Gender and Age Classification using Multi-task Learning with Convolutional Neural Networks

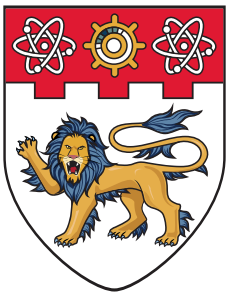
---

Adience Benchmark dataset

Neural Networks & Deep Learning

Kasper Jørgensen	N2202234F
Mads Ringsted	N2202230G
Tejas Rajagopal	U2023194K

November 11, 2022



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

1 Introduction

Computer vision is a field of artificial intelligence that has been growing in prevalence and it is now being used in several applications like surveillance, self-driving cars, and cancer detection. These are all critical tasks that, if made effective by computer vision, could improve the lives of many people. The fundamental building block of computer vision is the ability to perform image classification and this paper examines one such application and modern day techniques on the matter.

In the last decade, the performance of deep neural networks have come a long way and they are now being used extensively in several crucial applications. This paper examines leveraging convolutional neural networks (CNN) and state-of-the-art architectures to perform gender classification on real-world images. As shown in [1], a simple CNN with a rather basic architecture was able to outperform the former benchmark models of 2015 when classifying the age and gender of subjects in images. CNNs have turned out to be a very powerful model due to its ability to extract and learn generalised features of domain-specific data. This simply means that a CNN is able to learn the features of the images from scratch. Thus, figuring out which aspects of the image that are important for the classification task of the given domain [2]. The fact that CNN is a sub-type of neural networks inherently gives it a lot of flexibility and freedom in regards to the construction of the model, as there are a lot of options to increase/reduce the number of parameters of the model.

This paper seeks to explore properties of various types of CNN architectures and their advantages. This includes tweaking a few parameters and observing the change in training time versus model performance. This paper also examines multi-task learning (MTL) in neural networks. Rather than training a model on a single task, we explore training a model on multiple related tasks and examine how the model performs better on the original task. In this case, by considering age and gender recognition simultaneously, we were able to leverage both gender-specific age characteristics and age-specific gender characteristics to drastically reduce training time while achieving similar gender and age classification performance. Lastly, the paper examines leveraging transfer learning simultaneously with MTL to exploit generalized features of the model in the new problem domain.

1.1 Related work

Gender and age classification have previously been done on the Adience dataset in [1]. They build a simple CNN model consisting of three convolutional plus pooling layers, followed by three fully connected layers. The shallow network structure as well as their implementation of data augmentation in the form of image cropping in conjunction with dropout layers greatly reduces the risk of overfitting. This model is able to drastically outperform former benchmark models and achieve accuracies of 86.8% and 50.7% for gender and age classification respectively.

Another attempt by Kim, T.S and Sohn, S.Y [3], examines the use of multi-task CNNs in order to predict the metric of Remaining Useful Life (RUL). RUL is a metric that represents the remaining time a system will stay functioning. They state that RUL has mostly been treated as an independent process, while they suggest that there is merit for it being related to health status identification. So they propose a multi-task CNN setup that utilizes a shared network to extract features and then individual feed forward networks to perform the classification tasks. They found that combining the two problems using MTL leads to a significant performance improvement when compared to various single task models.

Lastly, transfer learning, the act of utilizing previously trained models in order to exploit general features for problems that lack data, have been explored by Gopalakrishnan et al.'s paper. They construct a CNN model to detect cracks in pavements based on images. They only have access to a limited amount of labelled data—around 1000 images. To navigate this constraint, they use the keras VGG-16 deep convolutional network [5] trained on the ImageNet dataset [4]. After which, they append a simple model such as a single layer neural network or logistic regression to predict on the features generated by the VGG-16 model. The paper indicates that the VGG-16 model provided by keras is a nearly ready-to-use platform that allows for easy cross-domain image classification.

2 Data

This project makes use of the Adience dataset for training and testing our models. The dataset consist of images of over 2000 subjects posed in real-world settings, i.e containing various noise in the form of lighting, orientation etc. Moreover, not all images were taken from a frontal perspective. The images are labeled with both the age (given as an interval) and gender of the given subject. For this project, the aligned version of the dataset is used, since it removes the need for any data preparation which is not included in the scope of the project. The Adience dataset serves as a reliable benchmark for testing and, hence, the models mentioned in this paper were tested on both gender and age classification tasks. Figure 1 shows 5 random samples from the dataset.

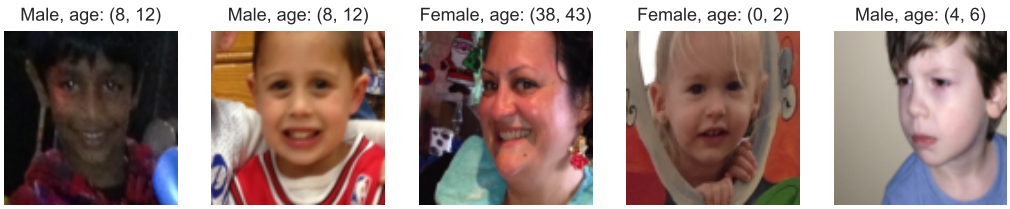


Figure 1: Sample data from the Adience dataset.

2.1 Data Preprocessing

In our preproccessing pipeline, we remove data that is uncategorized and data that does not follow the class labels: gender: {f, m} and age: {(0, 2), (4, 6), (8, 12), (15, 20), (25, 32), (38, 43), (48, 53), (60, 100)}. The dataset contains 16228 labeled facial images after removing invalid data. The class labels are then encoded with integers. The age distributions of male and females are illustrated in Figure 2. There are slightly more females in the data set $\sim 53\%$, but the distributions are quite similar. However, there is significantly more data of people in the age group of 25 to 32. Due to the greater amount of data, we expect our model to be most effective on people in this age range.

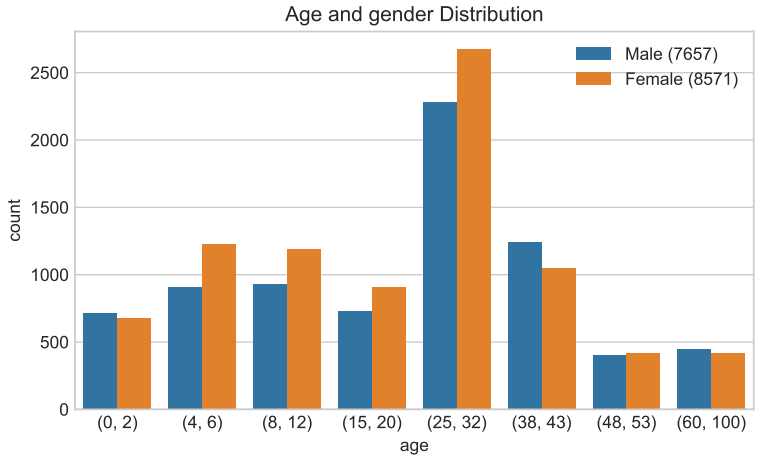


Figure 2: Age and gender distribution of the Adience dataset.

2.2 Data Split for Cross-Validation

Images were provided by different users and multiple pictures of the same person may appear in a user's gallery. It is also possible for images from the same user to include another person from the dataset in the background. Hence, the dataset was divided into five folds, with all images from one user only appearing in either the training or test set of that fold. The data splits are the same as those used in [1]. We leveraged five-fold cross-validation and ensured that test data was separate from training data by training on four folds and testing on the last. The performance of our various models were evaluated as the mean accuracy $\pm$ the standard deviation

3 Methodology

The purpose of this section is to explain the methods we used to build our age and gender classification models. We were particularly concerned about reducing training time without sacrificing performance when compared to the model described in [1]. This model served as our baseline, since it has already been thoroughly tested on the dataset and has shown acceptable performance.

3.1 Reducing Parameters in Baseline Model

The first property of the CNN that is tested is the impact of the amount of model parameters on performance. A reduction in the parameters count should decrease the training time and produce a less complex model while minimzing the risk of overfitting. For the gender and age classification tasks, this project proposes a CNN architecture similar to the baseline model. The major difference is that only one-fourth of the filters are being used in each convolution layer and only one-fourth of the hidden neurons in the fully connected layers are being used. The detailed structure of the reduced gender classification model is shown in Figure 1. This reduces the number of trainable parameters from $115,023,873$ to 7 192 065. For age classification, the model architecture is identical apart from the last layer which contains 8 nodes instead of 1, resulting in 1032 trainable parameters in the output layer. In order to prevent overfitting, dropout and local response normalization layers were implemented. Moreover, we used data augmentation to prevent overfitting, which will be discussed later.

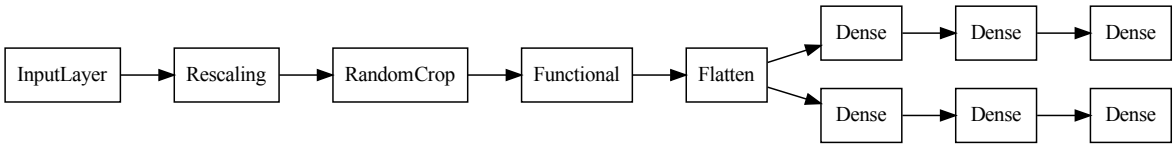
Table 1: Reduced baseline architecture.

Operational Layer		Number of Filters	Size of Each Filter	Stride Value	Padding	Size of Output Image	Parameters
Input image		-	-	-	-	$256 \times 256 \times 3$	-
Random cropping		-	-	-	-	$227 \times 227 \times 3$	-
Rescaling		-	-	-	-	$227 \times 227 \times 3$	-
Convolution Layer	Convolution	24	$7 \times 7 \times 3$	1×1	valid	$221 \times 221 \times 24$	3 552
	ReLU	-	-	-	-	$221 \times 221 \times 24$	
Pooling Layer	Max pooling	1	1×1	2×2	valid	$110 \times 110 \times 24$	-
Normalization	Local Response	-	-	-	-	$110 \times 110 \times 24$	-
Convolution Layer	Convolution	64	$5 \times 5 \times 24$	1×1	valid	$106 \times 106 \times 64$	38 464
	ReLU	-	-	-	-	$106 \times 106 \times 64$	
Pooling Layer	Max pooling	1	1×1	2×2	valid	$52 \times 52 \times 64$	-
Normalization	Local Response	-	-	-	-	$52 \times 52 \times 64$	-
Convolution Layer	Convolution	96	$3 \times 3 \times 64$	1×1	valid	$50 \times 50 \times 96$	55 392
	ReLU	-	-	-	-	$50 \times 50 \times 96$	
Pooling Layer	Max pooling	1	1×1	2×2	valid	$24 \times 24 \times 96$	-
Normalization	Local Response	-	-	-	-	$24 \times 24 \times 96$	-
Inner product layer	Fully connected	-	-	-	-	128	7 078 016
	ReLU	-	-	-	-		
Dropout Layer	Rate = 0.5	-	-	-	-	128	-
Inner product layer	Fully connected	-	-	-	-	128	16 512
	ReLU	-	-	-	-		
Dropout Layer	Rate = 0.5	-	-	-	-	128	-
Output Layer	Fully connected	-	-	-	-	1	129
	Sigmoid	-	-	-	-		

3.2 Multi-task Classification

The next property this project wishes to explore is a CNN's ability to learn multiple tasks at the same time by exploiting a CNN's inherent ability to learn generalized features of an image. If two tasks share some essential features for prediction, a multi-task CNN could drastically reduce the training time while maintaining nearly the same performance compared to two separate models. The multi-task CNN architecture constructed in this project is shown in Figure 3. The functional layer refers to the convolution layers of the reduced CNN shown in Table 1. After the convolutional layers, the model branches into two separate feed forward networks—one for gender classification and the other for age. The feed forward networks are identical to the fully connected layers in Table 1. This architecture allows the two branches to share and learn the same generalized features in the convolutional part of the network, and then individually tune feed forward networks to learn to classify their respective tasks. We may even see improvements in the classification of gender and age by learning age-specific gender features and gender-specific age features.

Figure 3: Multi-task classification architecture.



3.3 Transfer learning

The final aspect of a CNN this project explores, is its ability to share generalalized features learned on a different domain. This project utilizes a pretrained model called Xception [7], which is a large CNN model that consist of 36 convolutional layers. The network has a total of 20 861 480 parameters and has been trained on the ImageNet dataset [4]. This project extends the Xception network to a multi-task CNN by adding branches for both the age and gender classification. These branches are identical to the ones described in the multi-task section, i.e. the Xception network just replaces the functional layer in Figure 1. In order to learn some task specific features, the last 10 layers of the Xception network remain trainable, whereas the rest are frozen from training. In addition, a preprocessing layer is added to the Xception network to match the required input of 299 x 299 sized images. The training of this model is expected to converge rapidly as the base model already contains thoroughly trained parameters.

3.4 Training and experiments

To train the models in this project, NTU SCSE's GPU cluster was utilized. The models were trained using a single GPU core with access to 8 GB of memory. All the models were trained for 100 epochs using a batch size of 128 images. A callback function was used to implement early stopping with a patience of 10. This was done to help minimize the risk of overfitting. Adam was selected as the optimizer of choice for training. The optimizer was initialized with default values i.e $\alpha = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-7}$. The loss function that Adam had to optimize for was binary cross-entropy for gender classification and sparse categorical cross-entropy for age classification. For the multi-task CNNs, the individual loss function of age and gender classification were combined using a weighted sum:

$$L_{multi-task} = \gamma L_{gender} + (1 - \gamma) L_{age}, \tag{1}$$

Where γ is a hyper parameter that weighs the importance of the two loss functions. All the models were trained with the callback functions: Early stopping and Model checkpoint. Model checkpoints were used to retrieve the weights of the best performing model. As mentioned in the prepossessing section, the models are trained using five-fold subjective exclusive cross-validation.

In the training process we apply data augmentation techniques to prevent overfitting and to help the model with generalization. However, since [1] only performed cropping during data augmentation, we have also chosen to use only random cropping. This would allows us to better compare our model performance with the baseline, rather than comparing the best preprocessing method. Random cropping is a technique where we randomly crop the original image to create a set of additional images which the model may train on. This may help our model generalize better by introducing images wherein the face is not entirely visible.

3.5 Overview of Models

To give a brief overview of the size of the models, the following table presents the amount of trainable and non trainable parameters in each of the models as well as the functional layer (i.e the convolutional layers) used for all the models.

Model	Trainable parameters	Non trainable parameters	Total parameters
Gender	7 192 065	0	7 192 065
Age	7 192 968	0	7 192 968
multi-task	14 287 625	0	14 287 625
Transfer learning	31 220 873	15 365 160	46 586 033
Functional layer	97408	0	97408

Table 2: Model parameters counts and status of trainability

Table 2 shows that only a small part of the network parameters are used for feature extraction/generation (functional layer), and the majority of the parameters are used to make the prediction based on these features. The output from the convolutional part of the transfer learning model is approximately twice as large as that of the reduced models, but the feed forward networks are built with the same structure and parameters. This leads to almost twice as many parameters in the fully connected layers when compared to the other models.

4 Results

The results of the experiments will be presented in this section.

4.1 Models with Reduced Parameters

The learning curves for the separate Gender and Age classification models with reduced parameters are presented in Figures 4 and 5. All curves converges after about 60 epochs. We observe that the first fold yields a higher accuracy than the rest. We also note that the standard deviation of age accuracies across folds is higher than the for gender accuracies. This is probably due to the fact that there are more age classes and that age in general may be harder to predict than gender.

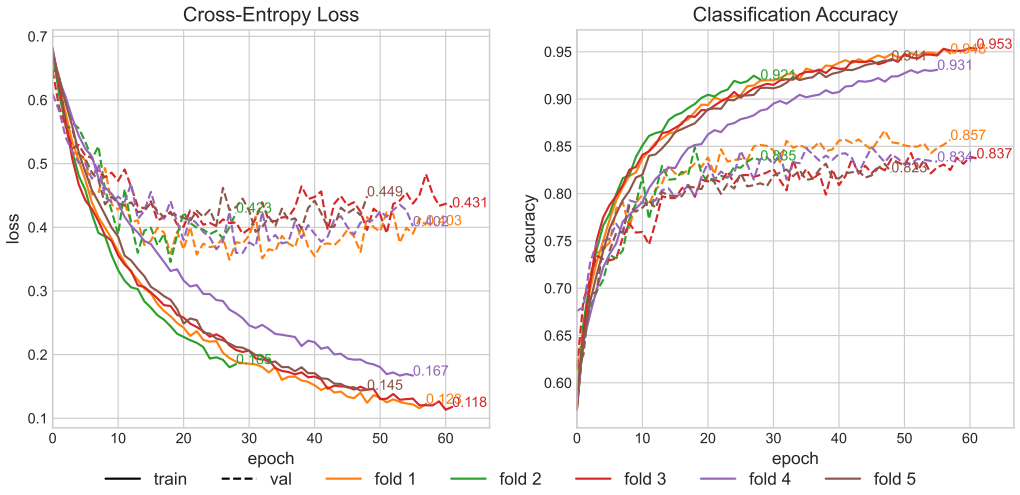


Figure 4: Learning curves for Gender classification

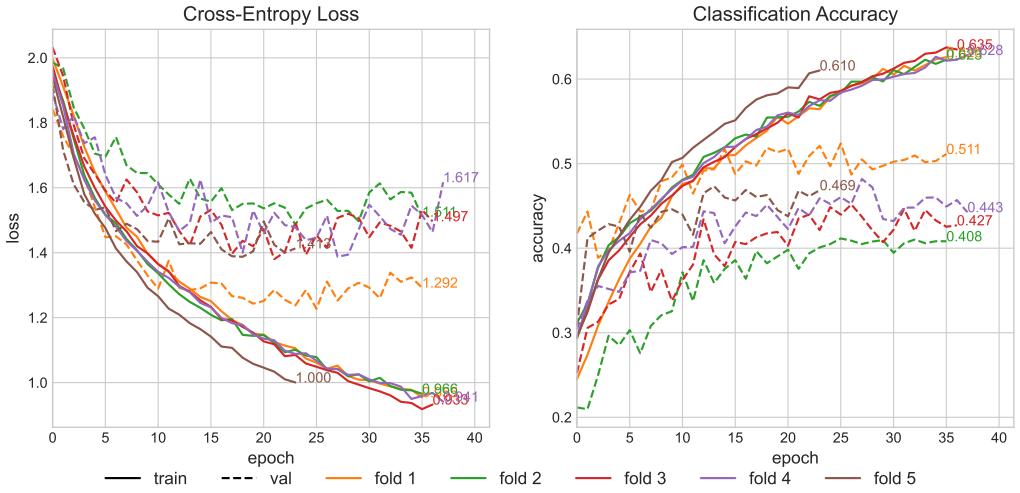


Figure 5: Learning curves for Age classification

4.2 Multitask model

The learning curves for the multitask classification model are presented in Figure 6. All curves converges after about 25 epochs. The curves looks a lot like the curves for the separate models, but with a little larger variation across folds. This indicates that the multitask model performs just as good as the separate models.

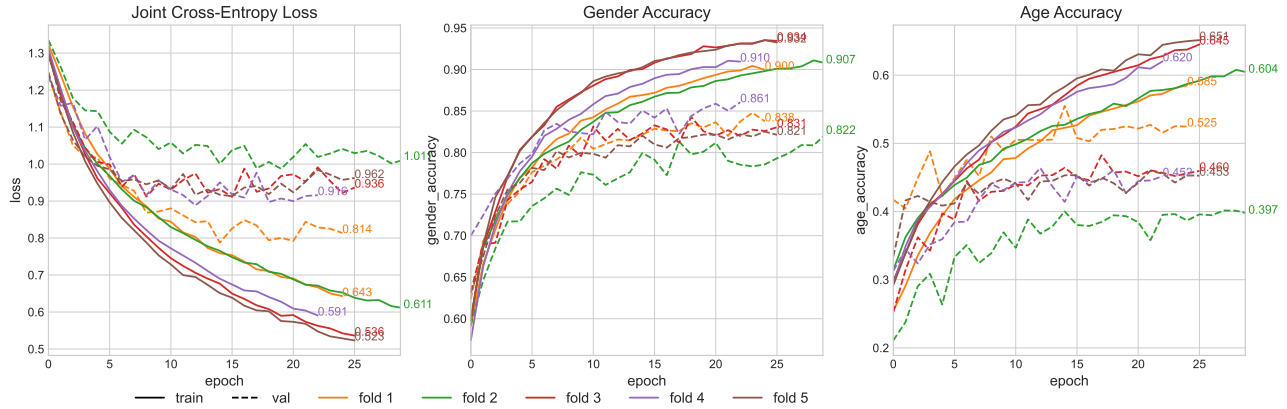


Figure 6: Learning curves for Multitask classification

4.3 Transfer learning Model

The learning curves for the multitask transfer learning model are presented in Figure 7. These curves looks different, as all curves converges after only a couple epochs. If we look at the loss function we see that the model starts after about 4 epochs. This shows that the weights in the Xception model are well trained for the beginning. We also note that the model performance in general is better than the former models.

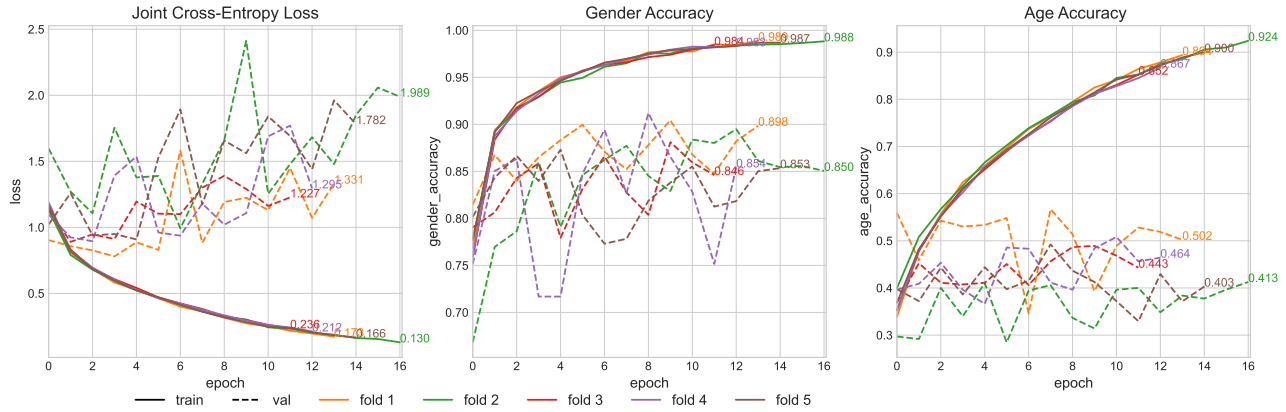


Figure 7: Learning curves for transfer learning classification

4.4 Model Performances

The mean accuracies and standard deviations across the five folds are listed in Table 3. We see that the performance for the seperate models and the multitask model are slightly below the baseline performance. However, the transfer learning model on the other hand outperforms the baseline model in gender classification and nearly the same for age classification.

Model	Gender accuracy	Age accuracy	Avg training epochs
Baseline*	86.8 ± 1.4	50.7 ± 5.1	-
Gender Classification	84.8 ± 1.1	-	50.8
Age Classification	-	46.8 ± 3.7	34.2
Multitask Classification	84.0 ± 1.4	47.3 ± 4.9	26.0
Transfer Classification	89.3 ± 1.4	49.4 ± 4.9	14.2

*Results from paper [1]

Table 3: Model performances based on mean and standard error of cross-validation

5 Discussion of Results

This section analyzes the results of the models and compares them to the baseline model. Additionally, we propose various ideas and suggestions for interesting research topics that could build on top of the findings in this project.

5.1 Reduced Baseline Models

The first models that will be discussed are the reduced age and gender classification models. From Table 3, its apparent that both the reduced age and gender classification models are performing well and are able to make good predictions within their classification domains. The two models seem to be on par with the baseline model as the span of their respective confidence intervals is overlapping. However, it does appear that the baseline is slightly favored if only the mean accuracy is considered. This finding illustrates that a model with far lesser parameters is able to perform well in this domain of classification. Another takeaway is that the age and gender classification only relies on a limited amount of features, and excess convolution layers simply find lesser meaningful features. This could explain the reason why the reduced models are able to achieve this level of performance while only relying on a sliver of parameters compared to the baseline model.

The main goal of these models are to reduce training time while maintaining good performance. The baseline has a approximate training time of 4 hours using the Amazon GPU machine [1]. It may be difficult to compare the training time of our models to that of the baseline since they were trained in different computing environments having access to separate resource profiles. Our models were trained using limited computing power only having access to a single GPU. This lead to a training time in the range of 45-90 minutes for each

fold for the two models. Since the training time is highly correlated with the amount of model parameters, it is safe to assume that the reduced models train faster. This is because the reduced models only have a fraction of the baseline model's parameter count. Even though the total training time of the models reach similar times as the baseline, this is due to computational resources and not the model architectures.

5.2 Multi-task Model

Table 3 shows that the mutli-task model performs equivalent to both of the reduced models in the respective classification tasks. The multi-task model obtains a slightly lower mean accuracy for gender classification and slightly higher accuracy for age classification. However, the confidence intervals are overlapping for both tasks showing that there is no significant difference in performance. These results indicate that the two classification tasks rely on similar features in order to make their predictions. However, since the multi-task model doesn't improve the performance it might indicate that there are no age-gender dependent features e.g a feature that helps classify men in (38,43) range. Although an advantage of the multi-task model is that it can save training time. In this project, the reduced individual models had an average training time of $80 - 90$ seconds for each epoch and averaged 50.8 and 34.2 epochs for each fold for gender and age classification respectively (Table 3). Similarly, the multi-task model used $90 - 100$ seconds for each epoch, but in contrast, only trained for around 26 epochs for each fold. This means that the multi-task model is able to drastically reduce the training time, while maintaining equal performance. For a multi-task network to be effective, it must be applied to tasks that share essential features, which may be hard to determine before having done any training. Another difficulty with the multi-task setup is that there is no heuristic to find the optimal place to partition the shared network nor the best value of γ. Furthermore, there is no interaction between the two branches after the split, which means that the task only shares "knowledge" at the higher level layers of the network that are more general. One could imagine that the fully connected layers could reveal some useful interdependencies between the tasks.

5.3 Transfer Learning

Table 3 shows that the transfer learning model performs 2.5% better than the baseline in gender classification. The performance in age classification is a little lower than the baseline although the confidence intervals overlap. This may indicate that only gender classification benefits from the multi-task structure. It is not surprising that this model have better performance than the baseline. The Xception is a huge network of 36 convolutional layers which is much deeper than the baseline model. This also means that much more complex relationships within the data could be learned. In spite of the transfer learning model being more advanced, the training time is much lesser than that of the baseline model. This is due to the pre-trained parameters, that we do not need to update during training. These parameters are not updated as we assume the Xception network has already been trained to extract useful features for our classification task. Based on our results, this assumption seems to hold.

5.4 Ideas for Future Work

In order to develop models with even higher performances there are a few techniques that may be considered. Firstly, hyperparameter tuning could be one such technique. By changing some parameters in the model architecture such as the filter sizes, strides or number of hidden units in the fully connected layers, the performance may improve. In our case, hyperparameter tuning may have been infeasible given our limited computing resources. Secondly, adjusting the learning parameters like batch size, learning rate, and different loss weights (γ factors) may yield to better convergence. Thirdly, implementing different data augmentation techniques may help the model generalize better. For instance, distortion, blur, flipping, and scaling are all techniques which may be considered in addition to random cropping. This project did not consider these techniques for the sake of comparing with the baseline model which only implemented random crop. Lastly, another task like human race detection could be considered in addition to age and gender to learn task-specific characteristics that may help with classification on each task.

6 Conclusion

To summarize, we have developed a multi-task classifier for age and gender classification that matches the performance of earlier research models. By using transfer learning, we were able to predict gender with an

accuracy of 89.3% and age with a precision of 49.4%. We also reduced the amount of parameters that need to be trained, without signifacntly compromising on accuracy. Our results showed that the multi-task classification did not perform better or worse than the separate models, making it much more efficient to implement the multi-task structure with shared convolutional layers.

References

[1] G. Levi and T. Hassner, “Age and gender classification using convolutional neural networks.” in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops, 2015

[2] O’Shea, Keiron, and Ryan Nash. "An introduction to convolutional neural networks." arXiv preprint arXiv:1511.08458 (2015).

[3] Kim, T.S., Sohn, S.Y. Multitask learning for health condition identification and remaining useful life prediction: deep convolutional neural network approach. J Intell Manuf 32, 2169–2179 (2021). https://doi.org/10.1007/s10845-020-01630-w

[4] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248–255).

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

[6] Kasthurirangan Gopalakrishnan, Siddhartha K. Khaitan, Alok Choudhary, Ankit Agrawal, Deep Convolutional Neural Networks with transfer learning for computer vision-based data-driven pavement distress detection, Construction and Building Materials

[7] Chollet F., 2017, Xception: Deep Learning with Depthwise Separable Convolutions, Google Inc.