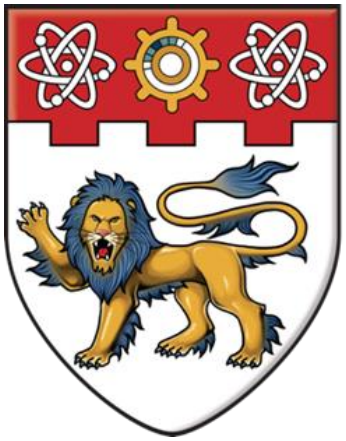


NANYANG TECHNOLOGICAL UNIVERSITY

CZ4042 Neural Networks and Deep Learning

Age and Gender Classification

Chen Xingyu U2021140F

Cao Qingtian U2020646L

Table of Contents

1. Introduction

1.1 Background and Motivation

Gender is basic biometrics and can be used in many applications. Hence, in this report, we first implemented gender classification models proposed by Gil Levi and Tal Hassner (refer as the original caffe model), and explored its variants on ADIENCE[1] data. Meanwhile, age is another valuable feature and the "hidden facial features" used to classify gender may be helpful for age as well. Therefore, joint classification using Caffe models for age and gender was also developed. Finally, a ResNet model pre-trained on CELEBA[2] was fine-tuned for joint classification as a comparison with the Caffe models.

1.2 Related work

Gil Levi and Tal Hassner proposed a model (Caffe model) with 3 convolutional layers and 2 hidden fully connected layers. It is able to predict gender for around 86% accuracy and age for around 50%. It has the advantage of needing no pre-training, and being a small network, making it cost-effective to train. In addition, Zhang et al. proposes a new way of age encoding that will improve accuracy for age classification. However, neither the authors did not carry out joint classifications for age and gender which we think the two tasks together might have the potential to give better results. We also investigated swapping the base Caffe model with ResNet (residual neural network), since it is a deep network that excels in feature extraction. It solves the problem of vanishing gradient through skipping layers, and has 25% type I error on ImageNet dataset. In this report, we aim to combine all these techniques to both models and test if the results are better.

2. Data Handling

Data cleaning and preprocessing were performed on ADIENCE only since CELEBA dataset is available within PyTorch and Tensorflow library.

2.1 Data cleaning

For gender, we first removed those with 'u' category, as we focus on finding if a portrait is a male or female. NaN values are also removed. For age, we add the pictures with singular age into their respective age groups. Below shows a mapping table for age group correction:

Original age group	Corrected age group
2, 3	(0, 2)
(8, 23), 13	(8, 12)

[1] ADIENCE refers to age and gender dataset provided by Gil Levi and Tal Hassner in *Age and Gender Classification Using Convolutional Neural Network*

[2] CELEBA is a large scale dataset containing over 200 thousand celebrity faces with 40 facial features such as 'big nose', 'chubby' being annotated, provided by Liu et al. in *Deep learning face attributes in the wild*

22	(15, 20)
23, (27, 32), 29, 32, 34,	(25, 32)
(38, 42), (38 ,48), 35, 36, 42, 45	(38, 43)
46, 56	(48, 53)
55, 57, 58	(60, 100)

Table 0. Age group correction mapping

2.2 Data preprocessing

For all Caffe-related models, we use the data transformation described in the paper (Levi & Hassner, 2015), namely resizing to (256,256), followed by a random crop to (227, 227), and then a random horizontal flip. For validation and testing, instead of random crop, we use centre crop, and we do not apply random horizontal flip.

For ResNet related models, we first use MTCNN to crop the images to the faces to reduce the noise of background features. The images are then resized to (224, 224) as required by ResNet models and pixels are normalised to (0,1) scale. Data augmentation such as RandomFlip and RandomContrast are also added to the models.

2.3 Train-test split

For CELEBA dataset, we follow the train-validation-test split in the dataset library. We will pretrain on the train set, and use the validation set and test set combined to select the parameters of pretrained models.

Split	Examples
'test'	19,962
'train'	162,770
'validation'	19,867

Fig 0. Train-validation-test split on CELEBA dataset

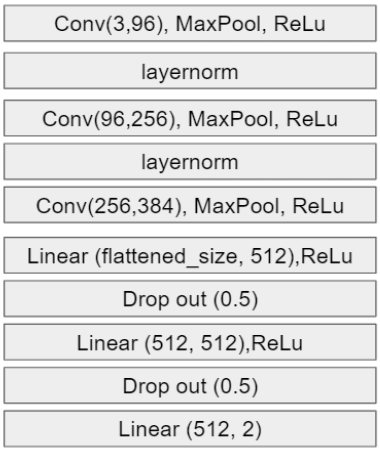
For ADIENCE dataset, the authors provided data in 5 folds since they trained the models using 5-fold cross-validation. In our experiments, instead of using 5-fold cross validation that takes a long time to obtain the results, we opt to use fold 0 to 2 as training data, fold 3 as validation data for model selection, and fold 4 for test where all our models are evaluated upon.

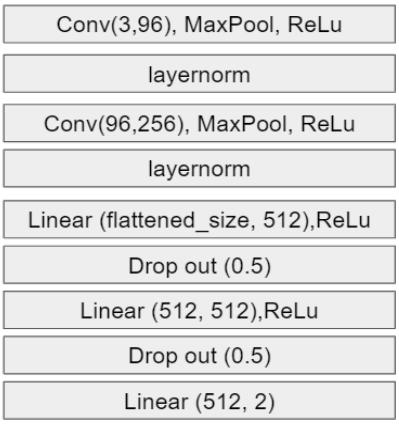
3. Caffe model

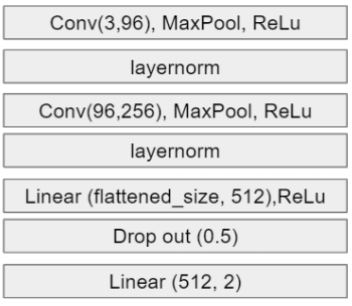
3.1 Reduction of layers on original caffe model

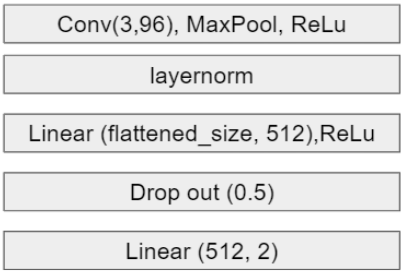
We hope to investigate the effect of reduction of layers on the model by Gil Levi and Tal Hassner (refer as the original caffe model). We implement the model (Levi & Hassner, 2015) in PyTorch, train through 10 epochs with 64 as batch size. We have a multistep learning rate scheduler that changes the learning rate from 1e-4 to 1e-5 at 10000th iterations - this is unlike the paper, for we find that in our case, this produced better results within 10 epochs with such a batch size.

After we implemented the original caffe model, we then derived 3 more models from original model, as shown below (table 1)

<table>
<tr>
<td>


Model 1(original Caffe model)
</td>
<td>


Model 2(remove 1 convolutional layer)
</td>
</tr>
<tr>
<td>


Model 3 (remove both a convolutional and a linear layer)
</td>
<td>


Model 4 (a model with two convolutional and a linear layer)
</td>
</tr>
</table>

Table 1: Models to investigate. For brevity's sake, we omitted some details of parameters of the original paper here, but in our code we followed them.

Model	Accuracy	f1	precision	recall
1	**0.827**	**0.822**	**0.784**	**0.864**

2	0.770	0.769	0.718	0.829
3	0.809	0.800	0.771	0.823
4	0.761	0.761	0.707	0.823

Table 2: result on the test set

We discover that the accuracies are decent, but there is the issue of slight overfitting for Model 1, 2, 3 at the end (train accuracies are at around 90% but the validation accuracies are at around 80%). Model 4 does not appear to overfit as much, probably because it is the most simplistic model, and it already cannot fit the train data well. For overfitting models, we try to modify the L2 regularisation and increase dropouts, but the result was not improved, so we just use the original parameters from the paper. Overall, we find that reducing the convolutional layer has a greater impact than reducing the linear layer, as we can see there is a significant drop in accuracy from model 1 to model 2, but not so from model 2 to model 3. In fact, there is a slight increase, but we suspect that it is because the model happens to perform badly on fold 4. If we use 5 fold cross validation and take an average, the result might turn out to be about the same.

3.2 Pre-train Caffe model on CELEBA for gender classification

We train the caffe model on CELEBA dataset for 10 epochs with batch size 64 and learning rate 1e-4 without the scheduler, and get the lowest loss on validation set. We save that model, and then fine tune on ADIENCE with batch size 64 and learning rate from 1e-4 to 1e-5 with multi-step learning rate (milestone = 10000), as seen in fig 3.

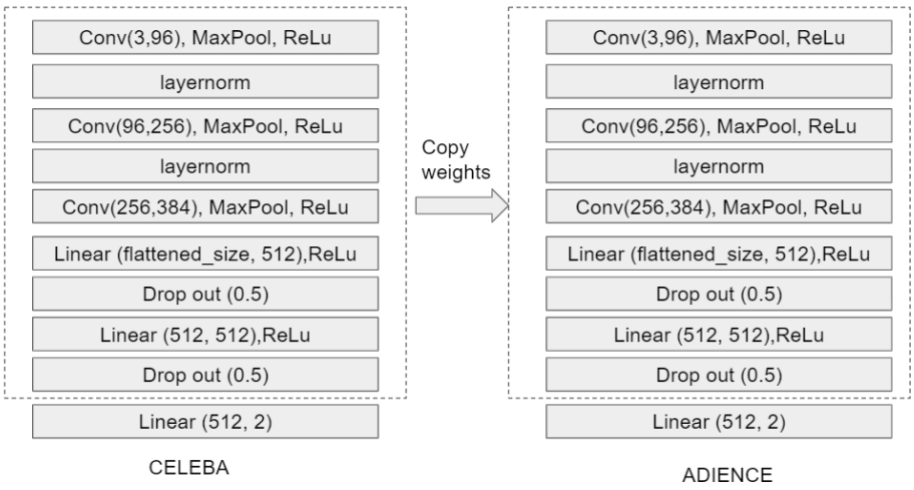


Fig 3: Fine tune model on ADIENCE after pre-trained on CELEBA

Model ID	Dataset	Gender accuracy	Gender f1	Gender precision	Gender recall
5	CELEBA	0.978	0.973	0.971	0.976
5	ADIENCE	0.792	0.771	0.785	0.756

Table 4: Accuracies of pre-training caffe on CELEBA, followed by fine-tuning of ADIENCE

Its overall accuracy, compared to the original caffe model, shows no significant improvement. However, we notice that the accuracy of ADIENCE is climbing up rapidly. At epoch 2, there are already 80% validation accuracies. Still, due to the difference between the nature of ADIENCE dataset and CELEBA dataset such as containing blurred images versus clear ones, the effect of pre-train is not apparent. The accuracy after 10 epochs fails to become higher.

3.3 Joint classification of gender and age on Caffe models

Based on the Caffe model, we propose various ways to allow models to integrate age and gender prediction together (fig 5, fig 6, fig 7). The batch sizes are all 64, with a multi-step learning rate starting from 1e-4 to 1e-5 at the 10000 milestone. Since the loss is usually high for such models, we train for 20 epochs.

3.3.1 16-categories

We consider age and gender jointly in a category. That means, a female from age group 1 is in a different category from a male from the same age group. Since we have 2 genders and 8 age groups, we will have 16 categories in this way. As such, we modified the original caffe model to have an output layer for 16 categories.

Fig 5

3.3.2 unweighted-caffe

We create two heads of age and gender, and then we compute the average of the two losses (each with a weight of 1) and perform back propagation on it. Then we select 3 models, one for each category (age, gender, both), based on the lowest loss. We then test the accuracy of each model.

Fig 6

3.3.3 weighted-caffe with new age encoding

As suggested by Zhang et al, age classification is generally harder as age groups are interrelated and people have inconsistent ageing processes in different age ranges. However, an easier question to answer is who is older or younger among the two. Hence, we proposed a new age encoding which considers the ordinal nature of age, i.e. a picture which is younger than group 2 is also younger than group 3. We create 8 age heads, each age head handling the query: is the picture younger than my age group? If the answer is yes, it will return 1, else it will return 0. All the 8 age heads

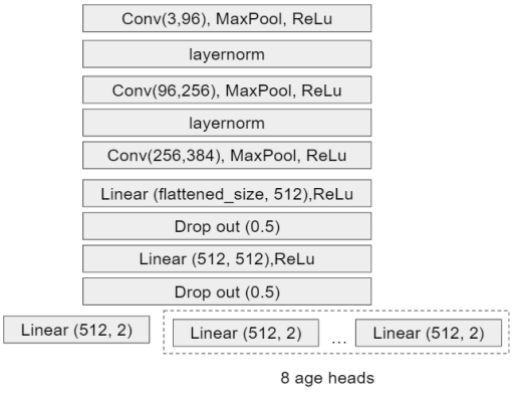


Fig 7

will then return something like [[0],[0],[0],[0],[0],[1],[1],[1]] (If the question is ‘is the picture older than my age group’, the vector will be a series of [1] followed by [0]). After that, we optimise a weighted loss, with weights of 8 heads’ losses: [1,1,1,1.3,1.5,1.5,1.3,1] (Zhang et al., 2017), with heavier weights on the age categories with highest mis-classification. We then put a weight of 9.6 on gender to balance age and gender classification. Like unweighted-caffe, we each select the lowest validation loss from age, gender and both, then test the best model on the dataset.

Model ID	Model	Gender accuracy	Gender f1	Gender precision	Gender recall	Age accuracy
6	16-categories	0.760	0.745	0.734	0.756	0.384
7.1	Unweighted-caffe : best loss on age	0.779	0.759	0.765	0.752	0.399
7.2	Unweighted-caffe: best loss on gender	0.806	0.800	0.764	0.840	0.406
7.3	Unweighted-caffe: best loss on both	0.779	0.759	0.765	0.753	0.399
8.1	Weighted-caffe : best loss on age	**0.838**	**0.818**	**0.849**	**0.789**	**0.458**
8.2	Weighted-caffe: best loss on gender	0.831	0.819	0.811	0.828	0.454
8.3	Weighted-caffe: best loss on both	0.836	0.829	0.802	0.858	0.448

Table 8: Results on 16-categories, unweighted-caffe and weighted-caffe for 20 epochs

As we can see from table 8, in 20 epochs, the weighted-caffe model (8.1, 8.2, 8.3) generally performs the best. Its gender accuracy is around 0.03 higher than the rest, and the age accuracy is around 0.05 higher than the rest. It is clear that the ordinal nature of age encoding helps to uncover hidden patterns and allow the model to perform better overall. It is also interesting to see that using a different age encoding improves the accuracy of gender classification, hinting that age and gender attributes might indeed have some relationships. Hence, doing joint classification this way will improve both of their accuracy. To further investigate the effect of multitasking, we compare weighted-caffe with a model that does not include an age head, like the original caffe model. Since it is unclear whether the extra accuracies are due to the extra epochs or the multi-tasking, we re-train the original caffe network and obtain the test accuracies for 20 epochs. The results are accuracy: 0.827, f1: 0.822, precision: 0.784, recall: 0.864, and none are better than weighted caffe. This shows that multitasking might indeed be helpful in increasing gender accuracies and age accuracies.

4. ResNet34 model

4.1 Pre-train ResNet34 model for multi-task classification of age and gender

We first tried out pre-training on CELEBA then used multi-task classification for age and gender, because multi-task model shows the best results from Caffe model, and we hope using pre-training technique will also improve the accuracy. CELEBA has 40 facial attributes, we removed the classification layer from ResNet34, added 40 binary classifiers for presences of each feature. It is expected that the pre-trained model will learn to identify facial features which can help speed up learning of age and gender classification. After pre-training on CELEBA, we use 40 features as input and add age output and gender output as shown in Fig 9.

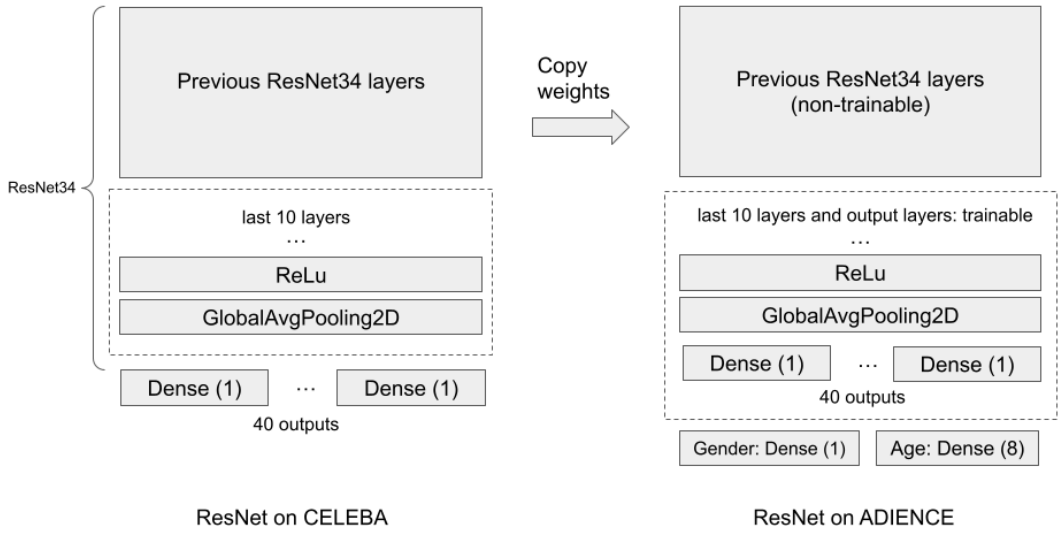


Fig 9. ResNet on CELEBA architecture (left) and ResNet on ADIENCE architecture(right)

ResNet34 converges fast on CELEBA, it was stopped at 6 epochs by early stopping on val_loss with patience of 3. The mean accuracy of 40 features on test dataset is over 90%.

The pre-trained model was then used to train ADIENCE with only the last 10 layers except output layers being trainable and age and gender outputs are added on top (shown in Table 10).

Model ID	Gender accuracy	Gender f1	Gender precision	Gender recall	Age accuracy
9	0.738	0.766	0.732	0.805	0.354

Table 10. Test results on ADIENCE

Above shows the test result on ADIENCE after training for 20 epochs. Compared with Caffe models on multi-task classification, the results are slightly worse. It shows that deeper CNN does not always outperform simpler models, probably because when the model is deeper, there are too many parameters and optimization becomes more complex and difficult. Overfitting was also observed during training and it indicates the model might require more data to perform better.

4.2 Modify ResNet34 model for multi-task classification of age and gender

Since ResNet34 pre-trained using CELEBA does not improve accuracy from the base model, more layers were added to see whether it could improve accuracy. The idea was inspired by Squeeze-and-Excitation Networks proposed by Hu et al. We added one Squeeze-Excitation block after the final activation layer in ResNet34, followed by age and gender classifiers on top. We still freeze training of all layers except the last 10 layers in ResNet34. Its structure is shown in Fig 11

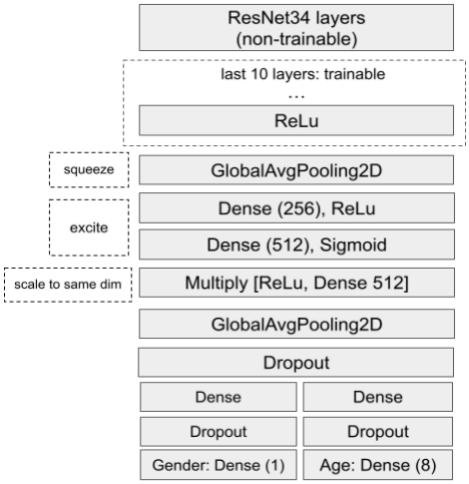


Fig 11

Model ID	Gender accuracy	Gender f1	Gender precision	Gender recall	Age accuracy
10	0.721	0.753	0.715	0.794	0.352

Table 12. Test results on ADIENCE

From test results shown in Table 12, adding one more squeeze-excitation block also did not help with the results. It is possible that due to additional layers, the model is prone to overfitting. As observed during training, training accuracy of gender and age rose faster, but was stopped by early stopping on val_loss with patience of 5. As a result, the model fails to perform better on test data.

4.3 Using loss_weights and new age encoding on ResNet34

We also adopt new encoding of age to train ResNet34 with and without additional layers as described in sections 4.1 and 4.2. Since 8 age tasks of whether the image is older than this age group are related, we use multi-task learning again and generate 8 outputs for different age groups and 1 output for gender. Different loss_weights are applied to model 11 & 13 and model 12 & 14, as models with additional layers are prone to overfitting and high loss_weights on gender drives gender accuracy to saturation (over 99%) very fast. Weights for gender are set higher at the first few epochs. Since there are many age outputs, if gender has weight at the same scale as every age group, the models did not learn gender classification. Hence, we made the models focus on optimising gender output first. After gender accuracy becomes high enough, we lower its weight to let the model focus on age outputs.

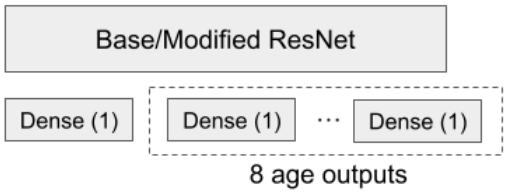


Fig 13 Model structure

Model ID	loss_weights								
	gender	age1	age2	age3	age4	age5	age6	age7	age8
11 & 13 (start)	4	1	1	1	1.3	1.8	1.8	1.3	1
11 & 13 (the rest)	0.9	1	1	1	1.3	1.8	1.8	1.3	1
12 & 14 (start)	2	1	1	1	1.3	1.8	1.8	1.3	1
12 & 14 (the rest)	0.5	1	1	1	1.3	1.8	1.8	1.3	1

Table 14. Loss_weights for different models

In addition, apart from freezing training of some layers (model 11 & 13), we also tried making all layers trainable (model 12 & 14). Below table documents the test results:

Model ID	Model	Pre-trained	Age encoding	Gender accuracy	Gender f1	Gender precision	Gender recall	Age accuracy
9	Base ResNet	yes	original	0.738	0.766	0.732	0.805	0.354
10	Modified ResNet	yes	original	0.721	0.753	0.715	0.794	0.352
11	Base ResNet	yes	new	0.742	0.765	0.743	0.788	0.383
12	Modified ResNet	yes	new	0.733	0.749	0.752	0.746	0.366
13	Base ResNet	**No, all layers trainable**	**new**	**0.808**	**0.825**	**0.804**	**0.847**	0.422
14	Modified ResNet	No, all layers trainable	new	0.784	0.793	0.812	0.775	**0.459**

Table 15: Results comparison for various ResNet34 models

Compared with models trained using original age encoding (model 9 & 10), the results generally increased. Especially the age accuracy which is hard to train. Models with original age encoding seem to have difficulty going beyond 40% accuracy. However, with new age encoding, models with some non-trainable weights slightly increase from original encoding models. The model with all trainable layers performed even better and was able to reach 46% test accuracy. During training, it was also observed that age accuracy on validation reaches 40% very fast and stays above 40%, as opposed to models with original age encoding which struggles to go over 40% accuracy. It shows that new age encoding helps age predictions.

In addition, comparing model 11 & 12 with model 13 & 14, training all layers seems to fit better on ADIENCE, probably due to the fact that ADIENCE has some blurred images as well as faces not properly aligned while CELEBA has clear and correctly positioned faces. Hence, pre-training on

CELEBA did not help much and a model with more layers trainable can fit better on this complex multiple tasks. Also, since CELEBA has few age-related features, keeping weights pre-trained on CELEBA might not help with age classification and using full layers on ADIENCE gave better results. Also, since model 13 is trained using higher gender weight, it could give better results on gender classification while model 14 uses lower gender weight and same age weights, so it tends to focus more on age classification. This may explain why model 13 gave best gender results and model 14 gave best age results.

It was also worth mentioning that the model was saved based on mean accuracy of age and gender on validation data, and it reached 85% for gender and 48% for age for both Model 5 and 6. Although the test performance did increase from models trained on original age encoding, it was not as high as on validation data. It could be due to an unfortunate split where test data is different from data use for training and validation.

5. Conclusion

In this report, we have investigated the modification of two base models, Caffe and ResNet34. We conclude that deeper models generally perform better, provided that it is optimised correctly. This is supported by the reduction in accuracy if we remove layers from the Caffe model, and the increase in age accuracy if we add layers to the ResNet34 model. However, the additional layers do not seem to help increasing gender accuracy in a joint model, and that might be because of the common train accuracy saturation problem mentioned by He et al.(2015). Additionally, we discover that removing convolutional layers has a greater impact than removing linear layers, probably because the former is responsible for feature extraction, which contributes more in accuracy of final prediction.

Regarding pre-training, we find that pretrained models do not seem to improve gender accuracy, mainly because of the difference of CELEBA and ADIENCE dataset. CELEBA has clear images, while ADIENCE does not. Pretraining on CELEBA does not seem to help the model to classify blurred images correctly, which is the main source of mis-classification.

We also discover that joint classification of age and gender does help with each other, seeing that the multi-task model with ordinal age encoding performs the best for Caffe base model. As a result, we adopt this method to all our ResNet34 models, and all of them achieve decent accuracy.

Lastly, we discover that the best model based on Caffe (model 8.1) and ResNet34 (model 13,14) are similar, with Caffe being better at gender classification. It is unexpected since ResNet is a deeper model, however, it could be due to the difference in gender weights. In ResNet34, gender weights are generally small, while caffe makes sure its weight and the sum of age weights are 1:1. This also shows that deep models may not always outperform simple models since having too many parameters makes optimization more complex and difficult. Another reason is that ResNet might require a more diverse pretraining dataset to perform better, as it will better learn how to extract features from non-standardised pictures (such as blurred, tilted pictures).

Reference

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

Levi， G. and Hassner T. Age and Gender Classification Using Convolutional Neural Networks. IEEE Workshop on Analysis and Modeling of Faces and Gestures (AMFG), at the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, 2015.

Zhang K., Gao C., Guo L., et al, “Age Group and Gender Estimation in the Wild With Deep RoR Architecture,” IEEE Access, vol. 5, pp. 22492– 22503, 2017.

Links

Codes without model files:

https://drive.google.com/file/d/1q0Aajl8WfvRL2Dcmbi78CZWHTL6ayP1s/view?usp=sharing

Codes with model files:

https://drive.google.com/file/d/1MpJu8xkiJKIVyO5_T44nwvry_7RnNpEL/view?usp=sharing

Tensorflow ADIENCE dataset link:

https://drive.google.com/drive/folders/1rzp8qdsuDFx6kuZYbqkmycUmoYkJjnv8?usp=share_link

Video:

https://youtu.be/NNFO7ZcgX70