Machine Learning 2: Coursework 1

Age and Gender Prediction using CNN

Contributions:

ID	Contribution
239684909	50%
199153693	50%

Do all members agree with the above contributions? Yes.

Links to two models:

Model A (defined and trained from scratch): **age_gender_A.h5**
Model B (pre-trained CNN model fine-tuned): **age_gender_B.h5**

Table of Contents

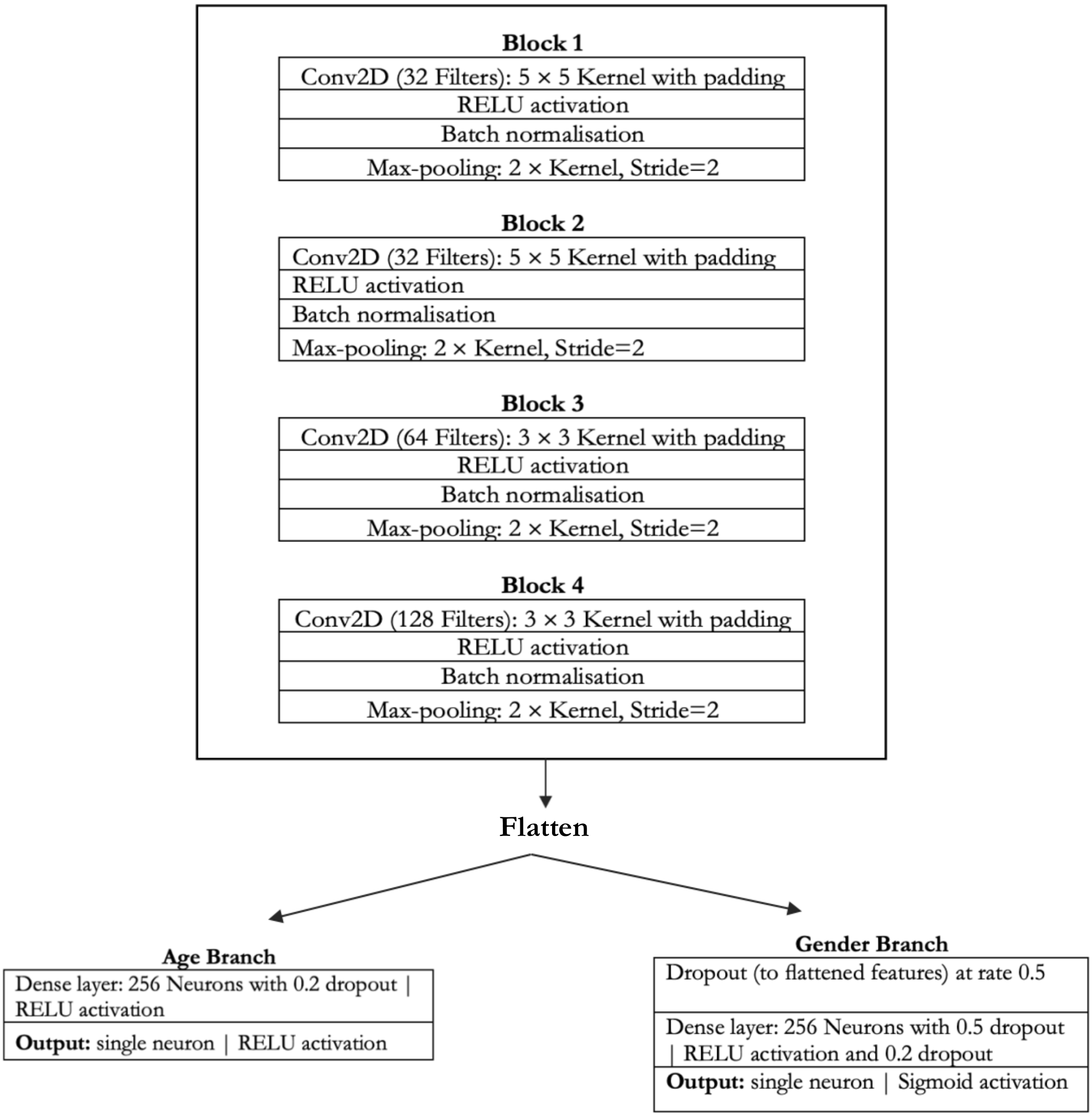
Section 1: Introduction

This assignment builds deep learning models A and B using Convolutional Neural Networks for image classification. Both models are trained and validated on a subset of the UTKFace dataset to predict the age and gender of human face images. All images are coloured and cropped into $128 \cdot 128$ pixels to optimise the central alignment of the face. Model A is designed and trained from scratch, and Model B applies transfer learning to the pre-trained VGG architecture. Classification tasks for both outputs are achieved simultaneously by splitting the model into two branches. Extensive model architecture exploration and hyperparameter optimisation were completed over the course of three weeks, upon which the model with the best performance is presented in this report.

Section 2: My Own CNN (Model A)

Model architecture:

As shown in the figure below, four sequential convolutional 'blocks' are shared between the age and gender branches, each beginning with a convolutional layer, then transformed using 'RELU' activation. This is followed by batch normalisation and max-pooling.



Block 1

Conv2D (32 Filters): 5 × 5 Kernel with padding
RELU activation
Batch normalisation
Max-pooling: 2 × Kernel, Stride=2

Block 2

Conv2D (32 Filters): 5 × 5 Kernel with padding
RELU activation
Batch normalisation
Max-pooling: 2 × Kernel, Stride=2

Block 3

Conv2D (64 Filters): 3 × 3 Kernel with padding
RELU activation
Batch normalisation
Max-pooling: 2 × Kernel, Stride=2

Block 4

Conv2D (128 Filters): 3 × 3 Kernel with padding
RELU activation
Batch normalisation
Max-pooling: 2 × Kernel, Stride=2

Flatten

Age Branch

Dense layer: 256 Neurons with 0.2 dropout | RELU activation
Output: single neuron | RELU activation

Gender Branch

Dropout (to flattened features) at rate 0.5
Dense layer: 256 Neurons with 0.5 dropout | RELU activation and 0.2 dropout
Output: single neuron | Sigmoid activation

Various model configurations were explored during development, including:

- Number of convolutional layers and their filters.
- Stacking multiple convolutional layers before pooling.
- Relative position of Batch Normalisation layers to the pooling layers (before/after).
- Kernel and stride sizes for all convolutional layers and max-pooling layers.
- Convolutional layers with or without padding.

It was discovered that simple architectures generally yielded better performance, given the small input sizes and homogeneity of image contents. Increasing the complexity of the convolutional base led to significantly more parameters (above 10 million), which yielded highly unstable outputs. Hence the quantity of filters for convolutional layers started from as little as 32 and were capped at 128. Under human perception, age and gender often manifest in subtle ways across the face rather than locally, in ways such as the presence of a moustache, wrinkles, curve of the jawline, and shape of the eyebrow. Mindful of such characteristics, we implemented a generous filter size of 5·5 in the first two convolutional layers, such that the receptive field is large enough to capture broader abstract patterns across the face and between facial features. Once such preliminary features are extracted, a reduced filter size of 3·3 is applied at deeper layers of the architecture (blocks 3 and 4) in an attempt to capture finer details to facilitate complex feature engineering.

The convolutional base extracts a reduced feature map of 8·8, which is flattened as input to the fully connected layers. Here, the model splits into the age and the gender branches. During initial training, the gender branch showed a tendency of faster converge, so we added additional dropouts to slow it down so that both outputs were synthsised.

Data pre-processing:

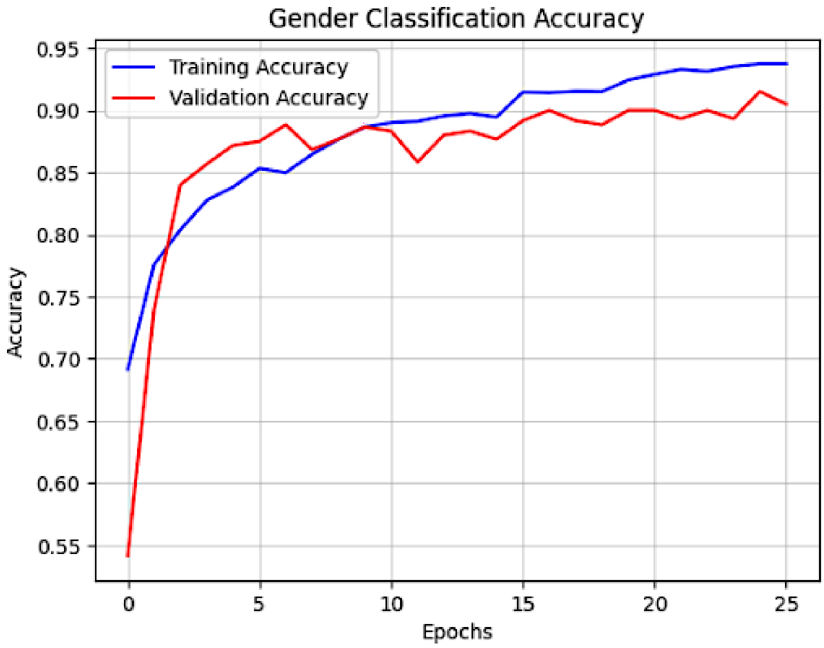
5000 images were shuffled and split into training and validation sets, with 4400 images (88%) for training and 600 (12%) for validation.

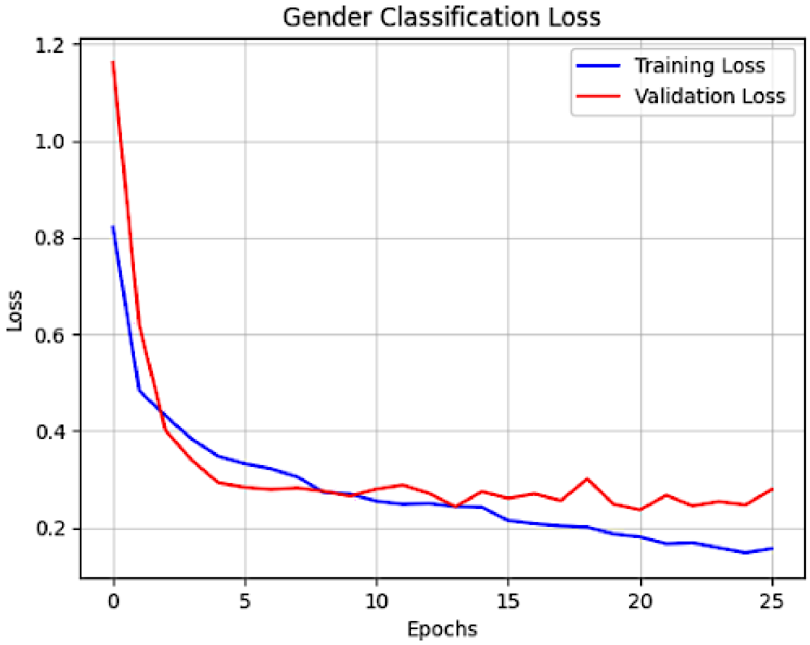
Upon initial examination, the majority of images have the subject's face centralised, occupying approximately 70-80% of the frame. To enrich samples, we explored several data augmentation techniques, e.g., flips, shifts, rotation, and adjusting contrast and saturation, concluding that minimal augmentation performs best. Only the horizontal flip is adopted in the final model. Since all image inputs are cropped for optimal alignment, any liberal modifications can distort the relative position of facial features, inadvertently introducing semantic differences between the training and validation sets.

Training:

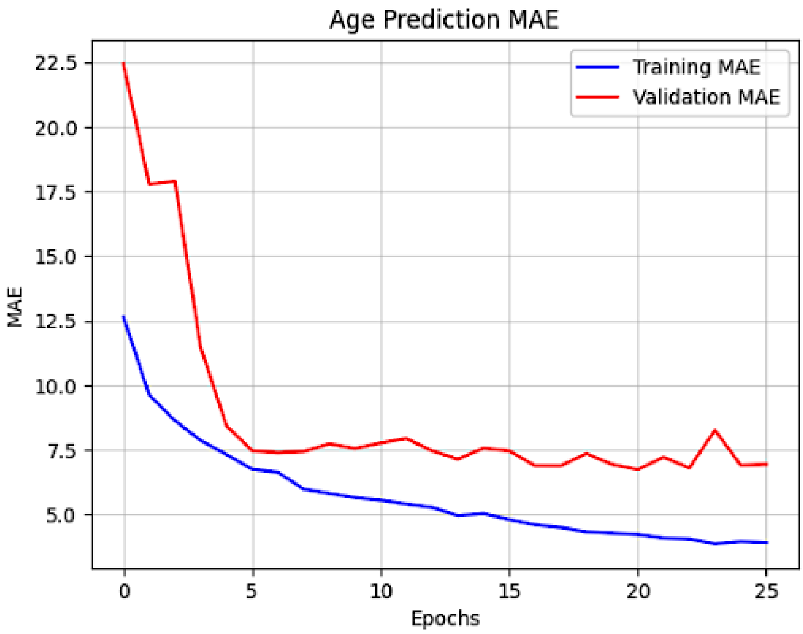
After exploring configurations of the SGD optimiser, the default Adam Optimizer was adopted for its built-in adaptive learning rate, which facilitated faster and more stable convergence (Kingma and Ba, 2014). To prevent overfitting, we applied Early-Stopping in call-back with patience=8 and min-delta=0.5. The "restore_best_weights" parameter was set to True, so weights for the best epoch are saved as the final model. The loss was measured in terms of MSE and MAE for age prediction as values are continuous and binary cross-entropy for binary gender classification. Model performance was measured by validation loss across age and gender. Batch size was set at 32 after trials of 64 and 128, both of which produced slower convergence, presumably due to fewer steps (less frequent updates to model weights) per training epoch (Kandel and Castelli, 2020).

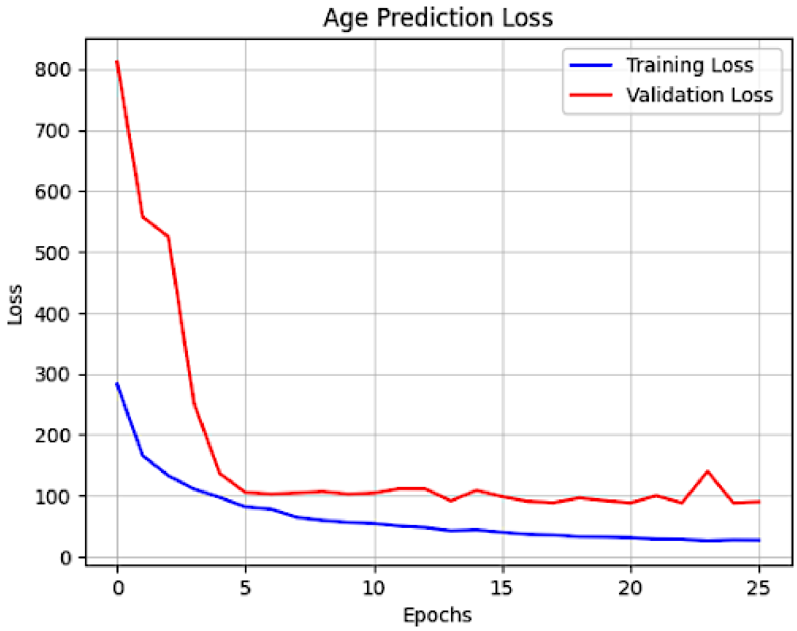
After 25 training epochs, the best weights are restored from Epoch 18. From the training curves below we can see that gender validation loss stabilised around Epoch 20 without recognisable signs of overfitting.





Age validation MAE also stabilised around epoch 16-20, suggesting Epoch 18 is a reasonable checkpoint to extract optimal weights.



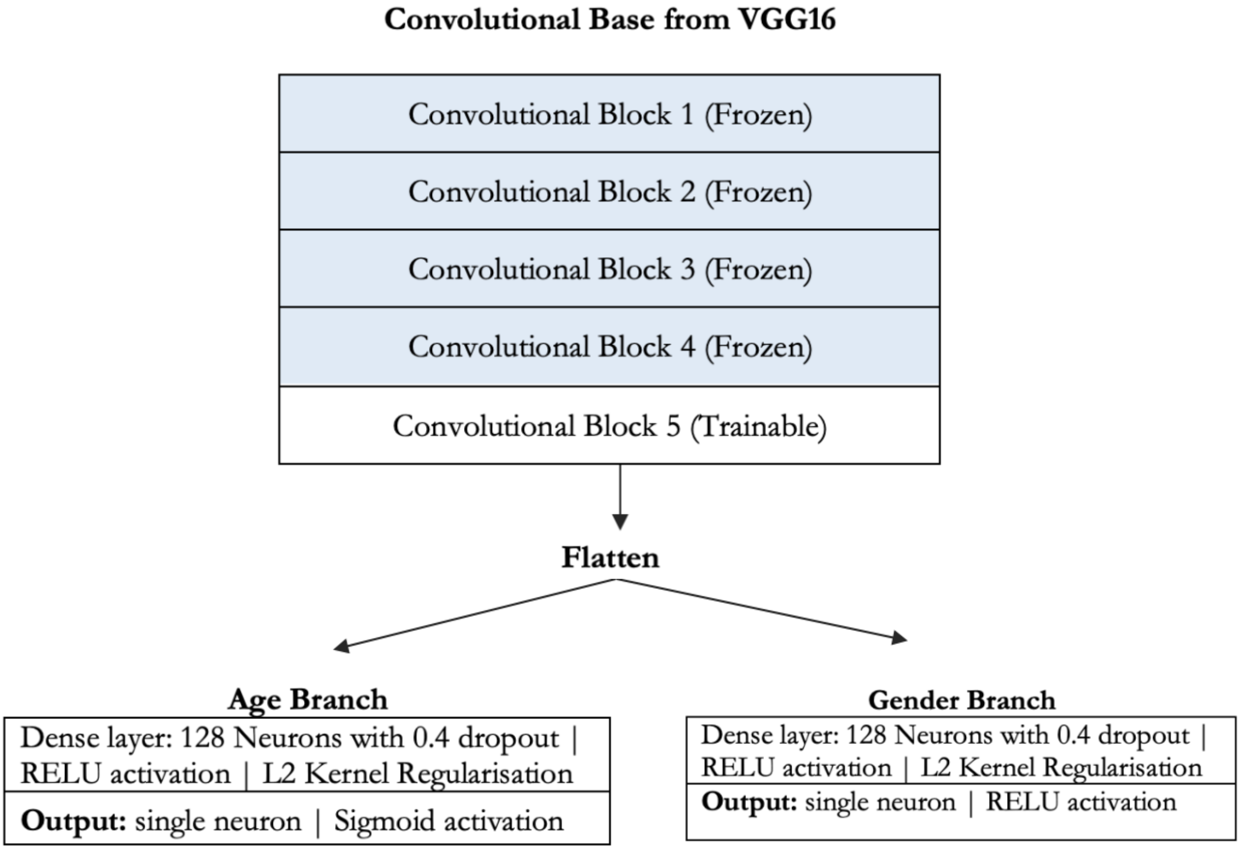


Model A training and validation performance metrics:

	Training Metrics		Validation Metrics	
Age	MSE (Loss):	26.7059	MSE (Loss):	89.6484
	MAE:	3.9122	MAE	6.9323
Gender	Binary Cross-entropy (Loss):	0.1563	Binary Cross-entropy (Loss):	0.2786
	Accuracy:	0.9373	Accuracy:	0.9050

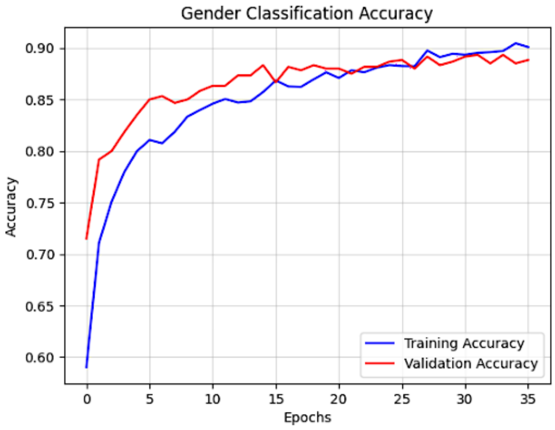
Section 3: Pre-trained CNN (Model B)

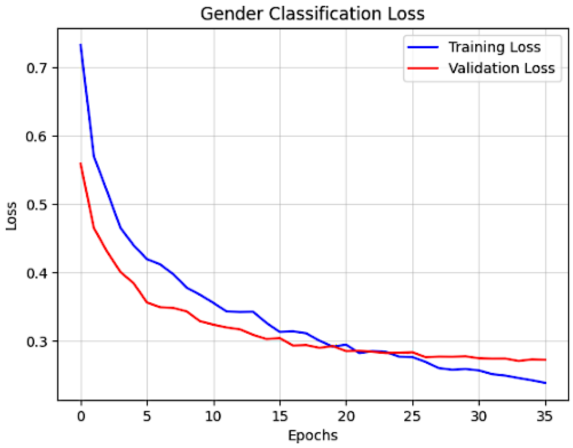
The pre-trained VGG16 model consists of 5 convolutional blocks followed by dense layers; each block has multiple convolutional layers with max-pooling. Recognising our dataset structurally differs from ImageNet, on which the model is initially trained, we take only the convolutional base architecture, freezing weights of the first three blocks for generic feature extraction, leaving weights of the last two blocks trainable. This outputs a 4·4 feature map, which we flatten and input into dense layers as two branches, just like model A.

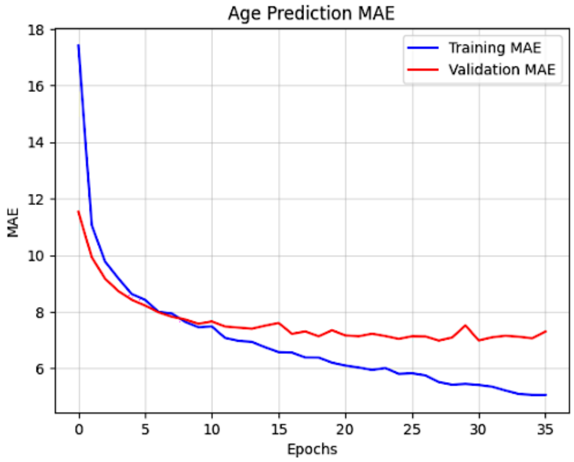

Convolutional Base from VGG16
Convolutional Block 1 (Frozen)
Convolutional Block 2 (Frozen)
Convolutional Block 3 (Frozen)
Convolutional Block 4 (Frozen)
Convolutional Block 5 (Trainable)
Flatten
Age Branch
Dense layer: 128 Neurons with 0.4 dropout | RELU activation | L2 Kernel Regularisation
Output: single neuron | Sigmoid activation
Gender Branch
Dense layer: 128 Neurons with 0.4 dropout | RELU activation | L2 Kernel Regularisation
Output: single neuron | RELU activation

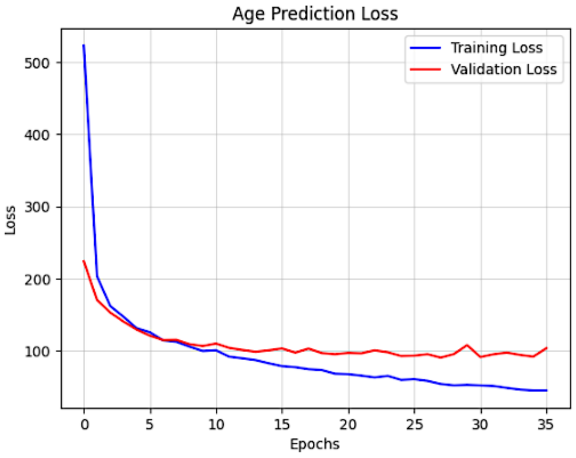
In addition to regularisation methods in Model A, kernel weights are regularised with L2 penalty for every dense layer. It was discovered that due to the increased complexity of the model, a reduced learning rate of 0.0001 significantly improved model performance and mitigated overfitting (Wilson and Martinez, 2001).

The same data pre-processing pipeline was adopted. The training curves for both outputs are shown below. After 36 training epochs, the best weights are restored from Epoch 28. From the graph we see that gender validation loss stabilised around epoch 27 without recognisable signs of overfitting. Age validation MAE also stabilised around epoch 27, suggesting that Epoch 28 is a reasonable checkpoint to extract optimal weights.









Model B training and validation performance metrics:

	Training Metrics		Validation Metrics	
Age	MSE (Loss):	44.8062	MSE (Loss):	103.5103
	MAE:	5.0649	MAE	7.3073
Gender	Binary Cross-entropy (Loss):	0.2386	Binary Cross-entropy (Loss):	0.2723
	Accuracy:	0.9009	Accuracy:	0.8883

Section 4: Summary and Discussion

In summary, the limitation to a distinct dataset with analogously structured images for this task prompted us to approach the model design from an intuitive and realistic standpoint, i.e., thinking about what actual features we hope to be identified by the filters, their relative sizes to the image frame, and how to engineer the convolutional layers best to do so. This particularly helped our design of Model A. Compared to model B, its outperformance reflected the relative rigidity we experienced when applying transfer learning to VGG16, a network with 134.8 million parameters and 16 layers trained on 14 million images. Fine-tuning this complex architecture to a small dataset for Model B required us to balance regularisation techniques with efficient learning, which gave us the opportunity to experiment with how different combinations of regularisation and hyperparameter configurations interact with one another. We eventually found a decreased learning rate and L2 kernel regularisation to be the compromise, which inevitably slowed down the learning process but produced a smoother learning curve, reaching a lower validation loss.

Additionally, to synthesise the learning trajectory between two outputs, we tried different architectures for each branch and did not discover noticeable improvements when splitting the branches earlier at the convolutional base. Our eventual solution was to increase regularisation for the branch that displayed overfitting relative to the other. This also confirmed our intuition that the features required to predict both outputs are highly similar, as both can be seen as meta-features that are meta-summarised from low-level facial observations.

References

Kandel, I., & Castelli, M. (2020) 'The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset', *ICT Express*, 6(4), pp. 312-315. https://doi.org/10.1016/j.icte.2020.04.010.

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations*, San Diego. Retrieved from https://doi.org/10.48550/arXiv.1412.6980

Mascarenhas, S., & Agarwal, M. (2021) 'A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification,' in *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*, Bengaluru, India, pp. 96-99. doi: 10.1109/CENTCON52345.2021.9687944.

Wilson, D. R., & Martinez, T. R. (2001) 'The Need for Small Learning Rates on Large Problems', in *Proceedings of the 2001 International Joint Conference on Neural Networks (IJCNN'01)*, pp. 115-119.