Transfer Learning: finetuning pre-trained networks to accurately predict age using faces dataset

Giovanni Mantovani and Andrea Palmieri

May 3, 2024

1 Introduction

The objective of this lab report is to explore the application of transfer learning for age prediction from facial images. Leveraging pre-trained MobileNet_v2 and ResNet18 architectures, we fine-tune models using the IMDb-crop dataset. Our goal is to assess the efficacy of transfer learning in enhancing age prediction accuracy.

2 Methodology

2.1 Metrics

The metrics taken into account to evaluate the models' performances are the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank Ordered Correlation Coefficient (SROCC).

PLCC (Pearson Linear Correlation Coefficient): Measures the strength and direction of a linear relationship between two variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, 0 instead indicates no correlation between the variables.

$$\text{PLCC} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

Where

- n is the number of data points.
- x_i and y_i are the individual data points.
- $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively.

SROCC (Spearman Rank Ordered Correlation Coefficient): As PLCC it evaluates the monotonic relationship between variables by comparing the ranked values of two variables. The SROCC has the same range of values as the PLCC [-1,1] where 0 indicates the absence of relationship between the variables.

$$\text{SROCC} = 1 - \frac{6\sum_i d_i^2}{n(n^2 - 1)} \tag{2}$$

Where:

- d_i is the difference between the ranks of corresponding pairs of data points.
- n is the number of data points.

2.2 Dataset

The dataset utilized is `IMDb-crop`, a subset of the IMDB-WIKI dataset, which comprises over 500,000 face images sourced from IMDb and Wikipedia profiles. Specifically, the dataset includes 460,723 face images from 20,284 celebrities sourced from IMDb and 62,328 face images from Wikipedia. These images were selected based on metadata such as date of birth, photo taken year, gender, and celebrity name. IMDb-crop contains only cropped faces of the IMDb dataset with 40% margin.
For more information about the dataset: https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/.
The images selected for training, validation and testing were filtered with the following criteria:

- `Presence of faces:` Only images containing faces were retained
- `Single faces:` Images with multiple faces were excluded
- `Face score threshold:` Only images with a face_score greater than 3.5 were kept
- `Age range:` Age between zero and one hundred, both included

After this filter process 71.828 were in the final dataset. Then the data partitioning was performed using 50% of the filtered dataset, with a 60/20/20 split.

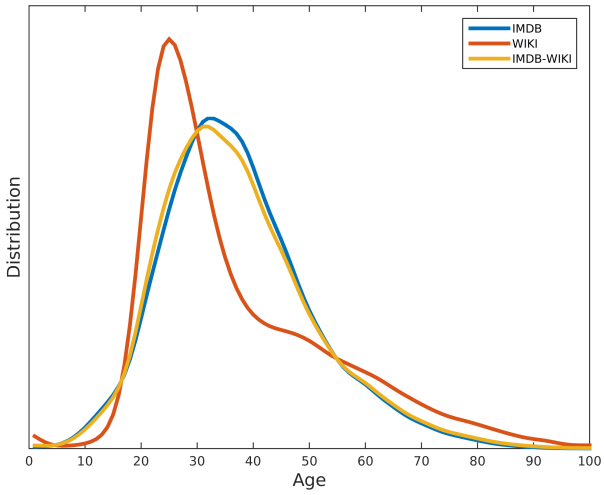

Figure 1: Age distribution of datasets

2.3 Models used for testing

We compare seven different models. The transfer learning approach was evaluated by comparing the PLCC and SROCC metrics of all the models on the test set. The models are the following:

- **Model 1:** MobileNet_v2 with unfrozen final classifier layers for finetuning and ImageNet1k_v1 weights.

  Architecture and Considerations:
  MobileNet_v2 employs an architecture based on inverse residual mapping, where the expansion step is followed by depthwise separable convolutions. In this configuration, only the final classifier layers were unfrozen for finetuning, while the rest of the network retained its pre-trained weights from the ImageNet1k_v1 dataset.

- **Model 2:** MobileNet_v2 fully unfrozen for training of the whole network.

  Architecture and Considerations:
  Unlike Model 1, in this setup, the entire MobileNet_v2 architecture was unfrozen and trained starting from ImageNet1k_v1 weights. By allowing all layers to update their weights, besides increasing the computational cost, the model can potentially learn more task-specific features related to the supervised learning task.

- **Model 3:** ResNet18 with unfrozen final classifier layers for finetuning and ImageNet1k_v1 weights.
  Architecture and Considerations:
  ResNet18 employs an architecture based on inverse residual mapping, where the expansion step is followed by depthwise separable convolutions. In this configuration, only the final classifier layers were unfrozen for finetuning, while the rest of the network retained its pre-trained weights from the ImageNet1k_v1 dataset.
- **Model 4:** ResNet18 with unfrozen final classifier layers and layer 4.
- **Model 5:** ResNet18 with unfrozen final classifier layers, layer 4 and layer 3.
- **Model 6:** ResNet18 with unfrozen final classifier layers, layer 4, layer 3 and layer 2.
- **Model 7:** ResNet18 with unfrozen final classifier layers, layer 4, layer 3, layer 2 and layer 1.

2.4 Training

Training was performed on the training set and for each model only the chosen layers were unfrozen by setting the `requires_grad` variable to True. This approach enabled the models to retain the learned representations from the source domain (ImageNet1k_v1 weights) while adapting to the target task of age prediction. By unfreezing progressively the layers, the aim was to preserve the high-level features learned by the networks while facilitating the learning of task-specific features from the target dataset.

3 Results

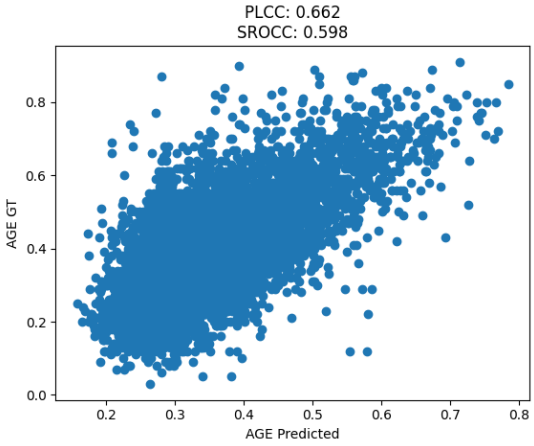

Figure 2: MobileNet_v2 with unfrozen final classifier layers

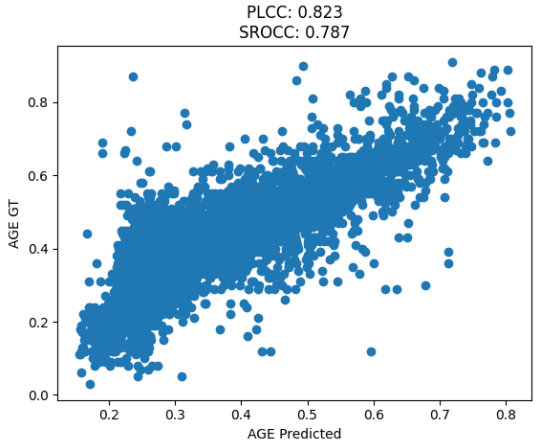

Figure 3: MobileNet_v2 fully unfrozen

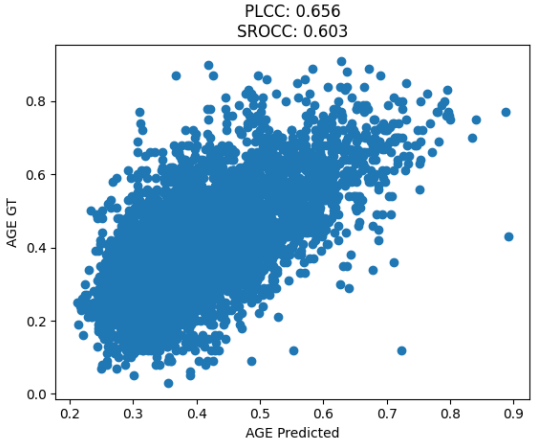

Figure 4: ResNet18 with unfrozen final classifier layers

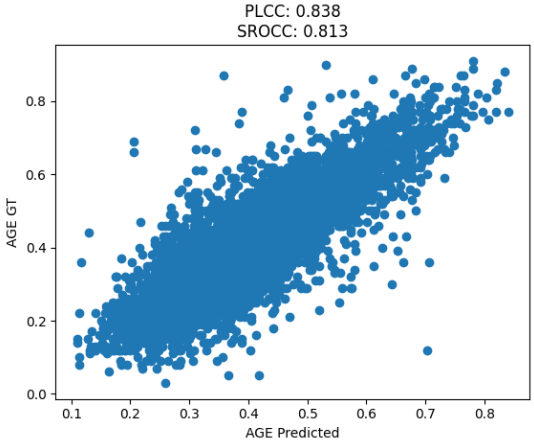

Figure 5: ResNet18 with unfrozen final classifiers layers and layer 4

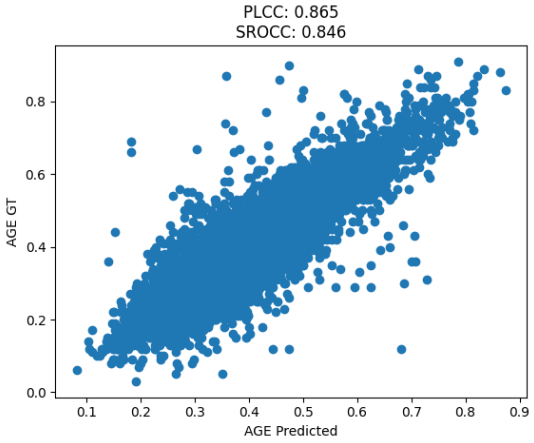

Figure 6: ResNet18 with unfrozen final classifiers layers, layer 4 and layer 3

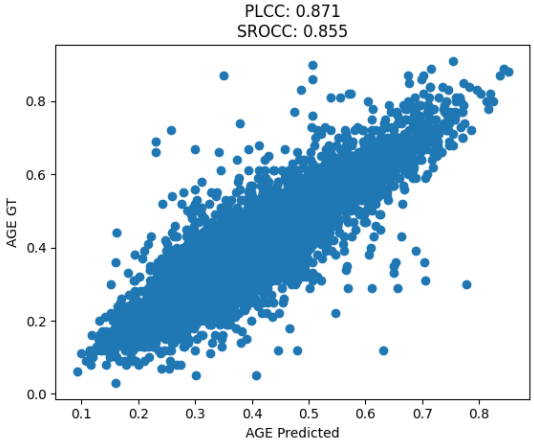

Figure 7: ResNet18 with unfrozen final classifiers layers, layer 4, layer 3 and layer 2

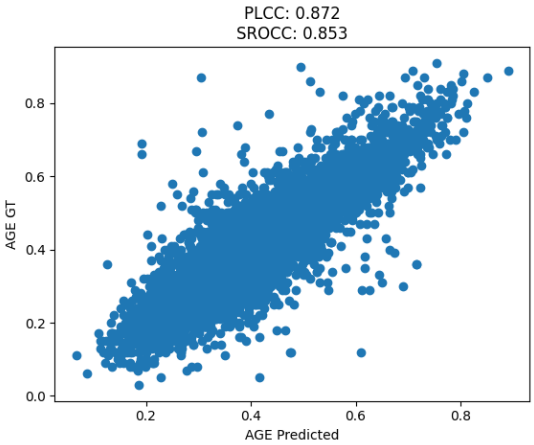

Figure 8: ResNet18 with unfrozen final classifiers layers, layer 4, layer 3, layer 2 and layer 1

Model	PLCC	SROCC
MobileNet_v2 only final layers	0.662	0.598
MobileNet_v2 fully unfrozen	0.823	0.787
ResNet18 only final layers	0.656	0.603
ResNet18 final layers and L4	0.838	0.813
ResNet18 final layers, L4 and L3	0.865	0.846
ResNet18 final layers, L4, L3 and L2	0.871	0.855
ResNet18 final layers, L4, L3, L2 and L1	0.872	0.853

Table 1: PLCC and SROCC metrics

4 Interpretation of results

The results presented in Table 1 provide insights into the performance of different models based on the PLCC (Pearson Linear Correlation Coefficient) and SROCC (Spearman Rank-Order Correlation Coefficient) metrics. Here's the interpretation of these results:

- `MobileNet_v2 only final layers:` This model achieved a PLCC of 0.662 and a SROCC of 0.598. These values indicate a moderate level of correlation between the predicted apparent age and the ground truth age labels.
- `MobileNet_v2 fully unfrozen:` The PLCC significantly improves to 0.823, and the SROCC also increases to 0.787. By training the entire MobileNet_v2 architecture from scratch, the model demonstrates enhanced performance, suggesting that fine-tuning all layers leads to better age estimation.
- `ResNet18 only final layers:` Similar to MobileNet_v2 with only the final layers unfrozen, this model achieves a PLCC of 0.656 and a SROCC of 0.603. Despite using a different architecture (ResNet18), the performance is comparable to MobileNet_v2 in this configuration.
- `ResNet18 final layers and L4:` The PLCC and SROCC notably increase to 0.838 and 0.813, respectively, by adding layer 4 to the training. This increase of about 20% suggests that features extracted from layer 4 contribute significantly to improving age estimation accuracy.
- `ResNet18 final layers, L4, and L3:` Further improvement is observed with the addition of layer 3, resulting in a PLCC of 0.865 and a SROCC of 0.846. This indicates that features from layer 3 also play a crucial role in enhancing age prediction performance.
- `ResNet18 final layers, L4, L3, and L2:` Adding layer 2 leads to a slight improvement in both PLCC (0.871) and SROCC (0.855). This indicates that features from layer 2 contribute marginally to the overall performance.
- `ResNet18 final layers, L4, L3, L2, and L1:` The performance remains consistent with the addition of layer 1, with a PLCC of 0.872 and a SROCC of 0.853. This suggests that layer 1 may not significantly impact the model's age estimation accuracy beyond the contributions of the previous layers.

5 Conclusion

The results highlight the effectiveness of fine-tuning the entire MobileNet_v2 network to significantly improve model performance compared to only fine-tuning the final layers. Leveraging the entire architecture facilitates better feature extraction tailored to the age prediction task. Furthermore, in the case of ResNet18, adding more layers for training progressively enhances performance, particularly with the inclusion of Layer 4. However, beyond Layer 3, the marginal improvements in performance suggest that training additional layers may not significantly improve the model's accuracy.