**Age, Gender, and Race Classification with Multi-outputs Convolutional Neural Networks**

**Abstract**

This project built a multi-output deep convolutional neural to classify the age, gender, and race for each image included in the UTK Face dataset, reaching an accuracy of 91.22% for gender and 81.23% for race. The model in this project consists of 16 convolutional layers, 3 fully connected layers for each class (age, gender, and race), and a final 8-way softmax for age class, 2-way softmax for gender class, and 5-way softmax for race class. To reduce overfitting, this project 1) added max-pooling layer and batch normalization between successive convolutional layers and dropout layer before each fully connected layer, 2) applied data augmentation, and 3) early stopping.

**1 Introduction**

The development of artificial intelligence enables machines to learn capabilities of humans, one of the capabilities is to recognize images (under domain of computer vision). Convolutional Neural Networks which were first introduced by LeCun [1] are preferred and perfected over time for image recognitions.

In 2012, Krizhevsky et al. built an impressive model, which is widely considered as the most influential model, largely improved the accuracy on the ImageNet classification [2]. Krizhevsky et al. suggested AlexNet architecture which consisted of 5 convolutional layers, max-pooling layers, dropout layers, 3 fully connected layers, and a 1000-way softmax layer.

In 2014, Karen Simonyan and Andrew Zisserman discussed the importance of depth and simplicity in Convolutional Neural Networks [3]. They created a 19-layer Convolutional Neural Networks named VGG-19 with only effective receptive field which is different from using effective receptive field in the first layer in AlexNet. They concluded that simple and deep Convolutional Neural Networks produce better performance.

In 2015, Kaiming He et al. argued that simply stacking convolutional layers is hard to reach optimal results due to notorious vanishing / explosion gradient problem [4]. They introduced “identity-shortcut-connection” that jumps over one or more layers to avoid vanishing gradient problem.

This project aim to find the optimal model for UKTFace dataset image classification by applying the methods argued by Krizhevsky He et al., Karen Simonyan and Andrew Zisserman, and Kaiming He et al..

**2 The Dataset**

**2.1 Overview**

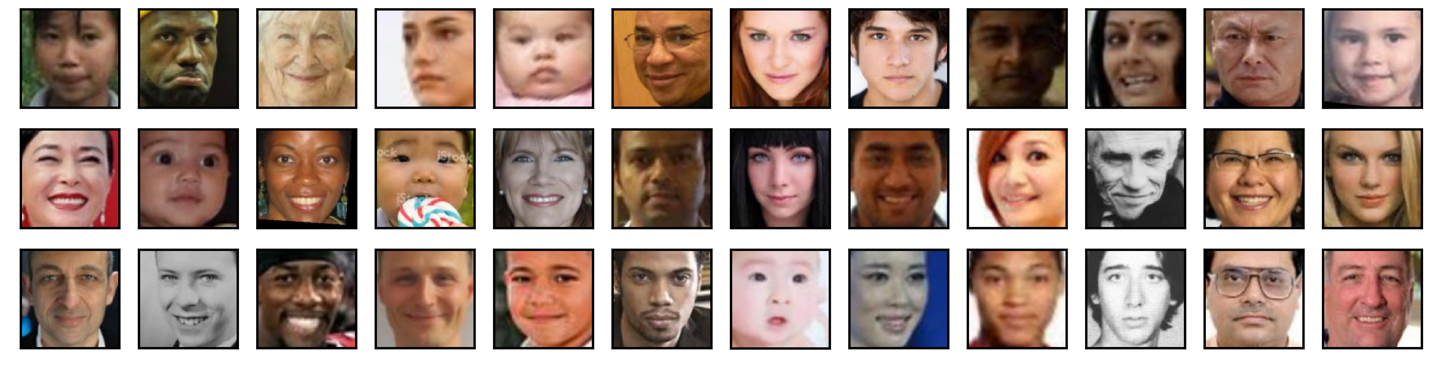


Fig.1 Face Image Samples from UTKFace Dataset

UTKFace dataset, which is released by Kaggle website, consists of over 20,000 face images with annotations of age, gender, and ethnicity. As shown in Fig.1, although the images are properly cropped and only contains the face region, there are variations in pose, facial expression, illumination. etc. among images, and thus this project thinks data augmentation is needed.

**2.1 Age**

Age label contains integers from 0 to 116.

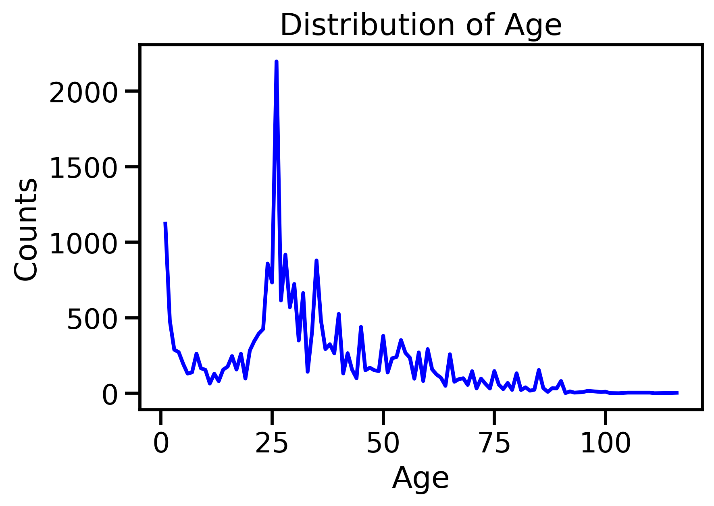


Fig.2 Distribution of Age in the UTKFace Dataset

I divided age labels into 8 groups: [0, 5], [6, 11], [12, 24], [25, 34]. [35, 44], [45, 59], [60, 79], [80, 116]. According to the age distribution displayed in Fig.2, [0-5] age group and [25, 34] may have the highest performances among the 8 age groups.

**2.2 Gender**

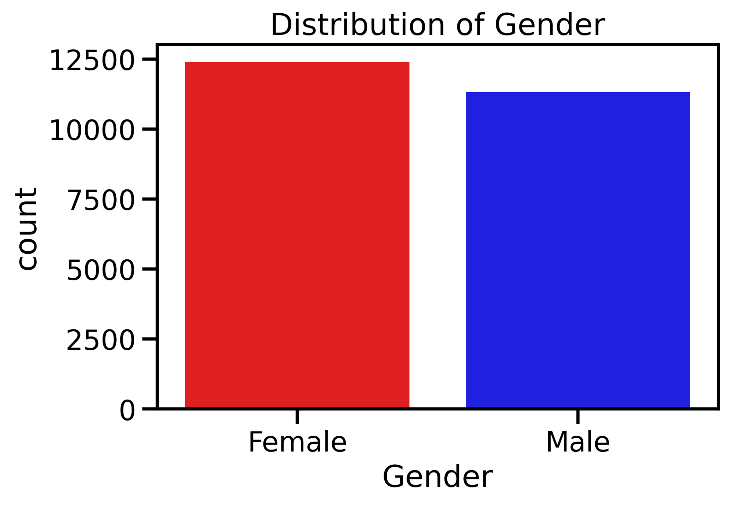


Fig.3 Distribution of Gender in UTKFace Dataset

Gender label contains 0s and 1s. 0 represents male and 1 represents female. According to Fig.3,, the distribution of gender is quiet even, indicating that we may have very high accuracy on predicting gender labels.

**2.3 Race**

**Chart, bar chart

Description automatically generated**

Fig.4 Distribution of Race in UTKFace Dataset

Race label contains integers from 0 to 4: 0 represents White, 1 represents Black, 2 represents Asian, 3 represents Indian, and 4 represents Others. From Fig.4, we can see that the proportion asian race is the largest and thus may have the highest accuracy.

**3 Image Processing and Model Training**

**3.1 Data Preprocessing**

**3.1.1 Image Resizing**

The image files are kept in vector image format with resized shape (original shape is . Age, gender, and gender class are one-hot encoded.

**3.1.2 Data Augmentation**

Data augmentation is a technique used frequently in deep learning that enlarge he size of training dataset and increase the amount of relevant features in the dataset by creating variations of the images, helping build more skillful models. Although deep leaning algorithms like Convolutional Neural Networks has invariance property which can classify objects regardless the objects are placed in different locations or orientations, data augmentation can boost the robustness of the model.[5]

Graphical user interface, text, application, chat or text message

Description automatically generated

Fig5. Augmentation Result Sample

As shown in Fig.1, variations exist in pose, facial expression, and illumination, etc. among the face data, easy to lead to overfitting problem, which can be reduced by applying data augmentation. In this project, the image data in the training set has been augmented by randomly rotating, shifting, horizontally flipping, and zooming, an example is shown in Fig.5.

**3.2 Strategies Applied for Reducing Overfitting**

**3.2.1 Max-pooling**

When the training dataset is not large enough to contain all the features in the whole dataset, overfitting happens. By adding max-pooling layer, the size of spatial size and the number of parameters will be reduced (et. only a subset of features which has the max value will be selected), as a result the model is less likely to learn false patterns.[6]

**3.2.2 Batch Normalization**

Regularization introduces additional information to the model and thus reduce overfitting problem [6]. Batch normalization is preferred by this project.

One of the problems encountered often when training Deep Neural Networks is internal covariate shift, which is caused by the different distribution of each layer’s inputs. Luckily, batch normalization can avoid the problem by reducing the amount of hidden unit values shift around. Also, each layer of the network can learn more independently from other layers [7].

**3.2.3 Dropout**

Introducing dropout method into Deep Neural Networks can avoid neuron interactions by ensuring that each feature is not always available and force the model to learn different potential patterns [8].

**3.2.4 Early Stopping**

One of the major challenges in training Deep Neural Network is deciding when to stop the training. With too short time of training, underfitting occurs, while with too long time of training, overfitting occurs. This project will also apply early stopping to reduce overfitting – stop training when performance on the validation dataset starts to degrade [9].

**3.2.5 Data Augmentation**

Data Augmentation reduces overfitting by enlarging the features in the training set and avoid the training model learning false patterns. Data Augmentation has been used widely and show effective results in image classification [2] [10].

**3.3 Model Architecture**

**3.3.1 Model 1**

**Diagram

Description automatically generated**

Fig.6 Model 1 Architecture

Model 1 share similar structures with AlexNet[2], it consists of 5 convolutional layers, 9 fully connected layers (3 fully connected layers for each class), and a final 8-way softmax for age class, 2-way softmax for gender class, and 5-way softmax for race class. Convolutional layers are followed by max-pooling layers. Instead of using Local Response Normalization after applying the ReLU nonlinearity in certain layers, model 1 use Batch Normalization. The architecture of model 1 is summarized in Fig.6.

**3.3.2 Model 2**

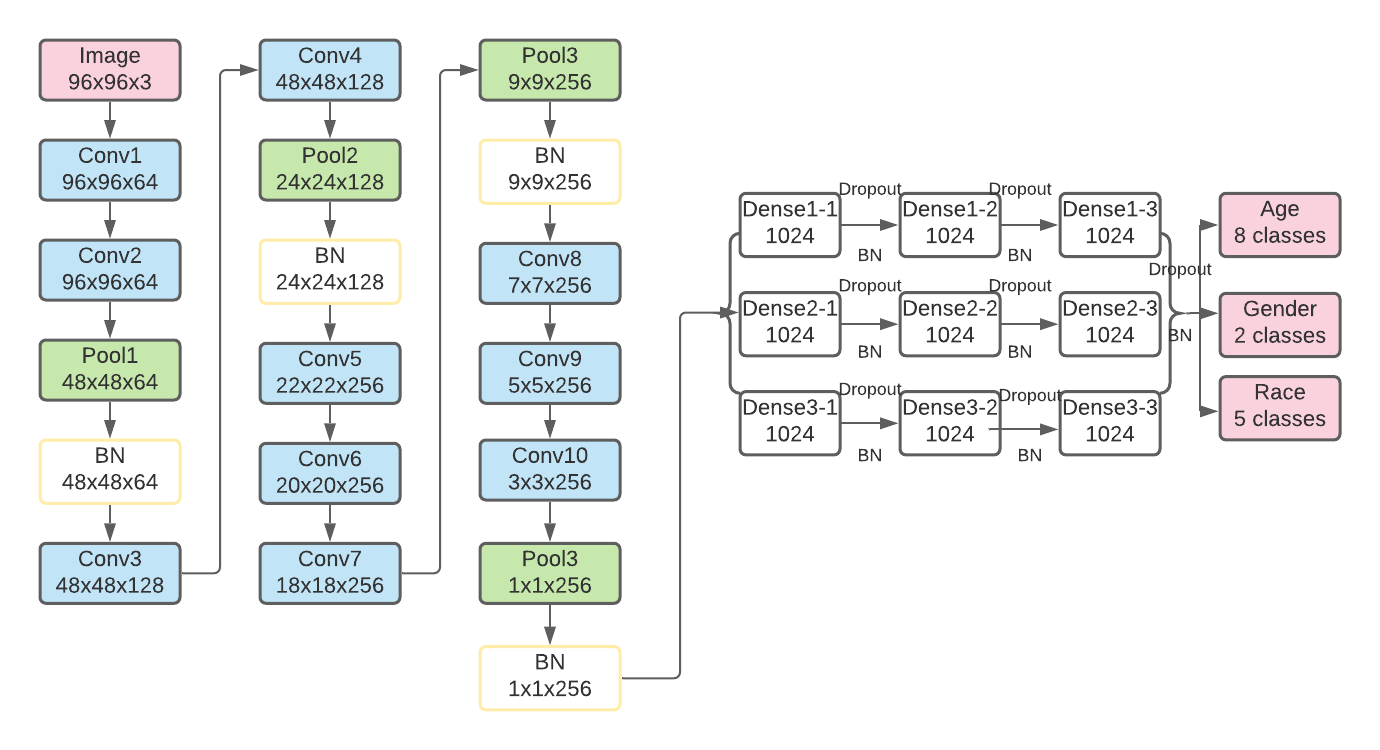


Fig.7 Model 2 Architecture

As argued by Karen Simonyan and Andrew Zisserman [3], simple and deep Convolutional Neural Networks perform better, Model 2 consists of 10 convolutional layers with effective receptive field only.

**3.3.3 Model 3**

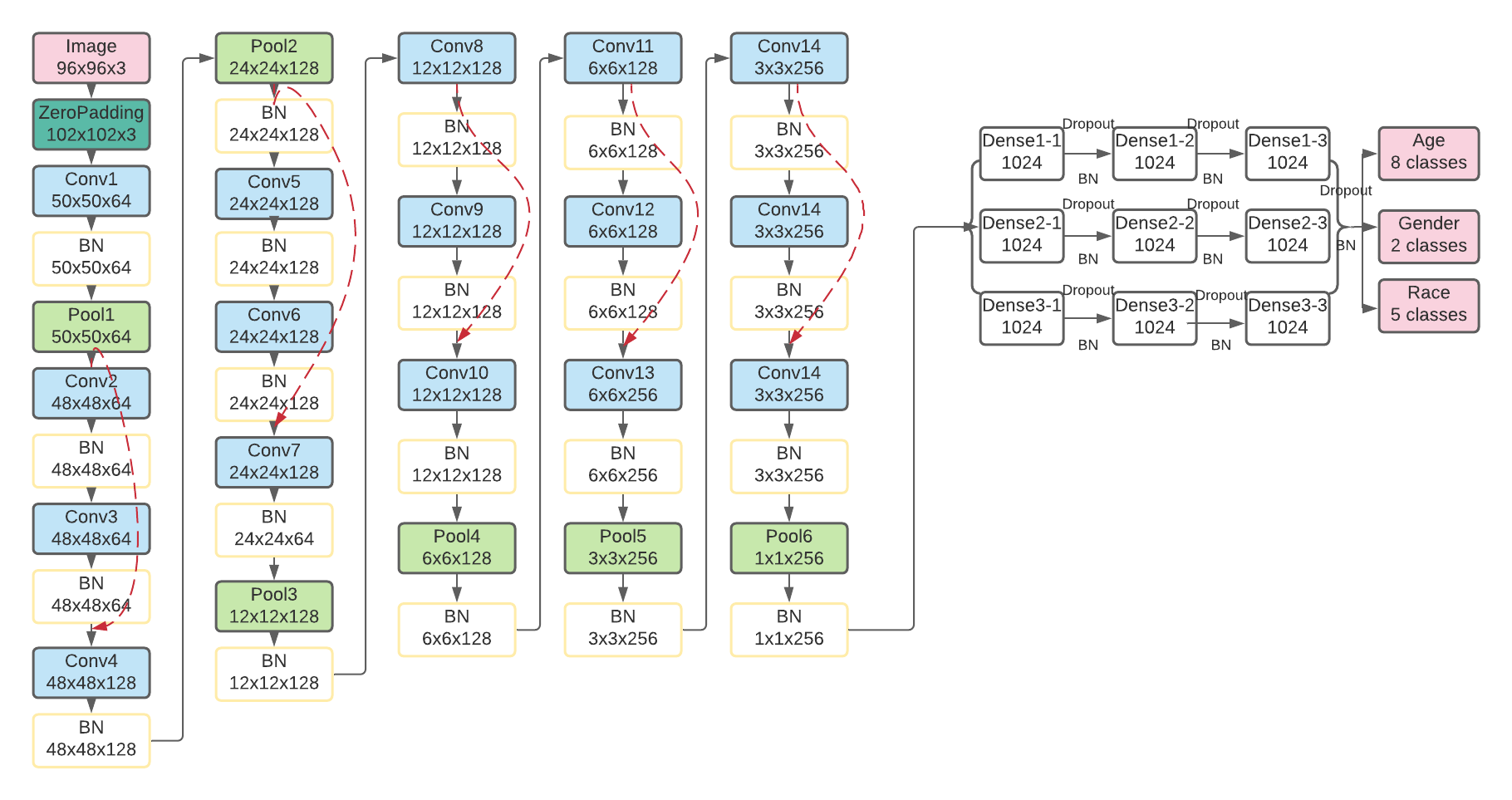


Fig.8 Model 3 Architecture

In model 3, more convolutional layers are added (16 convolutional layers) and the concept of residual learning [4] is applied to avoid vanishing gradient problem. The architecture of model 3 is summarized in Fig.8.

**3.4 Hyperparameter Tuning**

**3.4.1 Optimizers**

**Chart, histogram

Description automatically generated**

Fig.9 Loss Per Optimizer

Experiments were conducted in this project to choose the optimal optimizer among SGD, RMSprop, Adam, and Adamax with model 1 architecture. As shown in Fig.9, Adamax is the optimal optimizer with stable and excellent performance.

**3.4.2 Other Hyperparameters**

**Text

Description automatically generated**

**Table

Description automatically generated**

Fig.10 Keras Tuner Result

Other hyperparameters including number of filters, dropout rate, learning rate, and decay are tuned by Keras tuner which is a library for hyperparameter tunning in Tensorflow. The result is shown in Fig.10.

**3.5 Model Evaluation**

**3.5.1 Model 1**

**3.5.1.1 Age**

Chart

Description automatically generated

Fig.11 Model 1 Age Prediction Evaluation

From Fig.11, Although there are volatilities among accuracies in the validation dataset, the curve is overall stable. The model learns most of the patterns by the 20th epoch. Model 1 seems does not encounter overfitting problem predicting age class.

Table

Description automatically generated

Table.1 Model 1 Age Prediction Evaluation

As shown in Table.1, The overall and weighted accuracy for predicting ages in the testing dataset is 56.65% and 57.43%. We can see that the accuracy for predicting age group [0-6] is extremely high – 97.89%, which may be due to obvious features of children in the early age.

**3.5.1.2 Gender**

Chart, line chart

Description automatically generated

Fig.12 Model 1 Gender Prediction Evaluation

According to Fig.12, similar to age, model 1 learns most of the patterns to predict age class by 20th epoch. Overfitting does not exist in predicting gender. Training accuracy curve and validation accuracy curved converged by 60th epoch. The accuracy for predicting ages in the testing dataset is 90.95%.

A picture containing text, indoor, screenshot

Description automatically generated

Table.2 Model 2 Gender Prediction Evaluation

From Table.2, the overall and weighted accuracy are 90.95% and 91.06%. The accuracy for predicting male (93.32%) is higher than female (88.52%) may because there are more male images in the UTKFace dataset.

**3.5.1.3 Race**

**Chart, line chart

Description automatically generated**

Fig.13 Model 1 Race Prediction Evaluation

According to Fig.13, model 1 also learns most of the patterns to predict age class by 20th epoch. Overfitting does not exist in predicting race either. Training accuracy curve and validation accuracy curved converged very early – about 5th epoch. The accuracy for predicting ages in the testing dataset is 79.34%.

Table

Description automatically generated

Table.3 Model 1 Race Prediction Evaluation

From the report above (Table.3), the overall accuracy and weighted accuracy are 79.34% and 79.29%. Model 1 has very high accuracy of predicting Asian (90.96%) and White individuals (84.80%). Others class has the lowest accuracy due to its smallest sample size (Fig.4).

**3.5.2 Model 2**

**3.5.2.1 Age**

Chart, scatter chart

Description automatically generated

Fig.14 Model 2 Age Prediction Evaluation

From Fig.14, the volatilities among accuracies in the validation dataset is less than mode l , but the model may encounter overfitting problem after epoch 38, which could be solved by early stopping which has been applied in model2. The model learns most of the patterns by the 20th epoch. The accuracy for predicting ages in the testing dataset is 59.28%, higher than model 1.

Table

Description automatically generated

Table.4 Model 2 Age Prediction Evaluation

As shown in Table.4, The overall and weighted accuracy for predicting ages in the testing dataset is 59.29% and 59.28%, which are both slightly higher than model 1. We can see that the accuracy for predicting age group [0-6] is also the highest – 99.25% in model2.

**3.5.2.2 Gender**

Chart, line chart, scatter chart

Description automatically generated

Fig.15 Model 2 Gender Prediction Evaluation

As shown in Fig.15, model 2 encounter less overfitting problem predicting gender compare to the other classes. The model also learns most of the patterns by the 20th epoch. The accuracy for predicting ages in the testing dataset is 91.42%, higher than model 1.

Table

Description automatically generated

Table.5 Model 2 Gender Prediction Evaluation

Tabel.5 shows that model 2 has good performance on predicting age for both male or female (above 91% accuracy). The overall accuracy and weighted accuracy are 91.42% and 91.42%.

**3.5.2.3 Race**

Chart, line chart

Description automatically generated

Fig.16 Model 2 Race Prediction Evaluation

According to Fig.16, model 2 may encounter overfitting problem after epoch 45, however the accuracy is higher than model 1. The accuracy for predicting ages in the testing dataset is 81.16%, higher than model 1.

Table

Description automatically generated

Table.6 Model 2 Race Prediction Evaluation

By comparing Table.6 with Table.3, model 2 has much smaller accuracy on predicting Asian individuals (76.70% vs. 90.96%). Although model 2 has less accuracy on classifying Asian individuals compare to model 1, it has overall good performance on predict all the races, and has higher overall accuracy. The overall accuracy and weighted accuracy are 81.16% and 79.09%.

**3.5.3 Model 3**

**3.5.3.1 Age**

Chart

Description automatically generated

Fig.17 Model 2 Age Prediction Evaluation

Model 3 is the deepest model in this project, residual learning has been applied to avoid vanishing gradient problem. From Fig.17, both the volatilities and overfitting problem for predicting age are higher than the previous models. However, it has the highest overall accuracy – 59.49% for the testing dataset.

Table

Description automatically generated

Table.7 Model 2 Age Prediction Evaluation

As shown in Table.7, The overall and weighted accuracy for predicting ages in the testing dataset is 59.49% and 58.46%, which are both slightly higher than model 1. We can see that the accuracy for predicting age group [0-6] is also the highest – 92.81% in model3 (lower than model2).

**3.5.2.2 Gender**

Chart, line chart

Description automatically generated

Fig.18 Model 2 Gender Prediction Evaluation

As shown in Fig.18, model 3 encounter less overfitting problem predicting gender compare to model2. The model learns most of the patterns by the 10th epoch. The accuracy for predicting ages in the testing dataset is 91.22%, slightly lower than model 1 (91.42%).

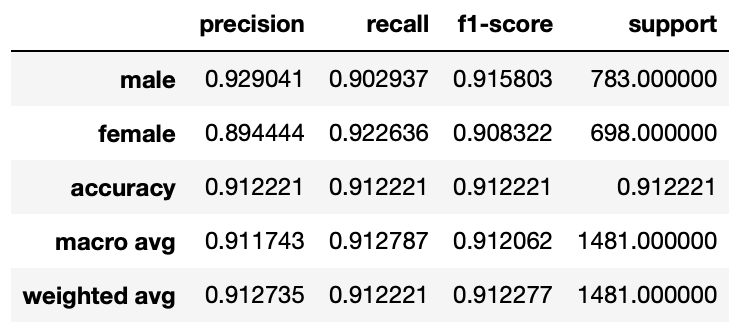


Table.8 Model 2 Gender Prediction Evaluation

Tabel.8 shows that model 3 has higher precision for predicting male than female (male 92.90% vs. female 89.44%). However. the model has higher recall rate (proportion of actual female was identified correctly) for female than male (male 90.29% vs. female 92.26%). The overall accuracy and weight accuracy for predicting gender of model 3 are 91.22% and 91.27%.

**3.5.2.3 Race**

Chart, line chart

Description automatically generated

Fig.19 Model 2 Race Prediction Evaluation

According to Fig.19, model 3 may encounter overfitting problem after epoch 40, however the accuracy is higher than both model 1 and model 2. The accuracy for predicting ages in the testing dataset is 81.23%.

Table

Description automatically generated

Table.9 Model 2 Race Prediction Evaluation

By comparing Table.9 with Table.6 (model2), model 3 has better performance on classifying White, Asian, and Indian individuals. The overall accuracy and weighted accuracy are 81.23% and 80.23%.

**3.5.3 Conclusion**

In this project, after adding simple convolutional layers and apply residual learning, the accuracies for image classification have been improved on age, gender, and race. The final model for this project is model 3.

**3.6 Model Output Visualization**

A picture containing text

Description automatically generated

Fig.20 Model 2 Race Prediction Evaluation

Fig.20 displays results of model 1 predicting randomly selected 16 face images.

**4 Conclusion & Discussion**

The accuracy for age is much lower than gender and race, two reasons may lead to the low accuracy. One is that the age groups of in this project were not divided properly. For example, in group 12-24, a 12 -year-old individual is much less mature looking compare to a 23-year-old individual, making it hard for the model to learn the common patterns. Therefore, the definition of each age group needs to be properly adjusted.

Another is that, in reality, people's aging speed differs, some people look younger while others look older at the same age. It is hard for humans themselves to predict an individual’s age, it would be even harder for machines to predict.

The final model of this project is model 3 – 16 convolutional layers with residual learning to avoid vanishing gradient problem, and 3 fully connected layers for each class (age, gender, and race). With model 3, the accuracy for classifying age, gender, and race are 59.49%, 91.22%, and 81.23%. Model 3 has better performance compare to model 1 and model 2, conforming that deep and easy neural network with residual learning can boost model performance.

**Reference**

[1] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D.Backpropagation applied to handwritten zip code recognition*. Neural Comput*., 1(4):541–551, 1989.

[2] Krizhevsky, A., Sutskever, I., and Hinton, G.E. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.[3] Karen Simonyan & Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In ICLR, 2015

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556,2014.

[5] Wang, J., Perez, L.: The Effectiveness of Data Augmentation in Image Classification using Deep Learning. <http://cs231n.stanford.edu/reports/2017/pdfs/300.pdf>

[6] Zhang, C., Vinyals, O., Munos, R., et al.: A Study on Overfitting in Deep Reinforcement Learning. <https://arxiv.org/pdf/1804.06893.pdf>

[7] Ioffe, S. & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.. CoRR, abs/1502.03167.

[8] Hernandez-Garcia, A., Konig, P.: Do deep nets really need weight decay and

dropout? <https://arxiv.org/pdf/1802.07042.pdf>

[9] Prechelt, L. (1996). Early Stopping-But When?. In G. B. Orr & K.-R. Müller (ed.), Neural Networks: Tricks of the Trade , Vol. 1524 (pp. 55-69) . Springer . ISBN: 3-540-65311-2.

[10] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell.Understanding data augmentation for classification: when to warp? CoRR, abs/1609.08764, 2016.