Contest di Visione Artificiale: Gruppo 08

Marco Preziosi, Salvatore Reina, Stefano Saldutti, and Bruno Vento

{m.preziosi9, s.reina1, s.saldutti1, b.vento}@studenti.unisa.it

1 Introduzione

Lo scopo del progetto è quello di realizzare una rete neurale per la stima dell'età di una persona. Per questo scopo, è stato addestrato un classificatore. Tutto il codice sorgente è disponibile al seguente link di github.

2 Descrizione della soluzione

2.1 Convolutional neural network

L'architettura della rete si basa su Fine Tuning della rete MobilenetV2 con input size pari a 160x160, partendo dai pesi ottenuti su "imagenet". Sia la tipologia di rete sia l'input size usata sono state scelte per ottenere un buon compromesso fra velocità di elaborazione e performance. Una rete con input size più grande avrebbe migliorato le performance del sistema, a discapito del tempo richiesto. Per quanto riguarda la rete, invece, la scelta è ricaduta su MobilenetV2 in quanto ha una dimensione piccola, bassa latenza e per soddisfare i vincoli di risorse di una varietà di casi d'uso, rispetto ad altre reti più "pesanti".

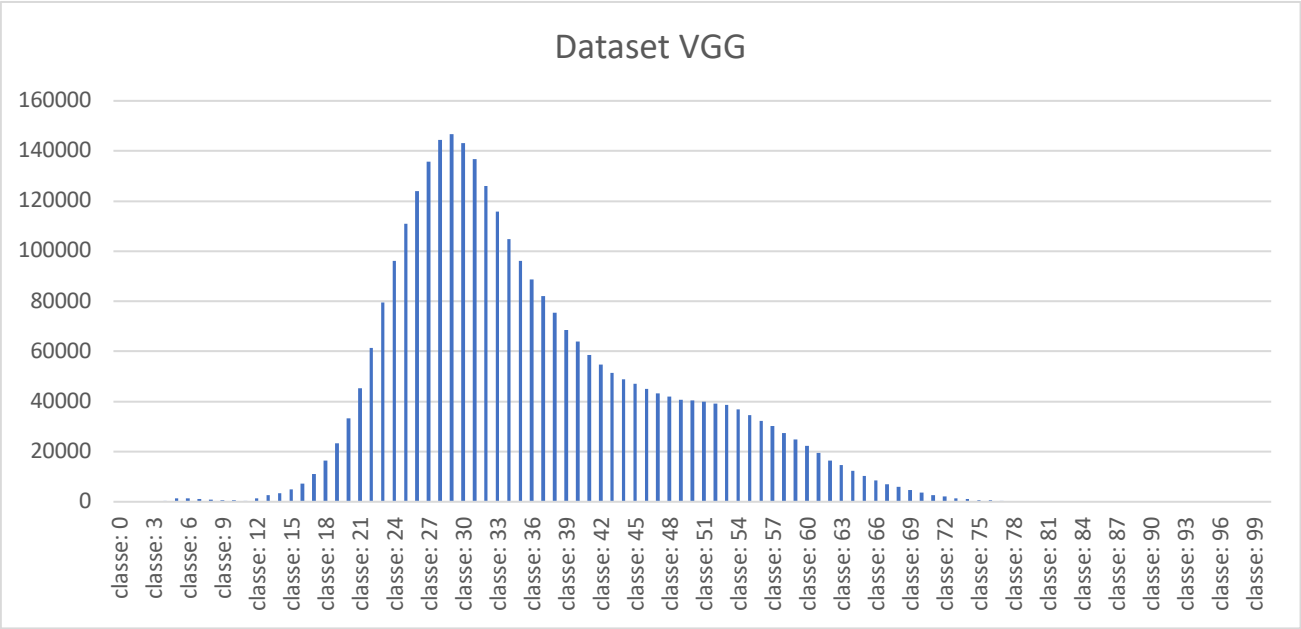
Per adattare il sistema al nostro scopo, l'ultimo livello della rete MobilenetV2 è stato sostituito con 2 strati densi di 512 e 128 neuroni ciascuno (con funzione di attivazione "relu"), entrambi seguiti da uno strato di Dropout impostato a 0.20, e infine con un ultimo strato denso con funzione di attivazione "softmax", costituito da 101 neuroni, uno per ogni classe da predire. I 2 strati densi che precedono l'ultimo costituito da 101 classi sono stati inseriti affinché la rete possa sfruttare e combinare in modo appropriato le feature estratte dalla CNN MobilenetV2.

La funzione di costo scelta è la "categorical cross-entropy", in quanto si adatta bene al task di classificazione multi-classe. Ogni probabilità di classe predetta viene confrontata con l'output 0 o 1 desiderato della classe effettiva e viene calcolato una perdita che penalizza la probabilità in base a quanto è distante dal valore atteso effettivo. Questa funzione di costo richiede che le labels siano codificate in one-hot encoding.

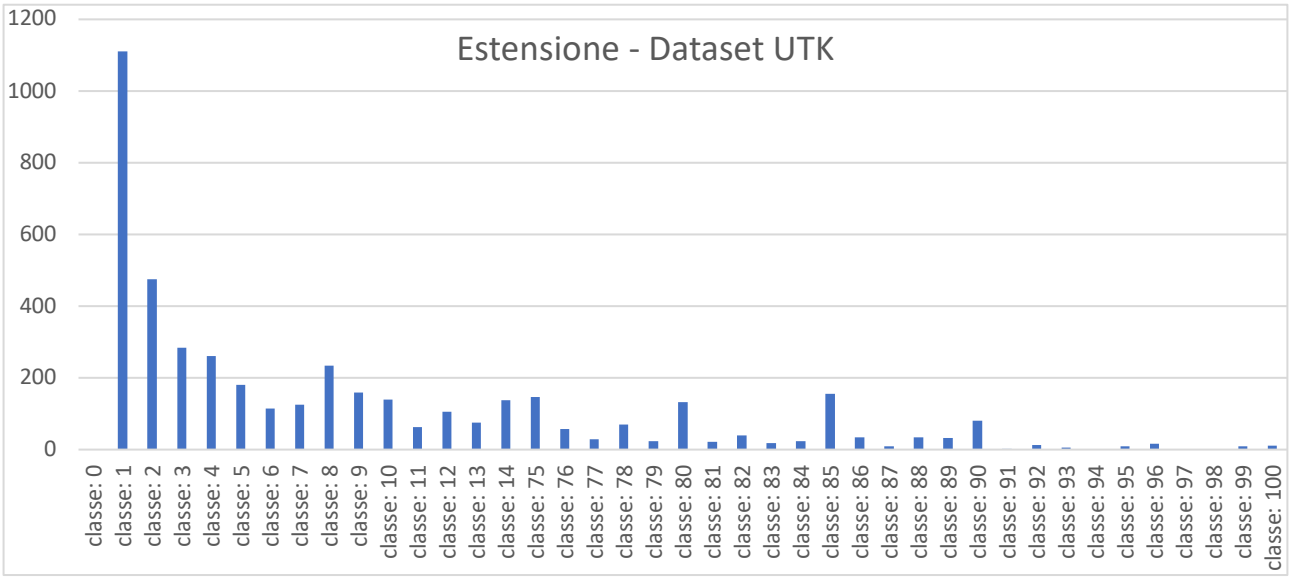
2.2 Procedura di addestramento

La procedura di addestramento consiste in:

- **Analisi del dataset.** La porzione di training del dataset VGG-Face fornitoci è stata analizzata calcolando il numero di campioni per ciascuna classe d'età. Da questa analisi preliminare è emerso che le classi con un maggior numero di campioni sono quelle comprese nella fascia 25 - 34 che superano i 100.000 samples ciascuna, mentre le classi meno rappresentate sono quelle comprese nella fascia 0 - 13 e 71 - 82 (con meno di 3.000 campioni ciascuna) e nella fascia 83-100 con 0 campioni. Pertanto, si è resa necessaria un'estensione del dataset VGG-Face con un altro che contenesse campioni di quelle specifiche classi.



- **Estensione del dataset**. Il dataset scelto per l'estensione è UTK-Face, un dataset che contiene più di 20.000 campioni di volti "in-the-wild" e con differenti pose, espressioni facciali, condizioni di illuminazione, occlusione e risoluzione. La scelta è ricaduta su tale dataset in quanto ha una buona reliability per le annotazioni, delle quali sono state prelevate le sole relative all'età. Inoltre, alle immagini scelte da questo dataset è già stato eseguito il crop.



- **Selezione del numero di campioni e divisione in training e validation**. Per ragioni di tempo necessario all'addestramento, si è scelto di considerare soltanto una porzione del dataset VGG-Face. Per la selezione dei campioni, si è tenuto conto del gran numero di immagini per ciascuna persona. Per evitare di prelevare troppi campioni da un singolo soggetto, ci si è limitati a prenderne solo 20. Inoltre, per ottenere bilanciamento fra le classi, sono state prese 2.000 per ciascuna età. In particolare, l'algoritmo di selezione del numero di campioni prevede:

    - i campioni appartenenti a UTK-Face sono stati considerati tutti e, di conseguenza, il numero di campioni di VGG-Face per ogni classe d'età è 2.000 – num_UTK-Face;

- si prendono k=20 campioni per ogni identità, fino a raggiungere la soglia di 2.000;
- se il numero di immagini selezionate al passo precedente non raggiunge 2.000, i rimanenti campioni sono scelti casualmente tra quelli ancora disponibili per ciascuna classe.

In totale, sono state prese 137599 immagini. Il training set è costituito dal 90% delle immagini di ogni classe e il validation set dal 10%, quindi rispettivamente 123839 e 13759 immagini. A tal proposito, sono stati realizzati i csv, in cui per ogni riga vi è il nome dell'immagine con la relativa annotazione.

- **Crop delle immagini e ridimensionamento**. Alle immagini appartenenti al dataset VGG-Face è stato eseguito il crop secondo le annotazioni fornite nel framework GenderRecognition, tramite un tool che legge il csv, estrae le coordinate e ritaglia l'immagine. Inoltre, per tutto il dataset scelto (sia immagini di VGG-Face sia quelle di UTK-Face), le immagini sono state ridimensionate alla dimensione 160x160 (l'input size della rete).
- **Codifica one-hot e realizzazione del dataset nel formato h5**. Le labels sono state trasformate in formato one-hot, ovvero con un vettore di dimensione 101 in cui sono presenti tutti 0, eccetto che per un 1 in corrispondenza della posizione relativa alla classe specifica. Con il fine ottenere delle batch bilanciate (ossia che contengano 100 elementi, uno per classe), il file csv del training set è stato ulteriormente riorganizzato in blocchi contigui di 101 elementi ordinati in ordine crescente in base alla classe. Successivamente, per l'addestramento si è scelto di generare due file in formato h5, che contengono uno le immagini di training e validation e l'altro le relative labels.
- **Addestramento della rete**, **parametri e data augmentation**. Per via del poco tempo a disposizione per l'addestramento, si è deciso di impostare 1.000 come numero di epoche, considerando una politica di early stopping con patience di 20 e monitorando la loss sul validation set. La scelta dell'early stopping è stata conveniente, in quanto l'addestramento è terminato prima della conclusione delle epoche previste, ovvero 70. L'ottimizzatore scelto è Adam e in combinazione con questo, è stata utilizzata una politica di "ReduceLROnPlateau", per decrementare il valore del learning rate (qualora le performance sul validation non aumentino per un certo numero di epoche) ed evitare grandi oscillazioni. In particolare, il learning rate iniziale è stato quello di default dell'ottimizzatore Adam. La singola batch, di dimensione pari a 100, è generata in modo customizzato: si sceglie un indice casuale fra 0 e la lunghezza del dataset e con un meccanismo a coda circolare si prelevano i 100 elementi successivi, di cui si effettua anche uno shuffle, ma preservando il vincolo per cui una batch contiene un elemento per classe. Per quanto riguarda la Data Augmentation, questa è stata svolta a runtime durante l'addestramento. In particolare, data un'immagine appartenente alla batch, si decide a caso se applicare una trasformazione e, in caso affermativo, si decide a caso la tipologia, scegliendo fra flip orizzontale, rotazione e cambiamento di luminosità. Come parametro di accuratezza è stata scelta la "contrastive_accuracy" in quanto tiene conto anche del grado di confidenza con cui la rete effettua la predizione della classe.

3 Risultati sperimentali

Per la valutazione dei risultati, è stata utilizzata la metrica del Mean Absolute Error, ossia la media delle differenze in valore assoluto fra l'età stimata e quella reale. In particolare, la suddetta metrica è stata calcolata sia sul training set sia sul validation set nel seguente modo:

- Dato l'insieme di campioni, effettuiamo la predizione;
- Per ciascuna predizione, troviamo l'indice che ha associata la probabilità maggiore, che corrisponde alla classe dell'età predetta;

- Si effettua la somma delle differenze fra la classe predetta e la classe vera (quest'ultima prelevata dal vettore di labels in one-hot encoding, trovando l'indice a cui è associato il valore 1);
- Infine, si effettua la media di tale somma.

Il Mae calcolato è:

- Training set: 2.32
- Validation set: 3.28

Inoltre, per testare il nostro modello con più campioni, abbiamo scelto di prendere le immagini presenti nel dataset VGG-Face che non abbiamo incluso in quello usato per l'addestramento della rete. Su un insieme di 600 mila samples, su cui è stato eseguito il crop, abbiamo ottenuto un mae pari a 3.02

4 Conclusioni

Considerando il basso valore del Mean Absolute Error ottenuto e la difficoltà dell'age estimation su 101 classi (anche per l'essere umano), ci riteniamo soddisfatti del risultato ottenuto. Per migliorare ulteriormente le performance della rete, si potrebbe addestrare il sistema su una quantità ancora maggiore di campioni rispetto a quelli già forniti. Se, invece, si vuole andare incontro agli stringenti vincoli imposti dall'utilizzo di un modello real-time, si potrebbe adottare una CNN più "leggera", cercando di non sacrificare le performance della rete.