Age, Gender and Ethnicity Prediction

Arijit Saha, Aryan Mathe, Akshat Singh & Yash Ruhatiya

***Abstract*—The overarching objective of this project was to develop a robust machine learning model capable of accurately predicting age, gender, and ethnicity based on facial images. Anchored in Convolutional Neural Networks (CNNs), the base models form a robust foundation for subsequent optimization strategies. The approach involves integrating gender labels to enhance predictions for age and ethnicity, culminating in a unified model capable of simultaneous, accurate predictions for all three attributes. Further refining predictive performance, ensemble methods such as bagging and boosting are employed.**

I. Dataset

We used the UTKFaceDataset [1] The dataset consists of over 20,000 face images with annotations of age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc. The dataset has long age span in the range 0 to 116 years old, binary gender labels and the following five ethnicity labels: White, Black, Asian, Indian and Others.

We carefully preprocessed the dataset to transform raw images into a suitable format for model training. Utilizing the dlib library for precise face recognition, facial regions were accurately identified and resized to meet model input specifications, optimizing computational efficiency. Labels for age, gender, and ethnicity were extracted from filenames, ensuring consistent and accurate annotation.

Fig. 1: An sample image before and after preprocessing

II. Models

A. Base Models

We developed base models for the prediction of Age, Gender and ethnicity. These models took image as input and predicted one of the respective traits regarding the image.

1) Gender Prediction: For the Gender Prediction Model, we harnessed the power of neural network architecture to achieve accurate and nuanced results. The architecture was meticulously crafted, featuring a sequence of essential layers such as Conv2D for convolutional operations, MaxPooling2D for down-sampling, and Dense layers for classification. The neural network design comprised three consecutive Conv2D layers, strategically employed to extract intricate patterns and features from facial images. To ensure a robust binary prediction of male or female, the sigmoid activation function was applied at the final layer. This activation function, tailored for binary classification tasks, provided a seamless transition from the convolutional layers to the output layer. For optimization and training, we utilized the Adam optimizer. The binary cross-entropy loss function was chosen to quantify the disparity between predicted and actual gender labels, guiding the model towards accurate predictions.

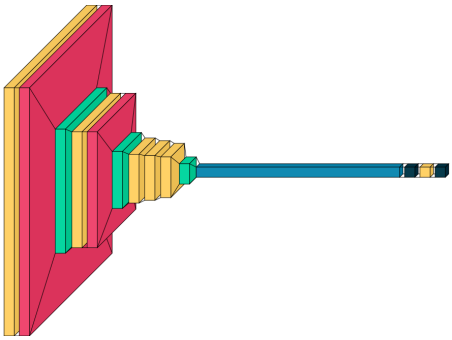


Fig. 2: Visual representation of the Gender Prediction Architecture

2) Ethnicity Prediction: For Ethnicity Prediction, we used a similar model to the one we used for Gender Prediction. Instead of having 3 consecutive Conv2D layers, we used 2 consecutive Conv2D layers, which resulted in a more accurate and robust prediction. We used the sparse categorical cross to calculate the loss on the vector of probability estimates, as Ethnicity classification is a multi-class classification.

3) Age Prediction: Age Prediction used a similar model as the above 2 tasks, the difference being it used a single Conv2D layer instead of multiple consecutive layers. As Age Prediction is essentially a regression task, we used Mean Absolute Error as the loss function, which is suitable for regression tasks.

B. Context-aware Models

To improve our prediction and increase accuracy, we employed context-awareness to models. While predicting Age, we provided the architecture with the gender. This led to an increase in the efficiency of the model as features corresponding to a particular gender at a certain age are different from the other gender. Hence, when gender is provided to the model, it utilises the features in a better way to make a more accurate prediction. The intuition for the above is that ageing in male and female have different effects on their facial feature at different ages. So, providing the gender helps the model classify them better.

Extending the context-aware approach to ethnicity prediction, the inclusion of gender information proved to be a valuable enhancement. By providing gender as contextual information, the model adapts its utilization of facial features,

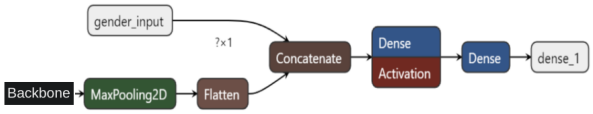


Fig. 3: Visual representation of the Age Given Gender Prediction Archietecture

recognizing gender-specific patterns. This refinement significantly contributes to the accuracy of ethnicity predictions, as facial features exhibit distinct characteristics for different ethnicities within specific genders.

C. Combined Model

To streamline our predictions and enhance overall efficiency, we devised a unified model capable of simultaneous Age, Gender, and Ethnicity prediction. This innovative approach leverages the dense feature vectors generated during gender prediction to augment the predictive capabilities for age and ethnicity. The initial architecture is shared across all predictions, encompassing Conv2D layers for convolutional operations, MaxPooling2D for down-sampling, and Batch Normalization layers to ensure stability and faster convergence. Following this common base, the model diverges to address gender prediction. The shared architecture is flattened and Dense layers are applied, creating a dense feature vector tailored for gender prediction. This vector encapsulates richer information about the image, acting as a comprehensive enhancement for subsequent predictions. This dense feature vector is concatenated with the flattened layer, seamlessly integrating gender-specific features into the model's understanding. This concatenated information is then utilized for predicting both age and ethnicity concurrently. This design choice is grounded in the recognition that gender prediction, with its inherently dense feature vector, encapsulates valuable insights that transcend the singular prediction domain. By combining these insights with the flattened layer, our model benefits from an understanding of facial characteristics, leading to more robust and accurate predictions for age and ethnicity.

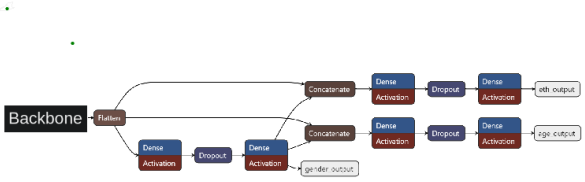


Fig. 4: Visual representation of the Combined Archietecture

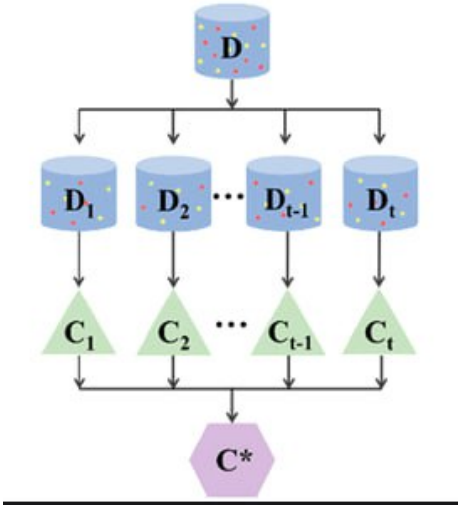


Fig. 5: Visual representation of the bagging algorithm

III. Ensemble Techniques

We employed ensemble techniques of bagging and boosting in pursuit of developing robust models with heightened accuracies.

A. Bagging

Our base models for age, gender and ethnicity are fairly complex by themselves and hence have a tendency to overfit to the training data. Bagging is useful in such scenarios as it reduces the overall variance of the model making it stable and hence helps generalize well. The bagging algorithm:

- Randomly select subsets (with replacement) from the training dataset. Each subset is of the same size as the original dataset,
- Train a base model independently on each bootstrap sample.
- Predict using either majority vote or averaging depending on nature of the task. For our model, we use averaging for age, majority vote for gender and ethnicity.

B. Boosting

In our pursuit of constructing an optimal predictive model, we initially considered boosting with our well-performing base Convolutional Neural Network (CNN) models. However, recognizing the inherent strength of these models, we pivoted to a more fitting approach. Given that both the gender and ethnicity prediction tasks are fundamentally classification problems, we employed Decision Trees (DTs) as our base classifiers for the boosting framework.

It was not plausible to directly pass images as input to DTs. To overcome this, we adopted a feature extraction strategy. For each image, we extracted a set of features (128/256/512 in number) and employed these feature sets as inputs to the boosting classifier. This innovative approach allowed us to seamlessly integrate boosting into our workflow. The next two subsection describe the techniques involved in detail.

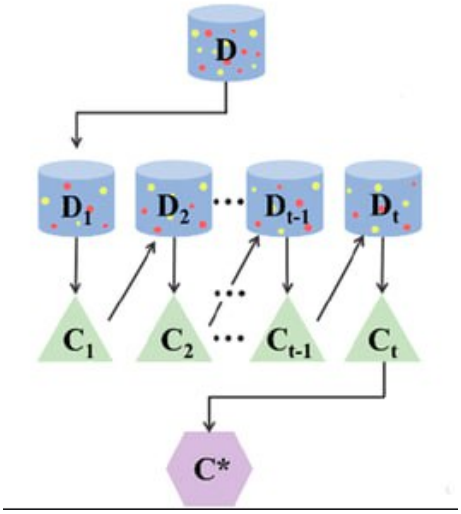


Fig. 6: Visual representation of the boosting algorithm

C. Feature Extraction

For getting features out of the facial images, we had 2 options: (i) Using the 68 facial landmarks or (ii) Extract features from an intermediate layer of the CNN model trained individually for the particular task of gender or ethnicity prediction. We went with the latter one as that would really be adapted to the task at hand and hence is expected to give better results.

We used a pretrained VGG16 model for getting the required features. By freezing all the layers of the VGG16 model, we retained the learned features from its extensive training on diverse image datasets. To tailor the pretrained VGG16 model to our specific prediction tasks, we augmented it with additional dense layers. This adaptation allowed the model to specialize in age or ethnicity prediction, leveraging the high-level features extracted by the pretrained layers.

The freezing of the initial layers ensured that the model retained its ability to recognize general image features, while the added dense layers enabled task-specific adaptation. We extracted features from an intermediate layer of the modified VGG16 model. This extraction process provided us with rich, abstract representations of facial attributes, serving as valuable inputs for subsequent stages of our predictive pipeline i.e the decision tree classifiers.

D. Boosting Algorithms

1) Gender Prediction with AdaBoost: For the task of gender prediction, we implemented the AdaBoost algorithm, a standard boosting technique, using a Decision Tree as the base classifier. After tuning, we set the max_depth parameter of the Decision Tree to 3. The base classifiers, despite their simplicity, contributed effectively to the boosted ensemble, enhancing gender prediction accuracy.

2) Ethnicity Prediction with SAMME Algorithm: For ethnicity prediction, which involves multiclass classification, we employed the SAMME (Stagewise Additive Modeling using a Multi-class Exponential loss) algorithm [2]. Again we utilized a Decision Tree with max_depth = 3 as the base classifier.

*3) **Algorithm.** SAMME:*

i Initialize the observation weights w_i = 1/n, i = 1, 2,...,n.

ii For m = 1 to M:

a Fit a classifier $T^{(m)}(x)$ to the training data using weights w_i.

b Compute error for the current estimator

$$err^{(m)} = \sum_{i=1}^{n} w_i \mathbb{I}(c_i \neq T^{(m)}(x_i)) / \sum_{i=1}^{n} w_i \quad (1)$$

c Compute weight for the current estimator

$$\alpha_{(m)} = \log \frac{1 - err^{(m)}}{err^{(m)}} + \log(K-1) \quad (2)$$

d Set sample weights

$$w_i \leftarrow w_i \cdot exp(\alpha_{(m)} \cdot \mathbb{I}(c_i \neq T^{(m)}(x_i))), \quad (3)$$

for i = 1, 2,...,n.

e Re-normalize w_i

iii Output class

$$C(\mathbf{x}) = arg\max_{k} \sum_{m=1}^{M} \alpha_{(m)} \cdot \mathbb{I}(T^{(m)}(x) = k) \quad (4)$$

where , K = number of classes

n = Total number of samples

M = Number of estimators

Note that when K=2, the SAMME algorithm reduces to AdaBoost.

IV. Experimentation

Our experimentation encompassed diverse architectural variations, involving the manipulation of kernel sizes, the number of convolution layers, dropout probabilities, and strides. Additionally, we explored different configurations for dense layers towards the end of the model. To enhance contextual awareness, we experimented with introducing information, such as gender, at various layers of the model. Within the combined model, we systematically adjusted the weights assigned to the loss or gradient of each attribute—age, gender, and ethnicity. Notably, the introduction of residual connections yielded performance improvements.

Furthermore, our exploration extended to ensemble techniques, specifically bagging and boosting. In boosting, we investigated varying the number of features, while for both bagging and boosting, we systematically adjusted the number of estimators. These comprehensive experiments allowed us to fine-tune our models, identifying optimal configurations.

V. Conclusion

Our models demonstrate notable accuracy in gender prediction and decent results in predicting ethnicity and age. The incorporation of contextual information yielded performance improvements, underscoring the distinct aging patterns between men and women. Additionally, our use of ensemble techniques provided the anticipated accuracy boost.

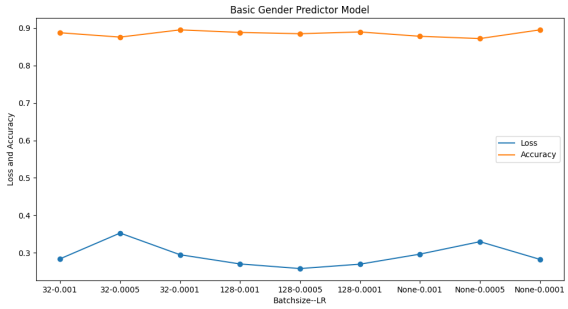
REFERENCES

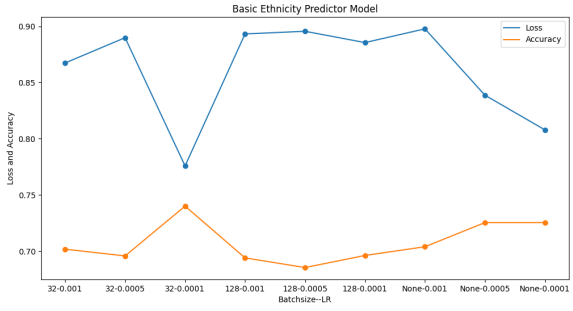
[1] UTKFace Large Scale Face Dataset, https://susanqq.github.io/UTKFace/
[2] Ji Zhu, Hui Zou, Saharon Rosset and Trevor Hastie (2009), Multi-class AdaBoost, Statistics and Its Interface Volume 2 349–360

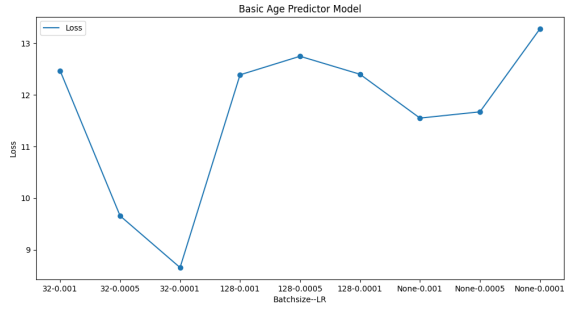
VI. RESULTS

The reported outcomes are based on training for a total of 8 epochs, utilizing the Adam optimizer with varying learning rates. Specifically, Binary Cross-Entropy (BCE) loss was employed for gender prediction, Sparse Categorical Cross-Entropy loss for ethnicity prediction, and Mean Absolute Error (MAE) for age prediction.

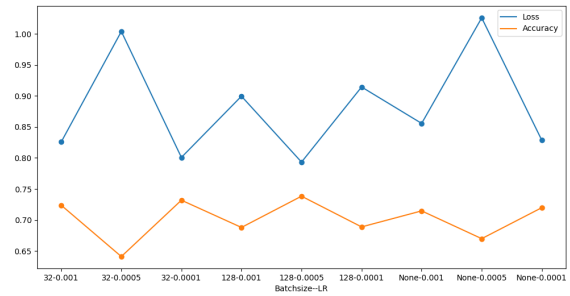
A. Base Models

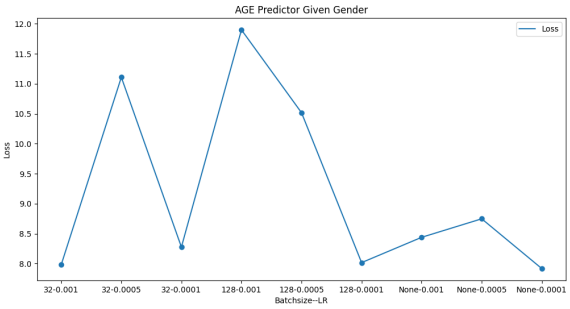




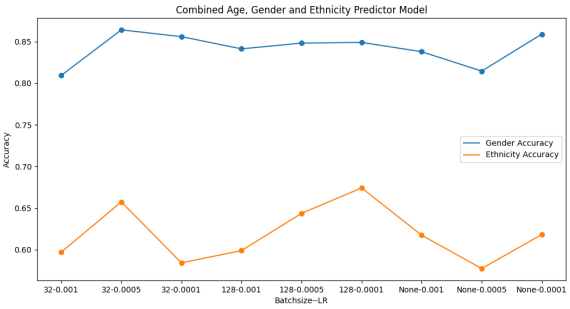


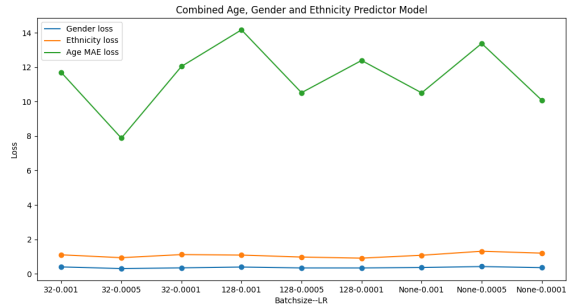
B. Context Aware Models



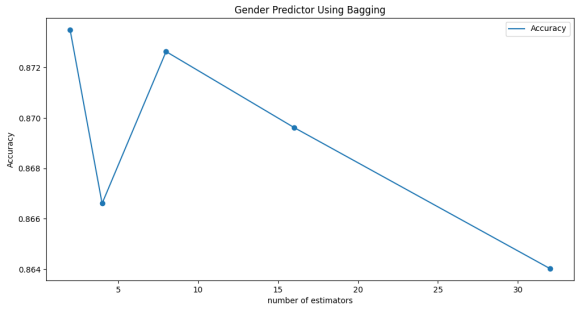


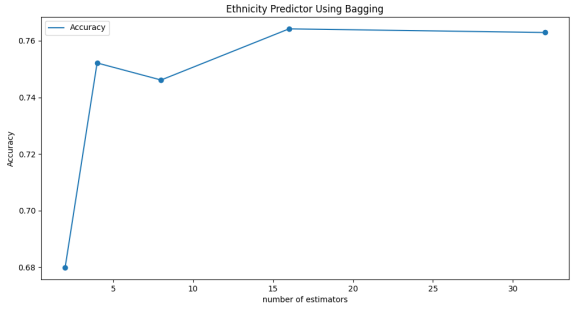
C. Combined Models

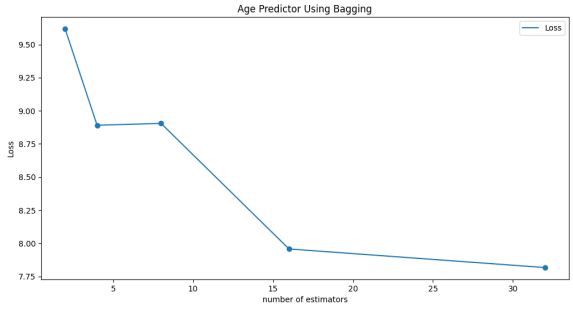




D. Bagging enabled model







E. Boosting enabled Models