracy con ANN2 e 200 epoche

Dataset 1 – CelebA

Sul dataset CelebA con la ANN 2 e con 100 epoche e batch-size 32 sono stati ottenuti i risultati seguenti:

CONFUSION MATRIX		Actual Value	
		positives	negatives
Predicted Value	positives	**19575**	**5703**
	negatives	**5785**	**19598**

METRIC ACCURACY	precision	recall	f1 - score
0	**0.77**	**0.77**	**0.77**
1	**0.77**	**0.77**	**0.77**
accuracy			**0.77**

con 200 epoche e batch-size 64 sono stati ottenuti i risultati seguenti:

CONFUSION MATRIX		Actual Value	
		positives	negatives
Predicted Value	positives	**19943**	**5335**
	negatives	**6323**	**19060**

METRIC ACCURACY	precision	recall	f1 - score
0	**0.76**	**0.79**	**0.77**
1	**0.78**	**0.75**	**0.77**
accuracy			**0.77**

Dataset 2 – Whe Dataset

Sul dataset Whe dataset con la ANN 2 e con 100 epoche e batch-size 32 sono stati ottenuti i risultati seguenti:

CONFUSION MATRIX		Actual Value	
		positives	negatives
Predicted Value	positives	**6839**	**4070**
	negatives	**2812**	**8150**

METRIC ACCURACY	precision	recall	f1 - score
0	**0.71**	**0.63**	**0.67**
1	**0.67**	**0.74**	**0.70**
accuracy			**0.69**

con 200 epoche e batch-size 64 sono stati ottenuti i risultati seguenti:

CONFUSION MATRIX		Actual Value	
		positives	negatives
Predicted Value	positives	**6842**	**4067**
	negatives	**2857**	**8105**

METRIC ACCURACY	precision	recall	f1 - score
0	**0.71**	**0.63**	**0.67**
1	**0.67**	**0.74**	**0.70**
accuracy			**0.68**

Dataset 3 – CelebA + Whe Dataset

Sul dataset combinato con la ANN 2 e con 100 epoche e batch-size 64 sono stati ottenuti i risultati seguenti:

CONFUSION MATRIX		Actual Value	
		positives	negatives
Predicted Value	positives	**31957**	**12572**
	negatives	**9260**	**35113**

METRIC ACCURACY	precision	recall	f1 - score
0	**0.78**	**0.72**	**0.75**
1	**0.74**	**0.79**	**0.76**
accuracy			**0.75**

con 200 epoche e batch-size 64 sono stati ottenuti i risultati seguenti:

CONFUSION MATRIX		Actual Value	
		positives	negatives
Predicted Value	positives	**32413**	**12116**
	negatives	**9721**	**34652**

METRIC ACCURACY	precision	recall	f1 - score
0	**0.77**	**0.73**	**0.75**
1	**0.74**	**0.78**	**0.76**
accuracy			**0.75**

Conclusioni

Bibliografia

[1] P. B. C. Bisogni, Gender Classification using a Spider Web Method, Presentazione, 2020.

[2] V. B. N. D. A. M. Burton, What's the difference between men and women? evidence from facial measurement, Perception, vol. 22, 1993..

[3] H. D. E. F. G. a. A. Y. V. Bruce, Parallel processing of the sex and familiarity of faces, Canadian Journal of Psychology, vol. 41, 1987..

[4] S. B. C. B. M. D. M. M. N. P. Barra, Web-Shaped Model for Head Pose Estimation: An Approach for Best Exemplar Selection, IEEE transaction on image processing, VOL. 29, 2020.

[5] D. E. King, Max-Margin Object Detection, 2015.

[6] V. K. a. J. Sullivan, One millisecond face alignment with an ensemble of regression trees, IEEE Conf. Comput. Vis. Pattern Recognit, 2014.

[7] P. W. P. M. R. a. H. B. M. Kostinger, Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization, Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops), 2011,.