JST: Smart Systems and Devices
Volume 35, Issue 3, September 2025, 008-015

Multi-task Learning for Age, Gender, and Emotion Recognition on Edge Processing

Ha Xuan Nguyen[1,2,*], An Dao[2], Duc Quang Tran[2], Tuan Minh Dang[2, 3]
[1]*Hanoi University of Science and Technology, Ha Noi, Vietnam*
[2]*CMC Applied Technology Institute, CMC Corporation, Ha Noi, Viet Nam*
[3]*CMC University, CMC Corporation, Ha Noi, Viet Nam*
[*] *Corresponding author email: ha.nguyenxuan@hust.edu.vn*

Abstract

In this work, a multi-task learning model for age, gender, and emotion recognition on edge processing is developed. The multi-task model is based on the backbone of MobileNetV2 in which the three last layers are customized to have three outputs for age, gender, and emotion. The model was trained and tested on a dataset which is the combination of the well-known dataset, namely Internet Movie Database (IMDB) and our self-collected dataset. The trained model is then optimized and quantized to be implemented on Neural Processing Unit (NPU) of the chip RK3588 from Rockchip on Orange PI plus hardware platform. Experimental evaluation on several testcase was performed. It is known that the multi-task model outputs prediction accuracy as high as single-task model while significantly reducing computational processing requirements. On Orange PI platform, the highest prediction accuracy for age, gender and emotion are 3.485 MAE, 98.281%, and 93.917%, respectively. The computational performance reaches 285.7 frames per second as the highest. These results have a high potential for many practical applications on edge devices.

Keywords: Age, gender, and emotion recognition, multi-task learning, NPUs, edge processing.

1. Introduction

The rapid development and advancement of deep learning and computing hardware have shown many advantages for the image processing problem. The age, gender, and emotion recognition issue has received much attention in recent years due to its very high potential for practical applications in customer-experiencing systems, public security surveillance systems, modern conversational robots, and electronic commercial transactions [1, 2].

The development of deep convolution neural networks has shown many remarkable achievements in predicting age, gender, and emotion both in accuracy and robustness. However, the simultaneous execution of three neural networks for the three recognition tasks on a single computer which normally has very limited hardware resources, would result in a computational bottleneck for many real-time applications. A solution to overcome this problem is the use of a single multi-task network for all recognition tasks. The basis idea of multi-task learning is parameter sharing among recognition tasks. The multi-task model is trained with data from three tasks simultaneously. The first layers of the network are used to extract high level features of the input data. Then, these features undergo several convolution layers to extract low-level features before putting into classification layers for each specific task.

The use of multi-task architecture has advantages of computational efficiency and hardware utilization in comparison to the use of independent neural networks for each task. Furthermore, the accuracy of the multi-task model would not be reduced even in some cases higher than that of single-task model. In fact, if the recognition tasks are significantly related, the representation of these tasks will be quite like each other. Thus, the computation of all single representations can be combined within one network only. For example, in our case, the age, gender, and emotion predictions have similar characteristics that they use low-level facial features for classification. These features can be shared among the three recognition tasks in a multi-task model. The multi-task model would even outperform the single-task model since it can exploit knowledge sharing among the different tasks. Although multi-task learning has many advantages, it is still a holistic research topic focusing on the improvement of recognition accuracy and computational efficiency [3].

There have been several investigations related to multi-task learning using facial features [3-9]. In [4], Sang *et. al* have introduced a multi-task learning scheme for age, smile, and emotion recognition. The model consists of a single convolution neural network with three branches in the output that produce predictions of age, smile, and emotion. The model is

ISSN 2734-9373
https://doi.org/10.51316/jst.184.ssad.2025.35.3.2
Received: Nov 3, 2024; revised: Feb 21, 2025;
accepted: May 16, 2025

compared with the best single-task one. In a similar approach, Vu *et. al.* [5] have proposed a multi-task network for age and gender prediction. In another approach [6, 7], the authors determine the relationship between recognition tasks and choose one task as the main task. They trained the network for this task and then made its features available to the networks of the other tasks. Although this task has the advantage of reducing effort to prepare training dataset, it would also be not a general case since it prevents tasks learning from each other. A better approach is to build a joint training network for all tasks and use a combined loss function for the prediction of each task in the last layers [3, 8, 9].

It is known that multi-task learning would be a good approach for problems with multi-task which corelate together. The development of a multi-task model which has high accuracy as well as computational efficiency is an urgent topic. Thus, in this work, a multi-task learning model for age, gender, and emotion recognition is developed. The multi-task model is based on the backbone of MobileNetV2 [10] in which the three last layers are customized to have three outputs for age, gender, and emotion. The model was trained and tested on a dataset which is the combination of the well-known dataset, namely IMDB [11, 12] and our self-collected dataset. The trained model is then optimized and quantized to be implemented on NPU of the chip RK3588 from Rockchip [13] on Orange PI plus hardware platform [14]. Experimental evaluation on several testcase is performed. The model's accuracy and computational efficiency will be thoroughly evaluated and analyzed. The contribution of this work is a new combined model for recognizing simultaneously age, gender, and emotion. Our model is not only computationally efficient, but also accurate and robust since it is trained by a diverse dataset.

2. Multi-Task Learning Architecture and Transfer Learning Process

The purpose of multi-task learning is hard parameter sharing in which all convolution layers are shared among the tasks with the target to minimize the processing time and the required hardware utilization. Our proposed network architecture is shown in Fig. 1, which is based on the backbone of MobileNetV2 [10].

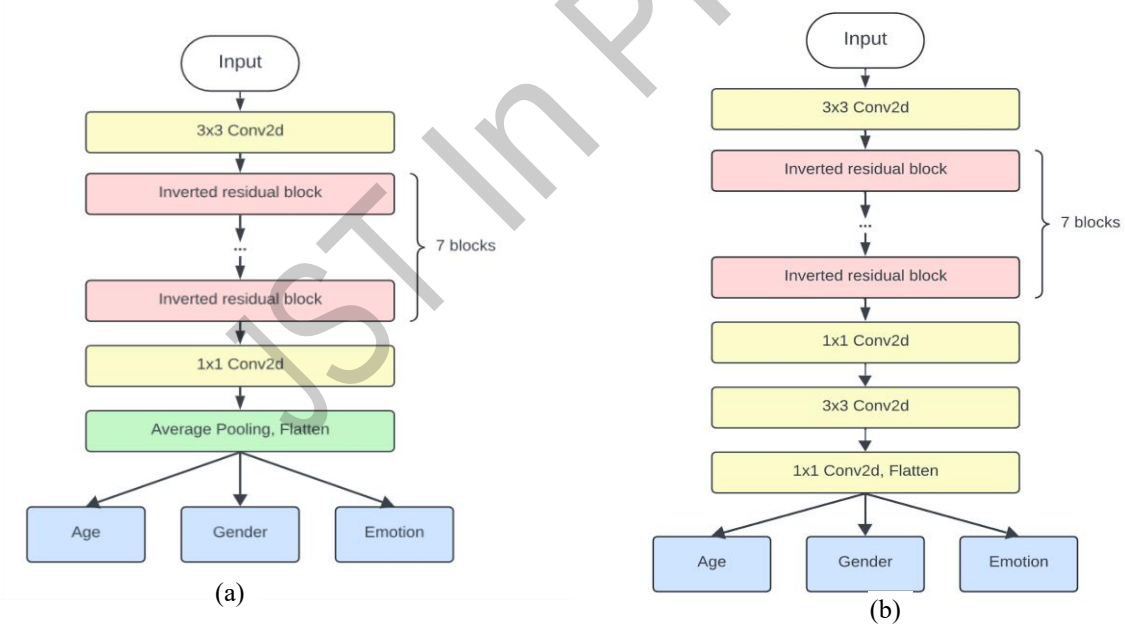

Fig. 1. Multi-task CNN architecture of the proposed system: a) MobileNetV2 backbone (left), b) customized MobileNetV2 (right)

It is known that the MobileNetV2 has a very lightweight structure while remaining sufficient accuracy. Thus, this network architecture would ensure computational efficiency when deploying on-edge devices with limited hardware resources. Detailed information of the backbone is listed in Table 1. The input size is 112×112 pixels. The convolution layers are defined by the filter size, for example 3×3, the type (conv2D or Inverted Residual Block (IRB) and number of the feature naps. The first convolution layer (conv1) is used to extract high-level features from input images. After that, high-level features will undergo seven blocks to extract low-level features. Each block has depth-wise separable convolution and residual connections. Instead of bottlenecks architecture, each block of MobileNetV2

is designed to have an expansion in the middle using IRB. Consequently, features maps are put into a pooling layer to extract important features $x \in \mathbb{R}^{1\times1792}$ before putting into the classification for each recognition task. Finally, features are put into a fully connected layer to predict the result for each task. In addition, for the task of age predictions, the activation function ReLU is used to ensure that the age has always a positive value.

With the purpose of deploying the model on edge devices, some layers of the model may be modified to operate accurately and efficiently-computationally. In this work, a hardware platform, Orange PI plus [14], is used. The platform has an artificial intelligent chip from Rockchip company. To successfully deploy the model on chip, we have to use the toolkit RKNN (Rockchip Neural Network) [15] of Rockchip for the model conversion and optimization. The chip has CPU, GPU and NPU. It is known that some neural layers of the model cannot be run on NPU. In this case, the calculation of these layers is implemented by CPU. The switching data between CPU and NPU for the calculation will slow down the processing. In other cases, although the neural layers can be run on NPU, the computation can be inefficient. For these reasons, we must customize the model so that it can be implemented on NPU efficiently. A customized version of MobileNetV2 backbone is proposed as illustrated in Fig. 1b and listed in Table 1. Compared to the original MobileNetV2, we replaced the pooling layer by two convolution layers. This replacement is empirically done by deep understanding the computational architecture of the chip and its corresponding toolkits.

Table 1. Architecture of the multi-task learning network based on MobileNetV2

Layer	MobileNetV2	MobileNetV2 customized	Output
input	$112 \times 112 \times 3$		$n \times 112 \times 112 \times 3$
conv1	$3 \times 3\ Conv2D, 48$		$n \times 56 \times 56 \times 48$
block1	$(IRB, 24) \times 1$		$n \times 56 \times 56 \times 24$
block2	$(IRB, 32) \times 2$		$n \times 28 \times 28 \times 32$
block3	$(IRB, 48) \times 3$		$n \times 14 \times 14 \times 48$
block4	$(IRB, 88) \times 4$		$n \times 7 \times 7 \times 88$
block5	$(IRB, 136) \times 3$		$n \times 7 \times 7 \times 136$
block6	$(IRB, 224) \times 3$		$n \times 4 \times 4 \times 224$
block7	$(IRB, 448) \times 1$		$n \times 4 \times 4 \times 448$
Conv2	$1 \times 1\ Conv2D, 1792$		$n \times 4 \times 4 \times 1792$
Pre_task	$Pooling\ 1792$	$3 \times 3\ Conv2D, 1024$ $1 \times 1\ Conv2D, 1024$	$n \times 1 \times 1 \times 1792$
Flatten			$n \times 1792$
Age head	$Linear(1792,1)$ $ReLU$		$n \times 1$
Gender head	$Linear(1792,1)$ $Sigmoid$		$n \times 1$
Emotion head	$Linear(1792,3)$		$n \times 3$

Note: IRB stands for Inverted Residuals Block; n stands for batch size.

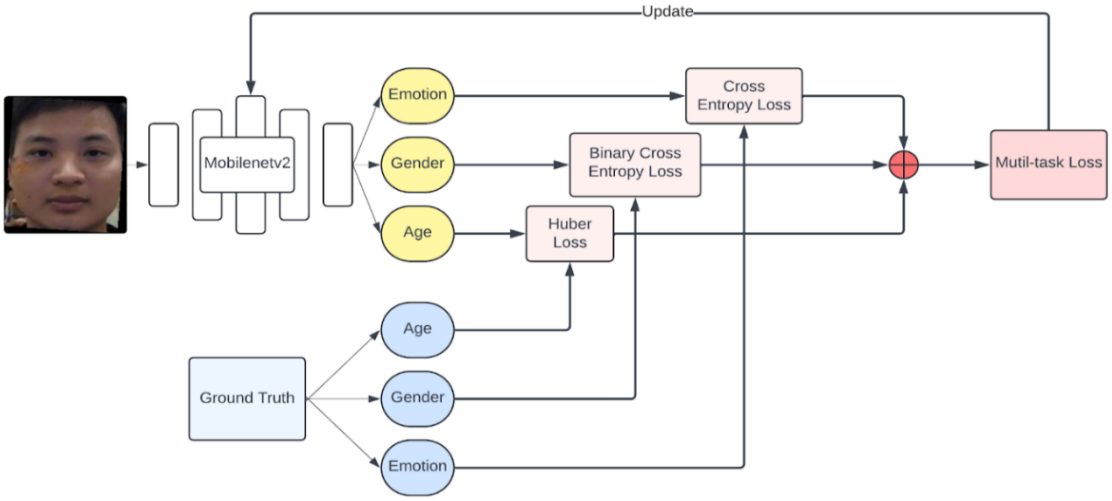

Fig. 2. Description of loss function in the training process

The loss function for the multi-task learning process is weighted loss which is defined as follows:

$$L_{Total} = \alpha_1 L_{Age} + \alpha_2 L_{Gender} + \alpha_3 L_{Emotion} \quad (1)$$

where:

L_{Age} is Huber loss [16],

L_{Gender} is Binary Cross Entropy loss,

$L_{Emotion}$ is Cross Entropy loss,

α_1, α_2, and α_3 are weights which are empirically chosen to $0.08, 0.95$, and 0.85 respectively. Detailed descriptions of the loss function are depicted in Fig. 2.

For the training and testing processes, the dataset IMDB-clean [11, 12] is used. In addition, we diversify the dataset by adding our self-collected dataset. Statistics of the dataset are listed in Table 2. Since the IMDB-clean dataset has only labels for age prediction and gender classification, we add the emotion label for the dataset. To make the emotion label, we first use the model EfficientNetB0 [17] for coarse label and then we post-check manually. Fig. 3 shows examples of the used dataset.

Table 2. Statistics of the testing and training datasets

Purpose	From IMDB-clean	Our self-collected
Training	183879 images	108265 images
Validation	45970 images	10774 images
Testing	56086 images	0

Based on the dataset and the proposed backbone, the model was trained from scratch using Nvidia GPU Tesla V100-PCIE. We use a batch size of 1024 and perform 50 epoch. During the training process, data augmentation techniques like random horizontal flip, random affine, color jitter, random erasing, and random grayscale are applied.

The trained model is saved to ONNX (Open Neural Network Exchange) format and is then converted to RKNN (Rockchip Neural Network) format using RKNN toolkit of Rockchip. Fig. 4 shows the whole training and converting process. The conversion process is not only the format but also the data type. In this work, the data type INT 8 is used for better computational performance.

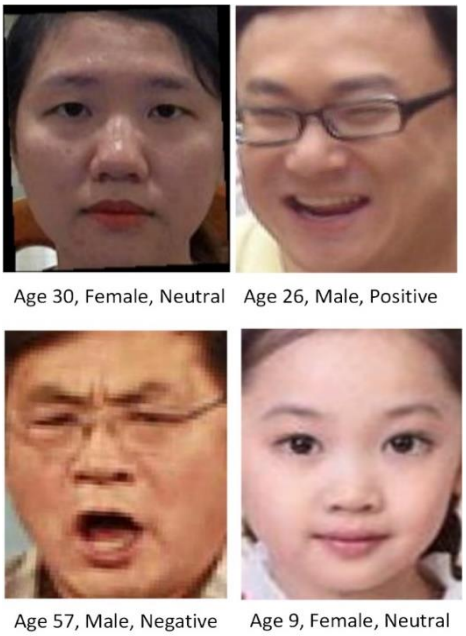

Fig. 3. Illustration of images in the dataset. Each image is labeled by its corresponding age, gender and emotion

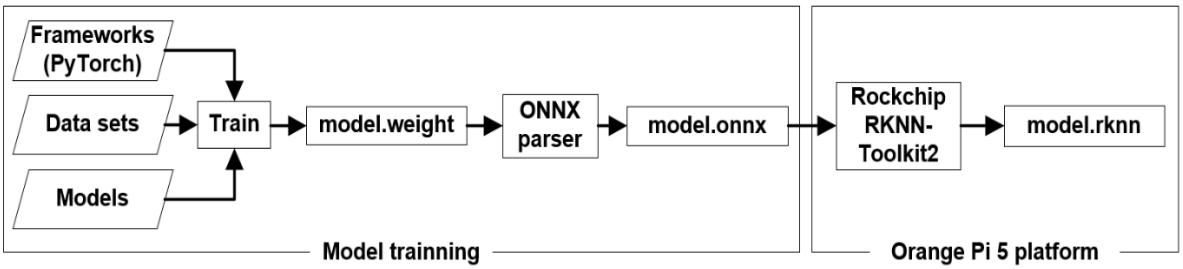

Fig. 4. Development and deployment pipeline of deep learning models on Orange PI 5

3. Results and Discussion

To evaluate the quality of the model, we implemented several testcases based on the testing and validation dataset as listed in Table 2. We introduce testcases as follows:

- Test case 1: we compare the accuracy of the trained models including the MoblieNetV2 and the customized MobileNetV2 with the Resnet50. The accuracy of prediction of age, gender, and emotion are compared. The datatype is set to floating point 32 bits. The computation was carried out on the Nvidia GPU Tesla V100-PCIE. The evaluations were based on 65744 images including 45970 images of the validation of IMDB and 10774 images of our self-collected dataset.

- Test case 2: we evaluate the accuracy of the customized MobileNetV2 implemented on Orange Pi 5 platform. The model was quantized to the data type INT 8 and converted to the corresponding format of the hardware using RKNN toolkit. We change the evaluating dataset in three cases as listed in Table 4.

- Test case 3: we compare the accuracy of models with the aspect of single task, dual task, and three task learning. The data type is set to floating point 32 bit.

- Test case 4: we evaluate the computational efficiency of the developed models on the orange PI hardware platform. The utilization of the hardware and computational performance on each model are compared.

$$Acc = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \quad (2)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \quad (3)$$

For gender and emotion prediction we use the accuracy metrics, which are calculated according to (2), where TP, TN, FP, FN are denoted for true positive, true negative, false positive, and false negative, respectively. For the age prediction the metric MAE (Mean Absolute Error), which is defined by (3), is used, where x is the numerical value from the actual values, and y is the corresponding numerical value from the predicted values for a total of number of predictions.

The results of test case 1 are shown in Table 3. It is seen that, compared to the Resnet50, the MobileNetV2 and customized MobileNetV2 have a slightly higher accuracy. The customized MobileNetV2 has two advantages. On the one hand, it outperforms others. On the other hand, its backbone is suitable for running on the NPU of Rockchip.

Table 3. Comparison of model's accuracy on test case 1

Backbone	Data type	Age (MAE)	Gender (Acc., %)	Emotion (Acc., %)
Resnet50	FP32	4.111	98.251	93.277
MobileNetV2	FP32	3.338	98.214	94.531
MobileNetV2 customized	FP32	3.216	98.351	94.959

Table 4 shows the results of the customized MobileNetV2 in which the model was converted to RKNN format and has data type of INT8. All implementation was carried out on NPU of Rockchip of orange PI platform. The accuracy of the model on three cases of evaluating dataset are considered. The evaluation on dataset of case 1 has highest accuracy while the one with case 3 has lowest accuracy. It is inferred that the testing dataset of IMDB is most challenging. The validation dataset of IMDB is less challenging. Our self-collected dataset is a little bit more challenging than the validation of IMDB which results in a slight reduction of accuracy in case 2. In general, all cases have quite good accuracy.

In test case 3, comparison of the model's accuracy with single task and multi-task learning is shown in Table 5. Model MobileNetV2 and customized MobileNetV2 are compared. It is seen that with the same model, single task learning, and multi-task learning has a very small discrepancy in accuracy. Thus, the use of multi-task learning has an advantage of computational efficiency while remaining accuracy as high as in the single task learning.

Table 4. Comparison of model's accuracy on test case 2

Case	**Dataset** (No. of images)	**Data type**	**Age (MAE)**	**Gender (Acc., %)**	**Emotion (Acc., %)**
1	45970 of the validation of IMDB and 10774 of the validation of ours	INT8	3.485	98.281	93.917
2	10774 of the validation of ours	INT8	3.535	96.937	92.414
3	56086 of the testing of IMDB	INT8	5.099	98.789	89.794

Table 5. Comparison of model's accuracy on test case 3

Checkpoint	**Backbone**	**Params**	**Image size**	**Age (MAE)**	**Gender (Acc., %)**	**Emotion (Acc., %)**
Age	MobileNetV2	4.3M	112	5.026	-	-
Gender	MobileNetV2	4.3M	112	-	98,61	-
Emotion	MobileNetV2	4.3M	112	-	-	**90.83**
Age and gender	MobileNetV2	4.3M	112	5.021	98,74	-
Age and emotion	MobileNetV2	4.3M	112	5.024	-	89.59
Gender and emotion	MobileNetV2	4.3M	112	-	98.76	89.61
Age, gender, emotion	MobileNetV2	4.3M	112	5.022	**98.88**	89.88
Age, gender, emotion	MobileNetV2 customized	4.3M	112	**5.013**	98.86	89.72

Table 6. Hardware utilization and computational performance of developed models

Type of Model	**CPU** (%)	**NPU** (kB)	**Performance** (FPS)
Multi-task MobileNetV2	2.69	7478	256.4
Multi-task customized MobileNetV2	1.92	**7735**	**285.7**
3 single-task models simultaneous execution	3.92	20455	232.6
3 single-task models in sequency execution	1.11	6934	106.5

Comparison of hardware utilization and computational performance of developed models is shown in Table 6. The evaluation was performed on orange PI platform equipped with Rockchip RK3588 8-core 64bit processor, quad-core A76+quad-core A55, 16GB Ram LPDDR4, embedded NPU supports INT4/INT8/INT16/FP16 mixed computing, with up to 6Tops of computing power. The utilization of CPU, NPU and processing performance FPS (frame per second) are calculated. It is seen that with the multi-task models the utilization of CPU and NPU is less than 3 single-task models which are executed simultaneously. Especially, the use of NPU of customized MobileNetV2 is three times less than that of single-task models. Also, the computational is the highest, which reaches 285.7 FPS, while the single-task model has only 232.2 FPS. Even in the worst case, if the single-task model is executed in sequency, it has very low performance with an FPS of

106.5. These results confirm the efficiency of the multi-task approach.

4. Conclusions and Outlook

In this work, a multi-task learn model for age, gender, and emotion predictions has been successfully developed. The use of multi-task model has an advantage of computational efficiency while remaining the accuracy as high as in single-task model. This allows us to deploy model on edge devices which have very limited hardware resources.

In the future, the system will be extended in some directions. The use of the Orange PI plus has demonstrated that it has a high potential for use as an edge processing device. Thus, a thorough evaluation of the device in the concept of hybrid edge-central computing should be investigated.

Acknowledgement

This research is funded by the CMC Applied Technology Institute, CMC Corporation, Hanoi, Vietnam.

References

[1] Wang, M., Deng, W., Deep face recognition: a survey, Neurocomputing, vol. 429, pp. 215-244, Mar. 2021. https://doi.org/10.1016/j.neucom.2020.10.081

[2] Face Analysis, Visage Technology, Diskettgatan 11A, 583 30 Linköping, Sweden; Accessed on: 03/08/2024, Online available at: https://visagetechnologies.com/face-analysis/

[3] Foggia, P., Greco, A., Saggese, A., and Vento, M., Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition, Engineering Applications of Artificial Intelligence, vol. 118, art no. 105651, Feb. 2023. https://doi.org/10.1016/j.engappai.2022.105651

[4] Sang, D. V., Cuong, L. T. B., and Thieu, V. V., Multi-task learning for smile detection, emotion recognition and gender classification in Proceedings of the 8th International Symposium on Information and Communication Technology, Dec. 2017, pp. 340-347. https://doi.org/10.1145/3155133.3155207

[5] Vu, D. Q., Phung, T. T. T., Wang, C. Y., and Wang, C., Age and gender recognition using multi-task CNN, in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18-21 Nov. 2019, pp. 1937-1941. https://doi.org/10.1109/APSIPAASC47483.2019.9023045

[6] Yoo, B., Kwak, Y., Kim, Y., Choi, C., and Kim, J., Deep facial age estimation using conditional multitask learning with weak label expansionin IEEE Signal Processing Letters, vol. 25, iss. 6, Apr. 2018, pp. 808-812. https://doi.org/10.1109/LSP.2018.2822241

[7] Xu, L., Fan, H., and Xiang, J., Hierarchical multi-task network for race, gender and facial attractiveness recognition in 2019 IEEE International conference on image processing (ICIP), Taipei, Taiwan, Aug. 2019, pp. 3861-3865. https://doi.org/10.1109/ICIP.2019.8803614

[8] Ming, Z., Xia, J., Luqman, M. M., Burie, J. C., and Zhao, K., Dynamic multi-task learning for face recognition with facial expression, arXiv preprint arXiv:1911.03281, 2019. https://arxiv.org/abs/1911.03281

[9] Han, H., Jain, A. K., Wang, F., Shan, S., and Chen, X., Heterogeneous face attribute estimation: a deep multi-task learning approach, IEEE transactions on pattern analysis and machine intelligence, vol. 40, iss. 11, Aug. 2017, pp. 2597-2609. https://doi.org/10.1109/TPAMI.2017.2738004

[10] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L. C., Mobilenetv2: inverted residuals and linear bottlenecks in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18-23 Jun. 2018, Salt Lake City, UT, USA, pp. 4510-4520. https://doi.org/10.1109/CVPR.2018.00474

[11] Lin, Y., Shen, J., Wang, Y., and Pantic, M., Fp-age: leveraging face parsing attention for facial age estimation in the wild, IEEE Transactions on Image Processing, vol. 31, pp. 5979-2992, 2022.

[12] IMDB-Clean: A Novel Benchmark for Age Estimation in the Wild, Accessed on 03/08/2024, Online available at: https://github.com/yiminglin-ai/imdb-clean

[13] Rockchip product introduction, Fuzhou, Fujian, PRC, Accessed on: 03/08/2024, Online available at: https://www.rock-chips.com/a/en/

[14] Orange Pi production introduction, Shenzhen, Guangdong, China, Accessed on: 03/08/2024, Online available at: http://www.orangepi.org/html/hardWare/computerAndMicrocontrollers/service-and-support/Orange-Pi-5-plus.html

[15] Introduction of rockchip toolkit, Accessed on: 03/08/2024, Online available at: https://github.com/rockchip-linux/rknn-toolkit

[16] Huber, P. J., Robust estimation of a location parameter in Breakthroughs in Statistics: volume II, Methodology and Distribution, 1st ed., New York, Springer, c.1992, pp. 492-518. https://doi.org/10.1007/978-1-4612-4380-9_35

[17] Tran Đuc Trung, EfficientNet: A new approach to model scaling for convolutional neural networks (In Vietnamese: EfficientNet: Cách tiếp cận mới về model scaling cho convolutional neural networks), Accessed on: 03/08/2024, Online available at: https://viblo.asia/p/efficientnet-cach-tiep-can-moi-ve-model-scaling-cho-convolutional-neural-networks-Qbq5QQzm5D8