A Comparative Analysis: Dual-Task CNN vs. Single-Task CNNs for Gender and Age Prediction in Facial Images

Paolo Speziali
paolo.speziali@studenti.unipg.it

***Abstract*—Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse non eleifend elit. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut ligula nulla, placerat ut porta vitae, efficitur ut ipsum. Aenean sodales lacus et mauris faucibus, at congue turpis consectetur. Mauris volutpat a velit ac commodo. In dui urna, pulvinar interdum varius at, volutpat non ipsum. Integer hendrerit convallis laoreet. Nunc in mattis diam. Donec at hendrerit risus, vel pellentesque tortor. In hendrerit malesuada elementum. Suspendisse nibh dolor, condimentum non nisi id, laoreet tincidunt mi.**

I. Introduction

In the realm of computer vision and image processing, the ability to accurately predict gender and age from facial images holds significant importance across various applications that benefit from the demographic data of their users.

Biometric information can, in fact, be used in a plethora of ways, ranging from targeted commercial use [1] to intelligent non-profit campaigns [2] and even extending to Orwellian credit scoring systems [3].

A heated debate continues to unfold regarding the application and potential abuse of Machine Learning and Computer Vision technologies in the daily lives of citizens, particularly heightened since the advent of Convolutional Neural Networks (CNN). The transformative capabilities of CNNs lie in their ability to process vast amounts of data and generate remarkably accurate predictions. Notably, these networks eliminate the necessity for manual feature engineering tasks, such as feature extraction, thereby rendering the implementation and utilization of these technologies more convenient than ever before [4].

This evolution raises significant questions about privacy, ethics, and the broader societal impact of seamlessly integrating advanced algorithms into various aspects of our lives [5]. Notably, entities such as the European Union Commission and Parliament have actively addressed these concerns by formulating the AI Act [6]. This legislative initiative seeks to classify the risks associated with ML and CV technologies, especially concerning the citizens of the confederation. The primary objective is to safeguard individuals from the inappropriate use of their biometric data, acknowledging the critical need to establish regulatory frameworks that balance the advancement of technology with the protection of individual rights and privacy [7].

Given the circumstances, it is crucial to understand the design techniques and architectures upon which these models are based to utilize and implement them with increased awareness and consideration for the effects and consequences on end-users. In particular, we will undertake a comparison between a multi-task CNN and two single-task CNNs for age and perceived gender detection to assess their results and differences in task execution.

II. Related Work

Several advancements have been made in the field of gender and age prediction from facial images, utilizing deep learning techniques. In this section, we present two notable works that have contributed significantly to the state-of-the-art in this domain.

The work by [Rafique et al., 2019] [8] introduces a deep learning framework based on an ensemble of attentional and residual convolutional networks with the primary objective of predicting gender and age groups with a high accuracy rate, treating both features as a classification problem. The proposed model, trained on the UTKFace dataset, employs attention mechanisms to focus on crucial facial regions, enhancing the accuracy of predictions. The multi-task learning approach is utilized, and the feature embedding of the age classifier is augmented with predicted gender information.

In a different approach, [Antipov et al., 2017] [9] explores improvements in existing CNN-based methods for gender and age prediction. The study investigates key factors that impact the training of CNNs, including target age encoding, loss function, CNN depth, pretraining necessity, and training strategy (mono-task or multi-task). The authors present state-of-the-art gender recognition and age estimation models designed according to benchmarks such as LFW, MORPH-II, and FG-NET. Notably, their best model won the ChaLearn Apparent Age Estimation Challenge 2016, significantly outperforming the solutions of other participants.

III. Proposed Approach

A. Dataset and Preprocessing

The dataset we will be utilizing is the UTKFace dataset [10], which consists of $23,708$ aligned and cropped facial RGB images, annotated with age, gender, and ethnicity labels. This dataset was created with the intention of covering a wide range of variations, including pose, facial expression, illumination, occlusion, resolution, and more. In our analysis, we will specifically concentrate on the first two attributes within the dataset: age and gender.

In our analysis, we will exclusively consider 70% of the dataset for our training set, with the remaining 30% designated for testing purposes. This division allows us to train our models on a substantial portion of the data while maintaining a separate, untouched set for rigorous evaluation.

The training set exhibits an age distribution depicted in the histogram shown in Fig. 1, along with a balanced gender distribution, as illustrated in Fig. 2.

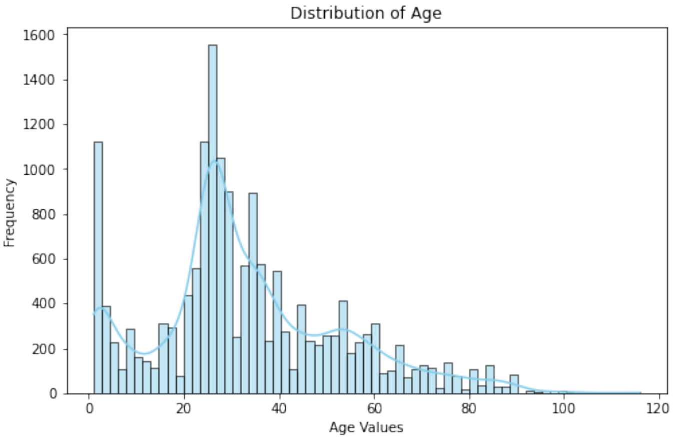

Fig. 1: Histogram of age distribution in the training set

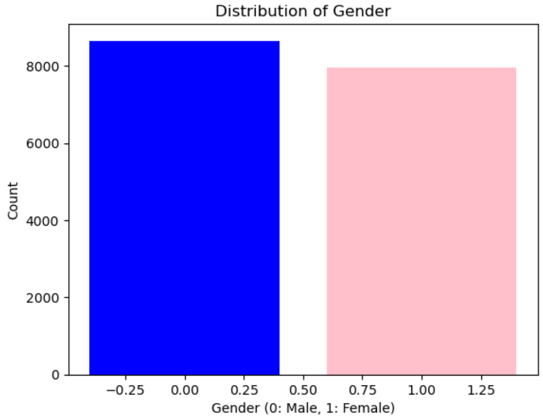

Fig. 2: Histogram of gender distribution in the training set

We further split the training set into a 70% training set and a 30% validation set. The validation set plays a critical role in refining and making our model more robust, preventing overfitting through techniques such as early stopping.

Despite this subdivision, we have taken measures to maintain the balance in the dataset: the accompanying histogram in Fig. 3 illustrates that, concerning gender, the distribution remains similar in both datasets.

As for age, a continuous value that poses challenges for balance assessment, we will employ the non-parametric Kolmogorov-Smirnov test to scrutinize the absence of statistically significant differences in the distributions of its values between the two datasets (its null hypothesys) [11].

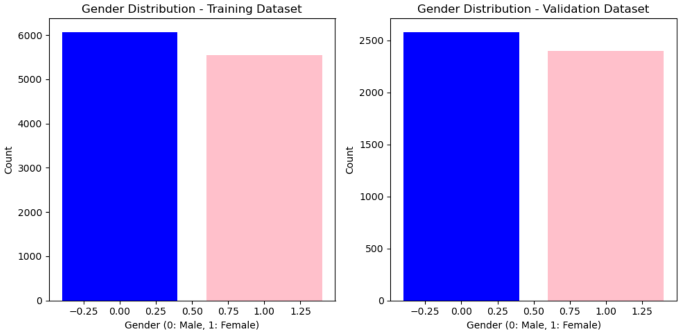

Fig. 3: Histogram of gender distribution in the new training set and the validation set

The outcome of the aforementioned test yields a p-value of approximately 0.38, this value is reasonably high, leading us to consider it sufficient evidence to accept the null hypothesis. Thus, we conclude that the two distributions of age in the two datasets do not exhibit statistically significant differences. In the plot shown in Fig. 4, we can observe that the cumulative distribution functions of the age values of the two datasets are almost identical.

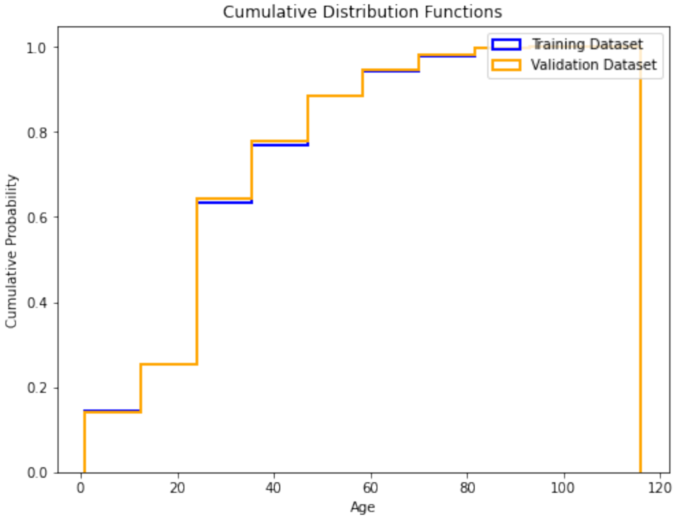

Fig. 4: CDF of age values in the new training set and the validation set

The concluding step in our dataset preprocessing involves the application of data augmentation techniques. Specifically, we expand the dataset by incorporating horizontally mirrored images and introducing random rotations of up to 10 degrees. Additionally, we employ random adjustments to the brightness, contrast, saturation, and hue of the images. This augmentation strategy has been employed based on empirical evidence suggesting that enlarging the dataset enhances the model's performance. By introducing these variations in orientation and color, we aim to expose the model to a more diverse set of examples, ultimately improving its ability to generalize and make accurate predictions on unseen data.

B. Model Architecture

As previously mentioned, in this project, we will compare two CNN architectures, one comprising a single multi-task network and another consisting of two single-task networks. The overarching goal is common between them.

Let's delve into the details of the both architectures.

1) Multi-task Architecture: The multi-task architecture, shown in Fig. 5, features a structure of the following type:

$$\texttt{INPUT} \rightarrow [\texttt{CONV} \rightarrow \texttt{BATCHNORM} \rightarrow \texttt{RELU} \rightarrow \texttt{POOL}] \times 4$$

it then divides it into two branches, each retaining the same structure:

$$\texttt{FC} \rightarrow \texttt{BATCHNORM} \rightarrow \texttt{RELU} \rightarrow \texttt{FC}$$

The key distinction lies in the purpose of each branch: one branch is designed to address the classification problem, specifically predicting gender, so it has two output neurons that perform a softmax function. The other branch is designed to solve the regression problem, involving the prediction of age, it has a single output neuron that performs a linear function.

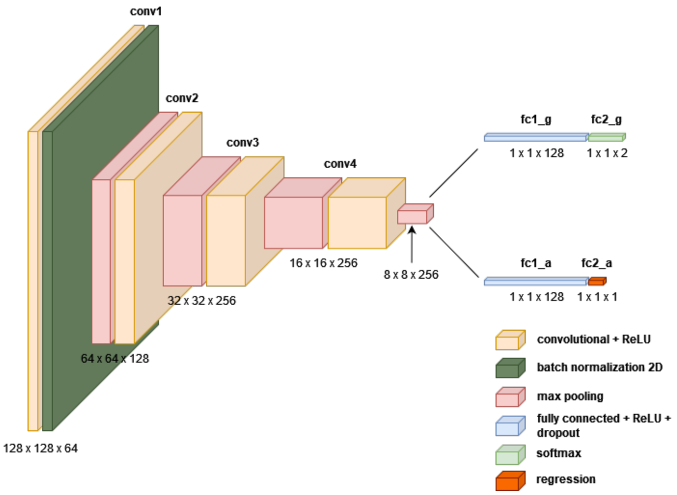

Fig. 5: Graphical representation of the Multi-task architecture

Let's explore in more detail the structure of the various layers:

- The input, with dimensions $128 \times 128 \times 3$, enters the first 2D convolutional layer comprising 64 filters of size 3×3. These filters apply a padding of 2 to preserve spatial dimensions after convolution. The stride is set to 1, consistent with every other convolutional layer, and set to 2 in the subsequent pooling layer. This setting avoids requiring the network to learn to subsample.
- Following this, a batch normalization layer is applied to the output of the previous convolutional layer. Batch normalization helps stabilize and accelerate the training of neural networks by normalizing the batch to its mean and standard deviation instead of doing it with the whole dataset. This is the only occurrence of batch normalization in the network.
- The output of the batch normalization then enters the Rectified Linear Unit (RELU) activation function, it introduces non-linearity to the model, allowing it to learn from more complex patterns in the data.
- Before entering another convolutional layer, the output of the preceding layer undergoes pooling through a max pooling layer with a window size of 2×2 and a stride of 2. As a result, the size of the produced map is halved.
- The previous steps are repeated three more times, with the only difference being the number of filters in the second convolutional layer (5 instead of 3), the number of filters, which is doubled in each subsequent convolutional layer, and the absence of batch normalization layers.
- The output of the last pooling layer is flattened and fed into two speárated branches of fully connected layers, each with 128 neurons, followed by a dropout layer (set at 0.25) and ReLU activation.
- The output of the last fully connected layer of the first branch is fed into a fully connected layer with two neurons, one for each class of the classification problem (`Male` or `Female`), followed by a softmax activation.
- The output of the last fully connected layer of the second branch is fed into a fully connected layer with one neuron for the regression problem (age), followed by a linear activation.

Since we have incorporated a batch normalization layer, it has proven unnecessary to include dropout layers in the convolutional part of the network. The combined use of these techniques is generally discouraged, as a rule of thumb, since it tends to yield worse results than when only one of them is employed [12]. However, we have included a dropout layer in the fully connected part.

We utilize Cross Entropy Loss as the loss function for classification and Mean Squared Error (MSE) Loss for regression. These losses are summed and employed for backpropagation at each step.

We have employed an Adam optimizer with a learning rate set to 0.01.

2) Single-task Architecture: The single-task architecture, shown in Fig. 6, features the same structure as the multi-task architecture, with the only difference being that, instead of branching out in two different layers for the tasks, it has a single branch that performs either classification or regression. For this reason, the single-task architecture is composed of two separate networks, one for each task.

The loss functions used are the same as the ones employed in the multi-task architecture, the only difference being that, instead of summing them and using the sum for backpropagation, we use them separately for each network.

We have employed an Adam optimizer with a learning rate set to 0.01 for the classification network and 0.001 for the regression network.

In both architectures, the batch size is 64. The metrics used for evaluation are accuracy, defined as the ratio of correctly predicted instances to the total instances, expressed as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

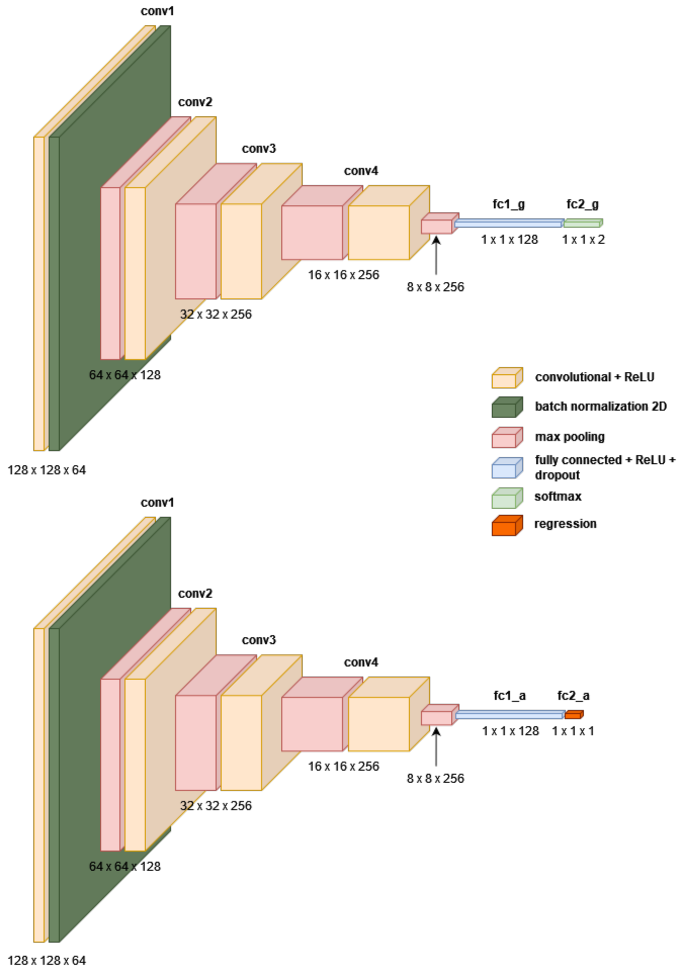

Fig. 6: Graphical representation of the Single-task architecture

for classification tasks, and R-squared (R^2) score, defined as the proportion of the variance in the dependent variable that is predictable from the independent variable(s), expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

for regression tasks.

IV. Experiments

A. Training Environment and Hardware Specifications

The model training was conducted within a Python Notebook environment in Anaconda, utilizing PyTorch libraries with GPU acceleration. The machine used features an AMD Ryzen 7 5800X CPU with 8 cores and 16 threads, operating at 4 GHz and it is equipped with 16 GB of RAM running at 3600 MHz and a PNY NVidia RTX 3070 graphics card with 8GB of VRAM and 5888 CUDA cores.

B. Training Process

The model training process utilized a total of seconds, as indicated in the following table:

Time (sec)	**Age task**	**Gender task**
Single-task CNN	1223.41	747.39
Multi-task CNN	640.34	*

Additionally, the epoch count for each training network is presented below:

Num. of epochs	**Age task**	**Gender task**
Single-task CNN	17	10
Multi-task CNN	9	*

In each training iteration, an early stopping mechanism was implemented with a patience value set to 3 for both architectures. The patience hyper-parameter was determined by monitoring the validation loss of the model and stopping the training process when the loss did not improve for a number of epochs equal to the patience value.

The training process was halted when the loss values on the training set reached:

Last loss value (training)	**Age task**	**Gender task**
Single-task CNN	0.7951	0.0027
Multi-task CNN	1.2228	*

and on the validation set:

Last loss value (validation)	**Age task**	**Gender task**
Single-task CNN	1.3360	0.0041
Multi-task CNN	1.4058	*

During the backpropagation, the loss value for the age task, was normalized by multiplying it by a factor $\lambda_{\text{age}} = 0.01$. By scaling down the loss for the age task, it effectively equalized the magnitudes of both losses, ensuring a more balanced and comparable training signal during the optimization process.

In Fig. 7, Fig. 8 and Fig. 9, we can observe the trends through the epochs of the loss in the single task for age, the single task for gender and the multi-task scenario, respectively. In Fig. 10 and Fig. 11, we can observe the trends through the epochs of the accuracy in the single task and the multi-task scenario, respectively.

The training process ended with the following metrics on the training set:

Training Metrics	**Age task ($\mathbf{R^2}$)**	**Gender task (Accuracy)**
Single-task CNN	0.8664	0.9278
Multi-task CNN	0.7955	0.9112

and the following on the validation set:

Validation Metrics	**Age task ($\mathbf{R^2}$)**	**Gender task (Accuracy)**
Single-task CNN	0.7693	0.8978
Multi-task CNN	0.7546	0.9020

We can observe how similar results are achieved with both the training and validation set. It is important to note that the true measure of the performance of the models will be revealed in the testing set and we will then analize if an architecture has an advantage over the other.

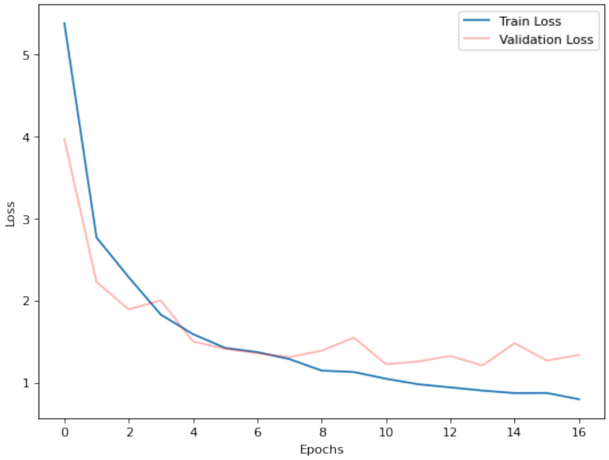

Fig. 7: Loss trend in the single task for age

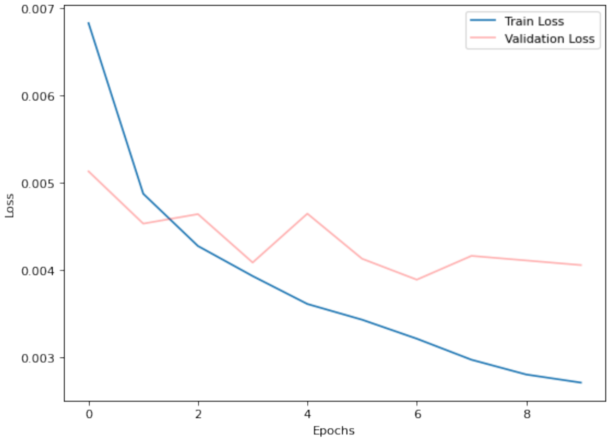

Fig. 8: Loss trend in the single task for gender

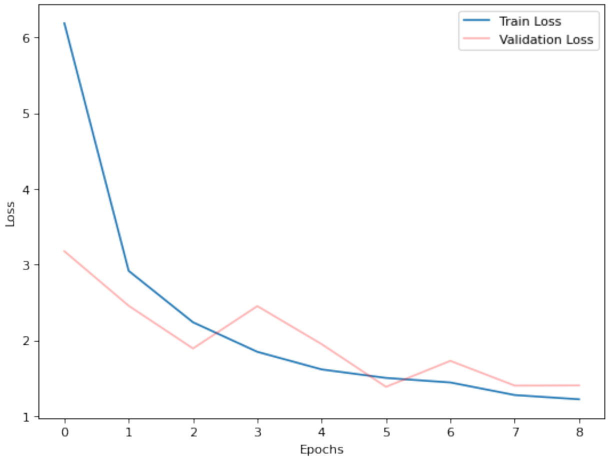

Fig. 9: Loss trend in the multi task

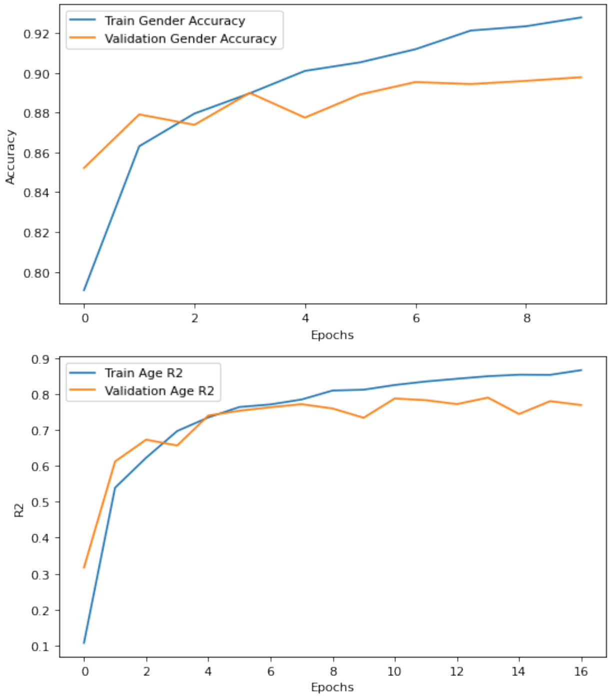

Fig. 10: Accuracy trend in the single task CNNs

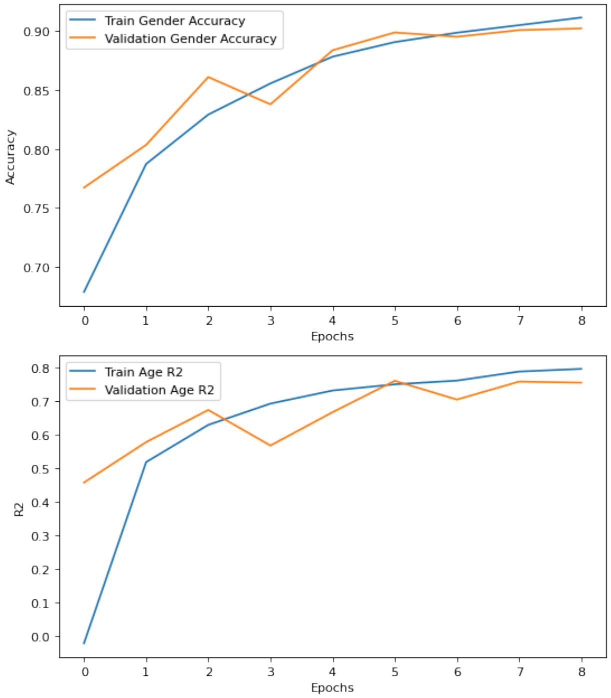

Fig. 11: Accuracy trend in the multi task CNN

C. Training Result

After completing the model training, we can now transition to the testing phase. During this phase, we load the testing images, which were initially set aside without undergoing any preprocessing, and evaluate the capabilities of the CNNs in explaining that data. Below, you can find the table of accuracy and R^2 results for both architectures:

Test Metrics	**Age task ($\mathbf{R^2}$)**	**Gender task (Accuracy)**
Single-task CNN	0.7832	0.8965
Multi-task CNN	0.7731	0.8986

along with Fig. 12 and Fig. 13 illustrating the confusion matrices for single-task and multi-task classification, respectively. To provide a visual example of the performance of the regressor, we present, in Fig. 14 and Fig. 15, a graph depicting the predicted age versus the ground truth for both the single-task and multi-task scenarios for the first 50 observations of the set.

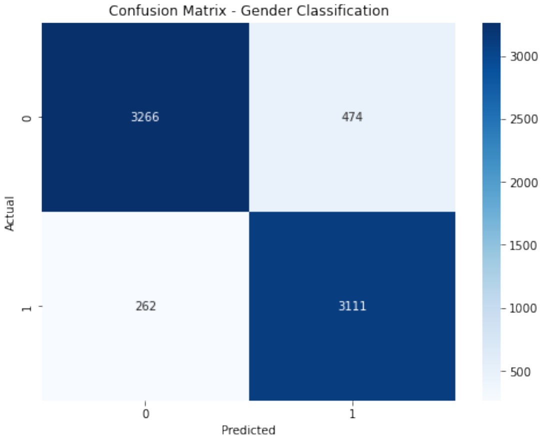

Fig. 12: Confusion Matrix for the single task

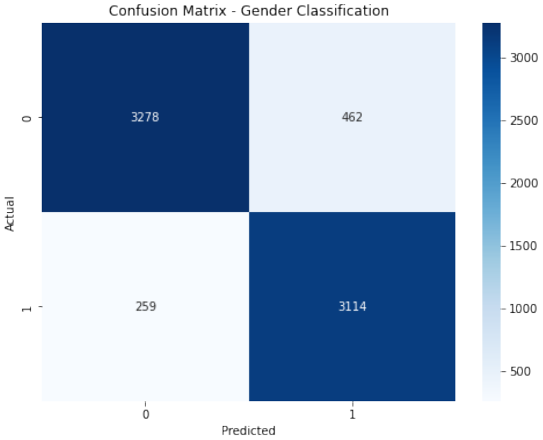

Fig. 13: Confusion Matrix for the multi task

In the context of the provided example image in Fig. 16, we can observe the corresponding mean attention heatmaps for each convolutional layer in the neural network. Specifically, Fig. 17 illustrates the single-task attention heatmap for age, while Fig. 18 depicts the single-task attention heatmap for gender and Fig. 19 showcases the multi-task attention heatmap. Notably, we can observe that the attention maps in the multi-task scenario appear to be a weighted average, with a notable bias towards the age-related features, as compared to the two individual single-task scenarios.

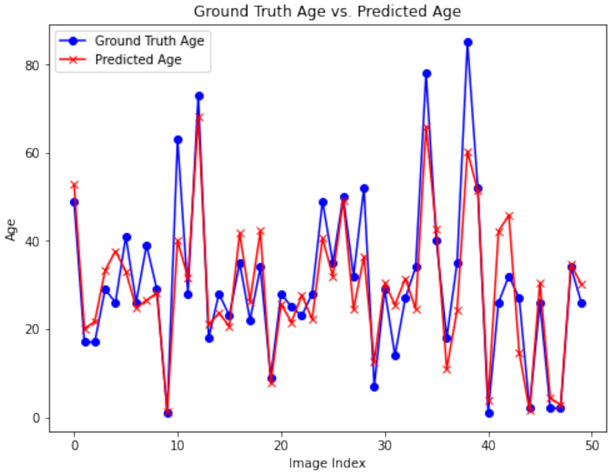

Fig. 14: Regressor performance example for the single task

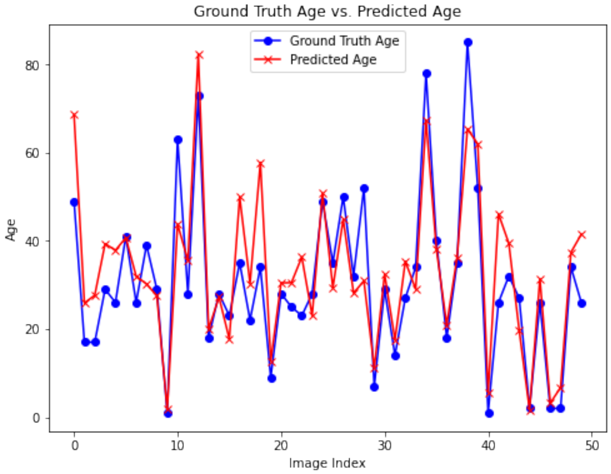

Fig. 15: Regressor performance example for the multi task

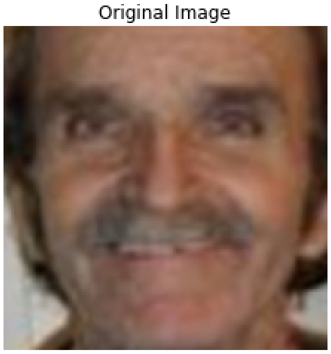

Fig. 16: Sample image from the validation dataset

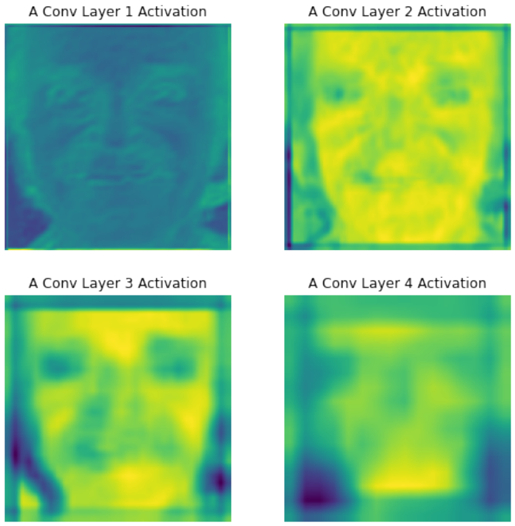

Fig. 17: Attention heatmap for the Age single-task CNN

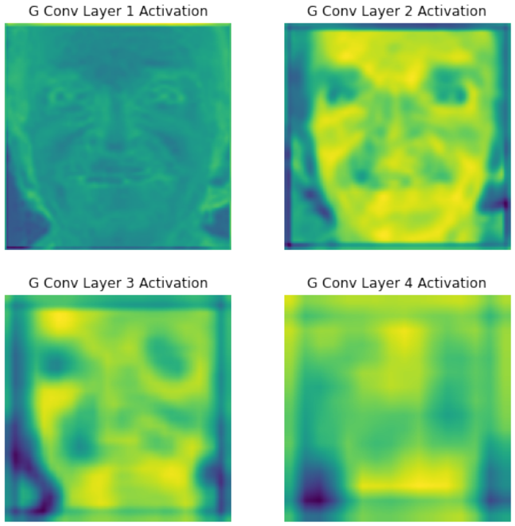

Fig. 18: Attention heatmap for the Gender single-task CNN

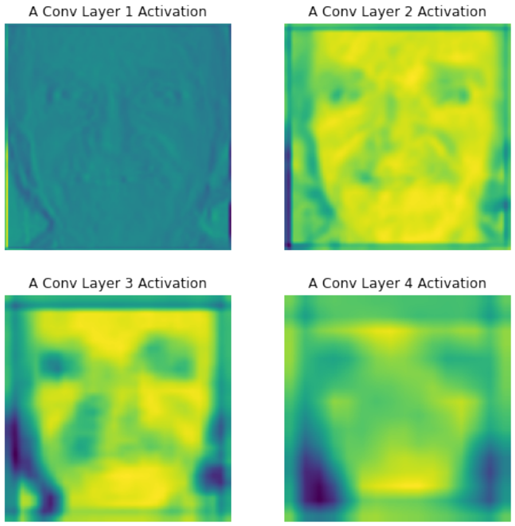

Fig. 19: Attention heatmap for the multi-task CNN

V. Conclusion

In this study, we have presented a comparison between two CNN architectures for the recognition of biometric characteristics, specifically age and gender, from facial images.

As we can observe, both through metrics and graphical representations, maintaining the same basic structure in both the single-task and multi-task architectures yields remarkably similar results. The distinction between the two models is largely attributed to the inherent randomness introduced during the model training process.

Nevertheless, it's important to note that the findings of this study are not universally applicable and may vary when altering the structure of the architectures, hyperparameters, or dataset.

References

[1] M. Fitzpatrick, "Advertising billboards use facial recognition to target shoppers," Sep 2010, accessed: 23/01/2024. [Online]. Available: https://www.theguardian.com/media/pda/2010/sep/27/advertising-billboards-facial-recognition-japan

[2] L. Chen, "Gender discriminating ads," Feb 2012, accessed: 23/01/2024. [Online]. Available: https://www.trendhunter.com/trends/facial-recognition-billboard

[3] N. Kobie, "The complicated truth about China's Social Credit System," Jun 2019, accessed: 23/01/2024. [Online]. Available: https://www.wired.co.uk/article/china-social-credit-system-explained

[4] Datagen, "Convolutional Neural Network: Benefits, Types, and Applications," May 2023, accessed: 23/01/2024. [Online]. Available: https://datagen.tech/guides/computer-vision/cnn-convolutional-neural-network/#

[5] N. Gladstone, "How facial recognition technology permeated everyday life," Sep 2018, accessed: 23/01/2024. [Online]. Available: https://www.cigionline.org/articles/how-facial-recognition-technology-permeated-everyday-life/

[6] European Union, "EU Artificial Intelligence Act," accessed: 23/01/2024. [Online]. Available: https://artificialintelligenceact.eu/the-act/

[7] L. Zorloni, "Abbiamo letto l'ultima versione dell'AI Act, la legge europea sull'intelligenza artificiale," Jan 2024, accessed: 23/01/2024. [Online]. Available: https://www.wired.it/article/ai-act-testo-ultima-versione-gennaio-divieti-riconoscimento-facciale

[8] I. Rafique, A. Hamid, S. Naseer, M. Asad, M. Awais, and T. Yasir, "Age and gender prediction using deep convolutional neural networks," pp. 1–6, 2019.

[9] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay, "Effective training of convolutional neural networks for face-based gender and age prediction," *Pattern Recognition*, vol. 72, pp. 15–26, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320317302534

[10] S. Subedi, "UTKFace Dataset," Aug 2018, accessed: 25/01/2024. [Online]. Available: https://www.kaggle.com/datasets/jangedoo/utkface-new

[11] J. Lopatecki, "Kolmogorov smirnov test: When and where to use it," Jan 2024, accessed: 23/01/2024. [Online]. Available: https://arize.com/blog-course/kolmogorov-smirnov-test/

[12] PyTorch and Keras, "Where should place dropout, batch normalization, and activation layer?" Oct 2023, accessed: 24/01/2024. [Online]. Available: https://androidkt.com/where-should-place-dropout-batch-normalization-and-activation-layer/