Accuracy Improvement of Age Prediction Model Based on AutoML

Shusong Xu¹, a, †, Zixuan Zhou², b, † and Make Tao³, c, †

¹Beijing Jiaotong University, Beijing 100044, China
²University of Illinois at Urbana-Champaign, Champaign, Illinois 61801, US
³Central South University, Changsha 410000, Hunan province, China

¹17281169@bjtu.edu.com, b,zixuanz3@illinois.edu, c,tmk8208181411@csu.edu.cn
†These authors contributed equally.

Abstract. Recommendations from websites are sometimes not suitable enough to fit the needs of different groups of people. One element in considering the recommended algorithm is the age range of users. To approach the goal of improving recommended algorithms in general, age prediction based on machine learning became the target of this research. We started by using AutoML from Google Cloud as the training tool and proportionally choosing 4,908 facial images of Asian people, which were categorized into six groups: images of the age below 10; images of the age from 10 to 20; images of the age from 20 to 30; images of the age from 30 to 40; images of the age from 40 to 50; images of the age from 50 to 60; and images of the age above 60. From the first training, we demonstrated 64.89 percentage of precision with the confidence threshold value of 0.5. After changing the threshold to 0.71, we optimized the result to 69.83 percentage of precision. By observing the result given by the deployment from Cloud training to an Android device, we found that the model differentiated Asian people aged from 10 to 40 most precisely, while it was relatively weak to distinguish people of the age from 50 to 60 and people of the age above 60. These results show an intelligent work of the prediction model on sifting people by ages and the strong relationship between the precision and the quantity of corresponding data. Its focus on distinguishing young people is suitable for our goal as young people are the main force of Internet users. By adjusting the composition of the dataset, we are looking for broader usage in other fields.

1. Introduction
Prediction plays an important role in our daily life, and it can be used in various fields, such as “Chronic Disease Risk Prediction” [1], “prediction model for dental caries” [2], and “age prediction.” Age estimation has various applications, ranging from customer relations to biometrics and entertainment [3] and a kind of semantic knowledge for multimedia content analysis and understanding [4]. There are many kinds of data to help train models, such as text [5], DNA [6], biochemical indicators, or images[7]. Facial morphology is one of the most important biological characteristics of human beings, which plays an important role in individual recognition and reflects the age of individuals most obviously. Therefore, we choose a deep learning model trained by human face images to achieve the age prediction. A large number of images are required to train most models based on face images for predicting ages, and they use some complex ways, like CNN or MTCNN, to
train their model [8-9].

CNN Face age prediction is nothing more than dividing faces into multiple categories, then training the convolutional neural network, and finally using the trained convolutional neural network for classification. However, there are several important problems in face age classification. First, face data sets are difficult to obtain; second, faces are highly sensitive to offset and light; third, features are not easy to extract. MTCNN enables people to produce more accurate models, but it still needs lots of images. We would like to focus on using age prediction on recommended algorithm. Thus, we find another way to solve these problems. The most contributions in this paper include 1) dataset: through artificial selection, we select about 5000 images. The proportion of each age group in the data set is different and is more skewed towards adolescents and adults since our model is used for the recommended algorithm, and the preference varies mostly in those ages. 2) training style: we use Google Cloud service to train it. Besides, Google Cloud provides a really convenient service; it is really easy for us to train a model and at the same time, deploying it on our edge devices like our cell phone.

2. Method

2.1 Experimental condition setup

We compare the traditional age prediction ML model in one article: https://github.com/serengil/deepface, with our deep learning model that we set up using some different techniques. The comparisons between the two methods are shown in Table 1.

Table 1. Comparison of Methods

| Main method | Feature Extraction | Learning Algorithm | Speed | Result | Optimize |
|-------------|--------------------|---------------------|-------|--------|----------|
| Baseline    | Traditional ML     | Human process of image extraction | Gradient Descent | Slow and sometimes stuck | Find the Precision and Recall in some mathematical ways manually | Simplify some algorithm in gradient descent |
| Proposed method | DL Auto process | CNN | Medium but acceptable | Automatically calculate the graph of precision and recall | Techniques like changing confidence threshold, multiple-labeled |

2.2 Dataset preparation

We download the human face dataset in the open-source website: http://mmlab.ie.cuhk.edu.hk/projects/MegaAge/, to provide a training dataset for our age prediction model. First, we design 7 labels of different age periods, then apply them to each image by constructing a CSV directory file, and the data distribution according to different labels are at Table 3. Then we use Google’s service to automatically divide them into three part: train set, test set, and validate set; the percentage distribution of images in each label is at Table 2.

Table 2. Percentage Distribution

| Percentage | Train | Test | Validate |
|------------|-------|------|----------|
| 70%        | 17%   | 13%  |          |
Table 3. Data Distribution

| Sample number | below_10 | 10s | 20s | 30s | 40s | 50s | over_60 |
|---------------|----------|-----|-----|-----|-----|-----|---------|
|               | 1,416    | 794 | 1,390 | 833 | 255 | 110 | 120     |

2.3 AutoML training
We upload the dataset to the Google AutoML and train the model. After training, we get the accuracy of the model learning at different thresholds and the confusion matrix for each label. Table 4 shows how the Precision and Recall change with the threshold varies. Table 5 shows the confusion matrix.

2.4 Optimization

2.4.1 Change the threshold to make the trade-off between precision and recall
Our purpose is to have higher precision to get the best prediction, so we change the threshold to get higher precision and still acceptable recall. Finally, we get a value of 0.71. Table 6 shows the Precision and Recall result when we optimize the threshold to the value of 0.71.

2.4.2 Enlarge the dataset to make the confusion matrix more accurate
We found our confusion matrix was not as accurate as we expected, so enlarging the dataset is our attempt. We add 1,000 more images into the dataset. Then we find the confusion matrix get improved. Table 7 shows the confusion matrix after optimization.

2.5 Deploy
We deploy the model to the Android edge device and run. Figure 1 shows one example of the model’s running result.

3. Results

3.1 Precision and Recall

| Threshold | Precision | Recall |
|-----------|-----------|--------|
| 0.1       | 43.76%    | 87.13% |
| 0.25      | 55.53%    | 65.43% |
| 0.5       | 64.89%    | 59.59% |
| 0.75      | 71.88%    | 49.20% |
| 0.9       | 80.45%    | 25.67% |

Precision means the proportion of the true positive instances (TP) in all instances that the machine predicts positive (TP + FP), Recall means the proportion of the true positive instances (TP) in all positive instances in the original sample (TP + FN), Threshold means the minimum confidence that the machine predicts one image as positive, any image with confidence less than the threshold will be removed in the calculation.

With threshold larger, Precision gets higher, and Recall gets lower.
3.2 Confusion Matrix

Table 5. Confusion Matrix

| True label/ Predict label | below_10 | 10s | 20s | 30s | 40s | 50s | over_60 |
|---------------------------|----------|-----|-----|-----|-----|-----|---------|
| below_10                  |          | 90% | -   | -   | -   | -   | -       |
| 10s                       | 16%      | 51% | 32% | -   | -   | -   | -       |
| 20s                       |          |    17% | 58% | 21% | -   | -   | -       |
| 30s                       |          | -    | 31% | 54% | 11% | -   | -       |
| 40s                       |          | -    | -   | 52% | 16% | 16% | -       |
| 50s                       |          | -    | -   | 18% | 26% | 18% | 36%     |
| over_60                   |          | -    | -   | -   | -   | 17% | 75%     |

Confusion matrix represents the frequency that the machine makes certain prediction on a given label, like 90% is the frequency that the machine predicts an image with label below_10 as below_10. On the diagonal is the frequency of the true prediction, which means the model’s prediction is exactly the truth. While the other areas are false prediction, which means the prediction is not the truth and the machine makes mistakes here.

We found in confusion matrix that some false prediction with high frequency always happens in adjacent age periods, for example, the 52% happens at true label 40s and predict label 30s, the 36% happens at true label 50s and predict label over 60s.

3.3 Precision and Recall at threshold of 0.71 after optimization

Table 6. Precision and Recall at threshold of 0.71

| Threshold / Accuracy | Precision | Recall  |
|----------------------|-----------|---------|
| 0.71                 | 69.83%    | 51.02%  |

We start at the threshold of 0.5 in Table 4, which seems a reasonable result for both Precision and Recall, and increase the threshold to increase the Precision, while in the process the Recall gets lower. Finally, we decide 0.71, at which the Recall is the minimum acceptance value. So with our goal to get higher Precision, the reason why the decrease of Recall can be accepted is that Precision represents TP/TP+FP, while Recall represents TP/TP+FN. The FN number, that the machine predicts negative is false, and do not influence the accuracy a lot when it is small. Therefore, we choose to decrease Recall to make the trade-off. Also, if we decrease the Recall lower than 51%, the FN number will get large enough to influence the accuracy, which gets the opposite effect. Thus, 51% is the minimum
acceptance value. The optimization is satisfactory; in the Threshold of 0.71, we have Precision close to 70%.

3.4 Confusion matrix after optimization

Table 7. confusion matrix after optimization

| True label/ Predict label | below_10 | 10s | 20s | 30s | 40s | 50s | over_60 |
|---------------------------|----------|-----|-----|-----|-----|-----|---------|
| below_10                  | 89%      | -   | -   | -   | -   | -   | -       |
| 10s                       | 15%      | 51% | 29% | -   | -   | -   | -       |
| 20s                       | -        | 15% | 55% | 26% | -   | -   | -       |
| 30s                       | -        | -   | 28% | 60% | 11% | -   | -       |
| 40s                       | -        | -   | -   | 39% | 32% | 14% | -       |
| 50s                       | -        | -   | -   | 26% | 27% | 30% | 15%     |
| over_60                   | -        | -   | -   | -   | -   | 32% | 58%     |

In our discussion, the possible explanation for these high false predictions is that human face feature does not change a lot in only one decade. That is where the machine easily makes confusion. So the solution is to enlarge the dataset to make the machine can distinguish the small difference between adjacent age periods.

So after we enlarge the dataset, we find the frequency of some false prediction greatly deducted, like 52% lower to 39%, and the true prediction on the diagonal get a little higher.

3.5 Running result

Figure 1. Running result example
The figure above is a 30s woman image, which we use as an example to test our model’s prediction. The model gives such predictions in 78ms time: 30s with 48.00% confidence, 40s with 26.00% confidence, below_10 with 16.00% confidence, so the prediction is quite reasonable.

4. Discussion

4.1 Add one more label to assist the age prediction
We add one more label to each image: male or female, to provide auxiliary assistance for the age prediction. Specifically, we know the human face at the same age differs between the two genders, so our model tries to distinguish certain age periods in accordance with their gender. In the training process, the model first classifies the images into two genders’ category, then distinguishes age periods in one certain category, so the prediction gets more accurate with the gender label.

4.2 Consider more feasible algorithms
As our model is based on the Google cloud platform, so if we want to duplicate our work in some new setting environment, it is better to know more feasible algorithms. We think about the Adagrad Gradient Descent algorithm to help the model find the minimum loss precisely, the classification method to allocate the images into different categories in several seconds, and the feature scaling technology to reduce the unstableness of the feature extracting process. We intend to try all of them in the future training process.

4.3 Add one more label for future application
We also add one more label: Adolescent or not. The reason we use this label is to apply it for future market use. We know that in some restricted areas, adolescents should not get admitted into. So we aim to predict user’s age to judge whether they are permitted to get access to some restricted Information. We believe this can be applied in some video game permission, which leads adolescents to healthier online surfing.

5. Conclusion
We have taken the first step to the improvement of recommended algorithms by classifying the user group by ages. We have approached the issue from a purely computational perspective and have shown that using AutoML from Google can perform the age prediction with accuracy close to 70%. We mainly trained the model by classifying the facial image to the appropriate age category, including all facial details into account. The model performed better when distinguishing young people than when categorizing older adults, which fits our goal as young people are the main consumers of the Internet industry. Though our work has achieved periodic success, the insufficiency of accuracy is still an issue waiting for a solution. The prediction is confused when distinguishing the person from two close groups i.e., from 20 to 30 and from 30 to 40, which may be solvable by improving the pre-training refinements of images and increasing the training data. Future work can be explored to more aspects of improving recommended algorithms by categorizing the user group.

References
[1] Yonsei Med. (2014) Development and Application of Chronic Disease Risk Prediction Models, 1-2
[2] Abernathy, J. R., Graves, R. C., Bohannon, H. M., Stamm, J. W., Greenberg, B. G., & Disney, J. A. (1987). Development and application of a prediction model for dental caries. Community dentistry and oral epidemiology: 15(1), 24-28.
[3] Can Malli, R., Aygun, M., & Kemal Ekenel, H. (2016). Apparent age estimation using ensemble of deep learning models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops: 9-16.
[4] Guo, G., Fu, Y., Dyer, C. R., & Huang, T. S. (2008, June). A probabilistic fusion approach to
human age prediction. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*:1-6.

[5] Nguyen, D., Smith, N. A., & Rose, C. (2011, June). Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*: 115-123.

[6] Anastasia Aliferia, David Ballard, Matteo D.Gallidabino, Helen Thurtle, Leon Barron. (2018) DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models.

[7] Akhtar, F., Li, J., Azeem, M., Chen, S., Pan, H., Wang, Q., & Yang, J. J. (2019). Effective large for gestational age prediction using machine learning techniques with monitoring biochemical indicators. *The Journal of Supercomputing*: 1-19.

[8] Chen, S., Zhang, C., Dong, M., Le, J., & Rao, M. (2017). Using ranking-cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*: 5183-5192

[9] He, X., Wang, P., Zhao, Z., Zhao, Y., & Su, F. (2019, July). MTCNN with Weighted Loss Penalty and Adaptive Threshold Learning for Facial Attribute Prediction. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*: 180-185.