AgeEstimAI Project Report

Yannik Gaebel

December 11, 2023

1 Introduction

In this deep learning project a model is developed to estimate a persons age from a face image. Age estimation (AE) has been a subject of research for many years. Over time, various features have been investigated for their relevance in age prediction, ranging from activity and blood data to medical imaging outputs. Historically, the predominant methods involved manual feature extraction combined with traditional machine learning techniques. However, newer approaches often shifted to deep learning. One prominent application of deep learning in this domain is the use of facial images for age estimation.

The UTKFace dataset, detailed in Chapter 2, was chosen for training the model. TensorFlow was used as the primary framework for implementing the models. The initial model implemented was EfficientNetB0, with hyperparameter tuning covering optimizers, loss functions, batch sizes, learning rate schedules, augmentation, and model scaling, as described in Chapter 4.

After optimizing the EfficientNet model, several strategies were considered to improve the performance. These included (1) Extending the training dataset with other available face datasets or using pre-training for transfer learning. A common approach is performing face recognition as a pre-training task. These approaches are well deonstrated to improve a models performance. (2) Doing more systematic hyperparameter tuning and experimenting with different augmentation techniques. This option is very computationally intensive. (3) Splitting the dataset based on specific demographic criteria, such as gender or ethnicity, and developing seperate models for each subgroup. (4) Implementing more advanced model architectures.

In this work I chose to focus on reimplementing a novel architecture that yielded a significant improvement in age estimation performance on the MORPH2 dataset, reducing the mean absolute error (MAE) from 1.97 years to 1.3 years. The architecture combines the generation of multiple augmented versions of a single image, processing each through a convolutional neural network and then aggregating the image embeddings via transformer-encoders. This architecture is described in more details in section 3.

2 Dataset

The UTKFace dataset, a common resource in the field of age and gender prediction, was selected for this project. Unlike certain datasets that require formal application

procedures, UTKFace can be downloaded easily from the internet. The dataset contains over 20,000 annotated facial images. It encompasses a broad spectrum of ages, ranging from 1 to 116 years. A strengths of the UTKFace dataset is its diversity in terms of a wide array of facial expressions, poses, and resolutions, offering a realistic and challenging dataset for age estimation algorithms.

The age distribution is represented in Figure 1.

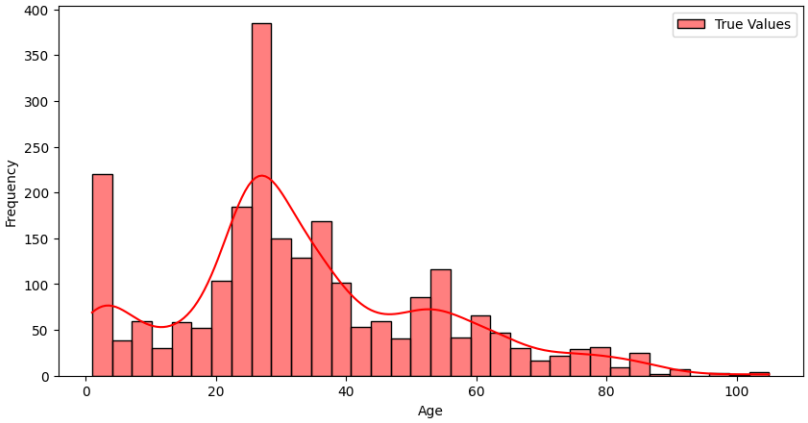

Figure 1: Age Distribution of the UTKFace Dataset

In alignment with other research a subset of the UTKFace dataset was used, focusing on the age range from 21 to 60 years. The dataset was split following an 80/20 train/test ratio totalling 16,431 images with 13,144 training and 3,287 test images. These decisions were made to ensure comparability with existing studies in the field.

The metric that is used is mean absolute error (MAE) in years of age. The state-of-the-art results in age estimation on the UTKFace dataset are 3.7 years when including additional training data and 4.23 years without additional data. Since no additional datasets are used in this project the relevant benchmark is 4.23 years.

3 Deep Learning Architectures

Initially I implemented the EfficientNetB0 model, known for a balance of accuracy and efficiency. To further enhance performance, I researched recent papers that showcase state-of-the-art (SOTA) results in age estimation using facial images on different datasets. A diverse range of different architectures can be found.

The MiVOLO architecture, introduced by Kuprashevich and Tolstykh (2023) integrates features extracted from the entire image with features from the face image to simultaneously predict age and gender. Employing the latest vision transformer technology, MiVOLO achieves SOTA results across various open-source datasets.

Shin et al. (2022) presents the moving window regression (MWR) algorithm. MWR iteratively refines age estimates by using reference images within a defined window. This method calculates a comparative value to determine if the subject image is closer to the younger or older reference image, offering a dynamic approach to age estimation.

The work of Hiba and Keller (2023) used two interesting techniques. Firstly, they propose an innovative transformer-based method for face image embedding, enhancing

the process through augmentation and embedding aggregation. Secondly, they introduce a hierarchical probabilistic age estimation scheme, optimizing the weighting of ensemble local age regressors' results. This paper marked significant improvements in SOTA results on two open-source datasets.

Given these advancements, I decided to implement the transformer-based augmentation and embedding aggregation approach, building upon the EfficientNet model I had previously implemented. My decision was also influenced by the observation that augmentation significantly boosts performance.

In the following chapters the EfficientNet and the Transformer-Encoder embedding aggregation model are described in more detail.

3.1 EfficientNet

The EfficientNet model family introduced the principle of compound scaling. The size of the models starting with EfficientNet-B0 to EfficientNet-B7 is scaled up systematically. The number of layers, number of channels and input image size are scaled using a compound coefficient. Using a bigger model might yield better performance, but comes with increased computationnal costs and the risk of overfitting. Which model size is optimal for the given task can be determined experimentally.

I first used the EfficientNetB0 model pretrained on the ImageNet dataset. The top layer for classification was replaced with a global average pooling layer to reduce spatial dimensions and two dense layers, with the last one being the regression output. In this way the architecture is tailored to predict a single continuous value, making it suitable for age estimation.

3.2 Transformer-Encoder embedding aggregation

This subsection delves into the architecture inspired by Hiba and Keller (2023), which merges augmentation techniques, CNN, and a transformer component to enhance image embedding capabilities.

The main idea of this approach is similar to test time augmentation (TTA), a technique where predictions for various augmented versions of the same image are combined to improve accuracy. The method introduced by Hiba and Keller (2023) extends this concept further. It utilizes a CNN backbone to generate embeddings for each augmented image version, and then employs a transformer-encoder for effective aggregation of these embeddings into a single, more informative representation, followed by a final regression layer for age prediction. The idea of the transformer encoder stack is to capture complex relationships before making a prediction.

4 Training and Hyperparameter Optimization

The training and optimization of the models involved a two-step process: firstly, optimizing the EfficientNet model, and secondly, employing the optimized EfficientNet as a backbone for the advanced transformer-encoder embedding aggregation architecture.

4.1 Training EfficientNet model

The training was conducted using Google Colab's T4 GPU with additional RAM, with each training session typically lasting between 20 and 40 minutes. The total cost for the entire training and hyperparameter optimization process was 20 euros.

My approach to optimization was focusing on one specific characteristic at a time and observing its impact on the test set loss, subsequently adopting the most effective setting for each characteristic.

The initial challenge was an unstable training process, indicated by fluctuating validation loss. To address this, I implemented an exponential learning rate decay schedule, which stabilized the training and improved results.

Initially, 20 percent of the data was used for validation. However, reducing this to 10 percent led to better test set performance. When attempting to scale up the model using EfficientNetB1-3, I encountered quicker overfitting without performance improvements.

Regarding batch size, I found that its impact on test set performance was minimal, so I settled on a medium size of 64. In terms of optimizers, Adam yielded the best results, outperforming RMSprop. Initially, I used mean squared error as the loss function, but switching to HuberLoss, which is more robust with outliers, proved more effective. The most impactful change came from implementing augmentation. I experimented with random adjustments in rotation, contrast, zoom, brightness, and translation.

4.2 Training Transformer-Encoder embedding aggregation model

Training the Transformer-Encoder Embedding Aggregation model proved to be a complex task. When attempting to run the model end-to-end, I encountered issues such as diverging validation loss or a notable drop in performance compared to baseline models.

To address training challenges, I adopted a strategy using pre-trained weights from the best-performing EfficientNetB0 model, keeping these weights frozen during training. This maintained the model's learned features while incorporating the new transformer-encoder architecture. To streamline the training, I simplified the model by reducing the transformer-encoder layers and the number of image augmentations. This approach balanced the model's complexity with its manageability. These adjustments initially matched the performance of the original CNN model. However, attempts to increase complexity, either by adding more augmentations or more transformer encoder layers, led to quicker overfitting and marginally reduced performance.

5 Results

The metric of evaluation of age estimation models is typically the mean absolute error (MAE). The state-of-the-art (SOTA) performance on the UTKFace dataset without extra training data is currently standing at 4.2, including extra training data it is 3.7. Initially prior to any fine-tuning, the EfficientNetB0 model achieved an MAE of 5.9. Through fine-tuning, the best-performing model developed was an optimized EfficientNetB0, which improved the test set MAE to 5.2. This represents a substantial improvement of 0.7.

The implementation of the more advanced transformer-encoder embedding aggregation

model did not yield an improvement in performance.

6 Discussion

The final model achieved a mean average error (MAE) of 5.2, which is 1 year higher than the current state-of-the-art (SOTA) value of 4.2 in age estimation on the UTKFace dataset. This gap suggests potential areas for improvement, possibly through the exploration of more advanced architectures and extensive fine-tuning. One notable area for enhancement is data augmentation, which emerged as the most effective strategy in my experiments. However, in this project, augmentation was limited to only one additional image per original image. The UTKFace dataset presents unique challenges due to its high diversity, making it a more demanding dataset compared to others like MORPH2, where an MAE of 1.3 was achieved.

Contrary to initial expectations, the transformer-encoder embedding aggregation model did not surpass the performance of the EfficientNetB0 model. Several factors could contribute to this outcome. Firstly, the transformer-encoder part was not pre-trained, and such pre-training might have been beneficial. Additionally, the advanced architecture resulted in a larger model. My earlier observations indicated that scaling up the EfficientNet variants often led to overfitting and diminished performance, a trend similarly observed in the transformer-encoder model. This suggests that the size of the UTKFace dataset might not be sufficient to benefit from a larger model. Notably, the new SOTA results achieved by the transformer-encoder model were on the MORPH2 dataset, which is about three times larger. An ablation study showed that using 10 augmentations and 4 transformer-encoder layers only improved the MAE by 0.1 year, which is only a marginal increase.

There are a number of key insights from this project. First, even basic fine-tuning can significantly boost a model's performance. Second, it's essential to match the architecture with the dataset's size and nature, as this project demonstrated the importance of aligning model complexity with dataset characteristics.

An interesting finding was the limitation of TensorFlow's shuffle function in this setup. The age-based ordering of images in filenames initially led to seemingly superior results. However, introducing a more effective shuffle revealed a performance drop, suggesting the model might have been learning some pattern from the data's sequential order. This emphasizes the importance of proper data randomization to avoid unintentional bias.

Moreover, in pursuit of improved model performance, obtaining more data could have been a more viable strategy. Larger datasets often provide richer learning environments, enhancing a model's accuracy and ability to generalize. Thus, alongside optimizations in architecture and processes, the significance of expansive and varied datasets in boosting model performance is clear.

Lastly, for the goal of enhanced model performance, acquiring more data might have been a more effective strategy. Extensive and diverse datasets at least as important as architectural and procedural optimizations.

7 Work breakdown

My initial estimates of the workload were:

- Dataset preprocessing (8 hours)
- Designing and building the network (35 hours)
- Fine-Tuning (25 hours)

The most time-consuming aspects of this project were hyperparameter optimization, re-implementing the advanced architecture using TensorFlow, and identifying and resolving the shuffle function bug.

References

Hiba, S. and Keller, Y. (2023). Hierarchical attention-based age estimation and bias estimation.

Kuprashevich, M. and Tolstykh, I. (2023). Mivolo: Multi-input transformer for age and gender estimation.

Shin, N.-H., Lee, S.-H., and Kim, C.-S. (2022). Moving window regression: A novel approach to ordinal regression.