Age and Gender Recognition Using Multi-task CNN

Duc-Quang Vu*, Thi-Thu-Trang Phung†, Chien-Yao Wang‡ and Jia-Ching Wang*§
* Dept. CSIE, National Central University, Taoyuan, Taiwan
E-mail: vdquang1991@gmail.com E-mail: jcw@csie.ncu.edu.tw
† Thai Nguyen University, Thai Nguyen, Vietnam
E-mail: phungthutrang.sfl@tnu.edu.vn
‡ Institute of Information Science, Academia Sinica, Taiwan
E-mail: x102432003@yahoo.com.tw
§ Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan

***Abstract*—The investigation into age and gender identification has been receiving more attention from researchers since social and multimedia networks are becoming more popular nowadays. Recently published methods have yielded quite good results in terms of accuracy but have also proven to be ineffective in real-time applications because the models were too complicated. In this paper, we propose a lightweight model that can classify both age and gender. The number of parameters used in this model is 5 times less than existing models. Experiment results show that the accuracy of the proposed method is equivalent to state-of-the-art methods, while the speed of age and gender recognition decreases by 4 times on the Audience benchmark.**

I. Introduction

Image processing and computer vision are two popular fields with many applications in the real world. For example, Alex et al. [1] proposed a model using Convolutional Neural Network (CNN) and won the ImageNet Large Scale Visual Recognition Challenge with 15.3% for the top-5 test set error rate in 2012. In 2013, Zeiler et al. [2] proposed a model with the name "ZFNet" and reduced the top-5 error from 15.3% to 14.8%. GoogleNet (Inception) and VGGNet was proposed in 2014 with top-5 errors of 6.67% and most7.32% respectively. At the ILSVRC in 2015, Kaiming He proposed ResNet, which is a novel architecture with skip connections and features heavy batch normalization. It achieves a top-5 error rate of 3.57% which is better than human performance. In the multi-object detection problem, there were many proposed models in recent years such as Faster R-CNN, R-FCN, SSD, FPN, RetinaNet, and YOLOv3, etc. In addition, the deep learning technique has been successfully applied to many other problems likes face recognition, image captioning, semantic image segmentation.

The face is a part of the human body and a detailed face annotation might include a lot of different information such as age, gender, emotional state, and ethnicity. Age and gender are two of the key facial attributes, they are useful for many applications in social interactions. Age and gender estimation systems from a single face image are used a lot in intelligent tasks such as marketing intelligence, recommendation, and computer interaction. This is a challenging issue and has great practical significance even for humans, very difficult to predict exactly the age of a person because younger people may look older than some older people and vice versa. Over the last decade, there were many proposed methods used image transformations, statistical techniques, feature extraction techniques and/or machine learning to estimate age and gender. More recently, popular deep learning techniques can achieve promising accuracy in age estimation. For example, Levi and Hassner [3] proposed a network model in 2015 using CNN with 50.7% and 86.8% accuracy for age and gender classification, respectively. Lapuschkin et al. [4] applied GoogleNet, CaffeNet and VGG-16 models for age and gender classification problem on Adience benchmark. The best result in their paper is VGG-16 with 138M parameters, VGG-16 is achieved 62.8% accuracy for age classification and 92.6% accuracy for gender classification. Duan et al. [5] proposed a model with combining between CNN and Extreme Learning Machine (ELM) and CNN-ELM model is achieved 52.3% and 88.2% accuracy for age and gender classification on Adience benchmark. However, most of the proposed methods are single-task learning that means each task is processed separately. Recently, new approaches involving the use of multi-task learning enable learning shared representations and keep correlations between tasks. In addition, to improve accuracy, models are usually built more complex with deeper layers, more parameters, but in many real world applications such as self-driving car, the recognition tasks need to process in the real-time. Therefore, to build a multi-task model with efficient network architecture is a challenging problem.

In this paper, we present a new multi-task model using CNN with the number of parameters in the range of 0.9M. The age classification accuracy is 66.82%, and gender classification accuracy is 87.07% for proposed model. With a lower number of parameters, our lightweight model can easily perform on mobile or embedded devices. Our key contributions are: (1) Build a lightweight model to solve the multi-task problem (age and gender classification). (2) The result of this paper shows that the method performs well in terms of accuracy and also reduce the number of parameters, which helps to improve the running time of the model.

The remainder of the paper is organized as follows. Section 2 provides a review of the related work. The lightweight multi-task model is proposed in Section 3 which also includes the network architecture, loss function, and the training/testing procedure. The experimental results, comparisons and com-

ponent analysis are presented in Section 4 and the final the conclusions are given in Section 5.

II. Related Work

A. Depthwise Separable Convolutional Neural Networks

Instead of using standard convolution, Depthwise Separable Convolutional Neural Networks uses 2 operations include: Depthwise convolution and Pointwise Convolution to reduce the number of parameters used and computational cost. Depthwise Separable Convolutional Neural Networks [6] is the special technique and usually used in Mobilenet [6] and Mobilenet v2. [7].

From input layer with $D_F \times D_F \times M$ where $D_F \times D_F$ can be the image size and M is the number of channels. With standard convolution, we use N filters with kernel size of each filter is $D_K \times D_K \times M$, the output size will be $D_P \times D_P \times N$. Standard convolutions have the computational cost of:

$$D_K \times D_K \times M \times N \times D_F \times D_F \tag{1}$$

With Depthwise Separable Convolution is broken down into 2 operations. In depth-wise operation, convolution is applied to a single channel at a filter per each input channel. Depthwise convolution has a computational cost of:

$$D_K \times D_K \times M \times D_F \times D_F \tag{2}$$

In point-wise operation, a 1×1 convolution operation is applied on the M channels. point-wise operation has a computational cost of:

$$M \times N \times D_F \times D_F \tag{3}$$

Comparison between the complexities of the Depthwise Separable Convolution and standard Convolution:

$$\frac{D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F}{D_K \times D_K \times M \times N \times D_F \times D_F} = \frac{1}{N} + \frac{1}{D_K \times D_K} \tag{4}$$

If we use 3×3 kernel size for Depthwise Separable Convolution, then the computational cost will be reduced 8 to 9 times less than standard convolutions.

B. Age classification

The age classification is one of the most popular problems in computer vision. Classifying age from facial images was first introduced by Kwon et al. Then Ramanathan and Chellappa proposed an improvement from Kwon's method. These traditional methods for age estimation are based on calculating ratios between different measurements of facial features. However, the methods are not suitable to distinguish the adults due to they are very sensitive to head pose. Another approach, Cootes and Edwards [8] proposed an active appearance model (AAM) to represent the face image. The AAM model can classify different age groups by using shape and texture information together. This is an advantage of AAM compare with previous methods. Then Geng et al. [9] proposed an AGing pattErn Subspace (AGES) model which was used to estimate age automatically. Although these methods have shown many advantages. However, this method needs a large number of cross-age images of one person, so that these approaches are not suited for many practical applications. After 2007, there were many proposed methods such as Gabor [10], Local Binary Patterns (LBP) [11], Spatially Flexible Patch (SFP) [12], and BIF [13]. Based on these features extractive techniques, machine learning methods are used to predict the age from an input image. [13] and [14] used SVM to age classification and SVR [15], PLS [16], and CCA [17] were used to age prediction. In recent year, many Deep Neural Network (DNN) models were proposed. For example, Levi and Hassner [3] proposed the first DNN model for the age and gender classification problem in 2015. In 2016, Zhu et al. [18] proposed a lightweight model to improve the age classification performance. Lapuschkin et al. [4] applied GoogleNet, CaffeNet and VGG-16 models for age and gender classification problem on Adience benchmark. Last year, Duan et al. [5] proposed a model with combining between CNN and ELM.

C. Gender classification

Golomb et al. [19] were some of the early researchers who proposed the model to classify gender based on neural network. Lyons et al. [20] built the model based on principal component analysis (PCA) and linear discriminant analysis (LDA) to predict gender. SVM was used by Moghaddam and Yang [21] for gender classification. Baluja and Rowley [22] based on AdaBoost to predict the gender from facial images. Although all of the methods mentioned above have a lot of advantages and achieved high performance. However, they were evaluated on constrained images dataset such as FERET, MORPH, FG-NET, the results on the all benchmarks show that they are not challenging for new methods. Eidinger et al. [14] not only designed a classification pipeline to classify gender but also presented new and extensive benchmarks named Adience to study age and gender classification. The Adience benchmark has often been used to compare results among recently proposed models because images in the Adience benchmark are more challenging than the old benchmarks mentioned above. The same with age classification, the models from [3], [18], [4], [5] also are built and evaluate the performance on gender classification.

Although all DNN models above have achieved a good result on performance. But the architecture of them are very complex with a lot of hidden layers and they are single-task models that mean we have two models with the same architecture but different about weights for age and gender classification, respectively. In order to this problem, we propose a multi-task deep model with a number of parameters and number of hidden layers very low, which conform with mobile and embedded devices.

III. Proposed Model

A. Multi-task learning

Multi-task learning has been applied successfully across all applications of machine learning, such as computer vision, natural language processing, and speech recognition. Multi-task learning has many techniques include: joint learning, learning to learn, and learning with auxiliary tasks. In the Deep Learning field, multi-task learning has two types is either hard or soft parameter sharing of hidden layers. In this paper, we build the model based on hard parameter sharing of hidden layers. This is the most commonly used approach to Multi-task learning in DNNs [23]. In hard parameter sharing of hidden layers, most of the hidden layers will be shared between all tasks, only the output layers are different. [24] proved hard parameter sharing greatly reduces the risk of overfitting.

B. Model Architecture

Our proposed model is shown in Figure 1. The model is divided into three main stages: Convolution, Depthwise Separable Convolution, and Fully Connected stage. In the first stage, we use standard CNN because in this state, the size of the filter matrix on each layer is not too big and we need to obtain more new features than the following state. All the operations in this stage include Convolution (Conv) + Batch Normalization (BN) + Rectified Linear Unit (ReLU) + Max Pooling (MaxPool) + Drop out (Dropout). In the second stage, Depthwise Separable Convolution is used to replace for standard Convolution to reduce size and complexity. All the operations in this stage include Depthwise convolution (DepthwConv) + Batch Normalization (BN) + ReLU6 + Pointwise Convolution (1x1 Conv) + Batch Normalization (BN) + ReLU6. Finally, Fully Connected Box is used with Global Average Pooling (GA Pooling) and Fully Connected (FC) operations. More detail about model is present in Table I (we remove some operations with none hyperparameter or only one hyperparameter such as ReLU, GA Pooling, FC (number of units = 512) and dropout (drop rate = 0.25).

TABLE I
Proposed model architecture

Type	Hyperparameters	Output Size
Input Image	None	64×64×3
Conv	No. of Filer = 32, filter size = 3×3, stride=1, padding = same	64×64×32
MaxPool	Filter size = 2×2, stride=2	32×32×32
Conv	No. of Filer = 64, filter size = 3×3, stride=1, padding = same	32×32×64
MaxPool	Filter size = 2×2, stride=2	16×16×64
Conv	No. of Filer = 128, filter size = 3×3, stride=1, padding = same	16×16×128
MaxPool	Filter size = 2×2, stride=2	8×8×128
DepthwConv	Filter size = 3×3, stride=1, padding = same	8×8×128
1×1 Conv	No. of Filer = 256, filter size = 1×1, stride=1, padding = same	8×8×256
DeptwConv	Filter size = 3×3, stride=2, padding = same	4×4×256
1×1 Conv	No. of Filer = 256, filter size = 1×1, stride=1, padding = same	4×4×256
DepthwConv	Filter size = 3×3, stride=1, padding = same	4×4×256
1×1 Conv	No. of Filer = 512, filter size = 1×1, stride=1, padding = same	4×4×512
DepthwConv	Filter size = 3×3, stride=2, padding = same	2×2×512
1×1 Conv	No. of Filer = 256, filter size = 1×1, stride=1, padding = same	2×2×512

C. Training and Testing

Initialization. The input RGB images are downsampled to the resolution of 64 × 64. The weights of all layers are initialized with Xavier uniform initializer and the bias of all layers are initialized to zeros. The target is a vector *y* with two values, the first value is the estimate for age and the second value is the estimate for gender.

Loss. Suppose we have N training samples, and T tasks (T = 2 in this problem). We assign ŷ as the output vector from the model. The loss function is calculated in Equation 5.

$$\mathcal{L}(w,b) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}||y_i^t - \hat{y}_i^t)||_2^2 \quad (5)$$

where w and b represents the corresponding network parameters. We use L_2 norm regularization to penalizes the complexity of $w^{[l]}$ to avoid model overfitting with $w^{[l]}$ as the parameters of layer l. L_2 norm is formulated as:

$$L_2 = \lambda\sum_{l=1}^{L}||w^{[l]}||^2 \quad (6)$$

where λ is the regularization parameter. In our work, λ is set to 0.001. From Equation 5 and Equation 6 we construct the cost function in Equation 7

$$\mathcal{J}(w^{[l]},b^{[l]}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}||y_i^t - \hat{y}_i^t)||_2^2 + \lambda\sum_{l=1}^{L}||w^{[l]}||^2 \quad (7)$$

In the testing phase, our model can directly estimate the result of both age and gender from an input image by forwarding the network.

IV. Experiment and Result

A. Dataset and Setting

As mentioned in Section 2.2, we use the Adience benchmark from [14] to test and evaluate the model. Adience dataset is mainly constructed for age and gender recognition. The dataset contains 26K images of more than 2K different people and

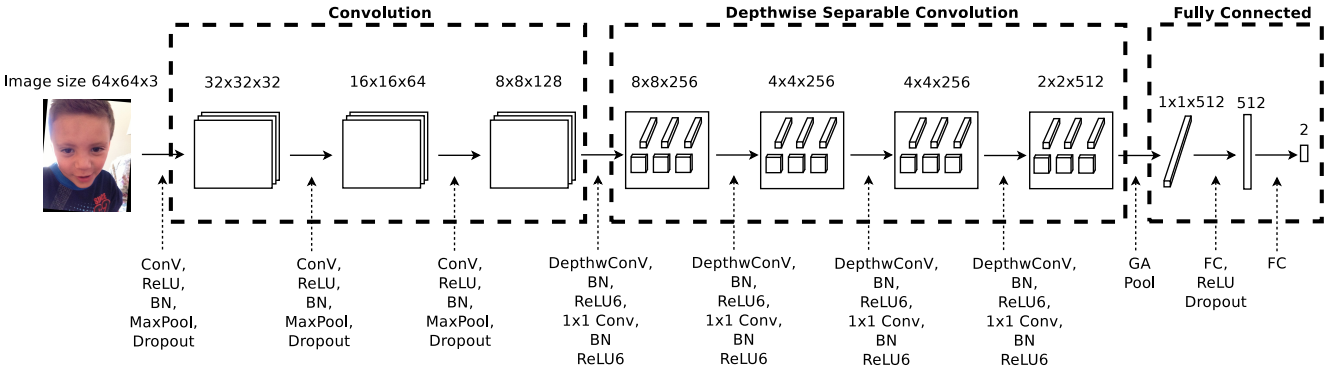

Fig. 1. Proposed Model

each image has resolution of 816 × 816. Most images from the dataset are automatically uploaded to Flickr and they are directly collected from mobile devices without prior manual filtering. The dataset categorizes the subjects into 8 groups based on the age in years 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60 and above. Figure 2 is an example of images with poor lighting conditions, facial obstructions, extreme variations in head poses etc. Our model is trained from scratch

Fig. 2. An example about Adience dataset

with Adam optimizer function. Training images are split into multiple parts with a batch size of 32 images and learning rate of 0.001. We use five-fold cross-validation to obtain the recognition accuracies of both age and gender. The results are compared with respect to accuracy and efficiency with four state-of-the-art methods in [3], [4], [18], [5].

B. Results and Comparison

As can be shown in Table II, the result of our model is compared with AdienceNet [3], Lightweight [18], CaffeNet [4], GoogleNet [4], VGG-16 [4] and CNNELM [5]. Experiment results show that the accuracy of our model is 4% higher than the best state-of-the-art method in age classification. In gender classification, the accuracy of our model is 5% lower than deeper models such as VGG-16. The results indicate that, our model extracts features corresponding to age rather than gender. Table III shows the number of parameters used in each model. Our model used just 0.9M parameters, which is 4.4 times lower with GoogleNet, 6.7 times lower than Best [18] and 10 times lower than AdienceNet [3]. We compared with AdienceNet [3] and Best [18] since these are also the lightweight models with shallow-layer and a low number of parameters. We rebuild the models and executed them on the machine with 3.6GHz CPU and no GPU. The AdienceNet and Best models required 0.64 and 0.48 seconds, respectively, to estimate the age and gender from an image. On the other hand, our model only required 0.08 seconds to perform the same task. On a mobile phone, our model runs 4 and 6 times faster when compare with AdienceNet and Best, respectively.

TABLE II
AGE AND GENDER ACCURACY ON ADIENCE BENCHMARK

Method	Age	Gender
AdienceNet [3]	50.7% ± 5.1%	86.8% ± 1.4%
Lightweight [18]	46.0% ± 0.6%	86.0% ± 1.2%
CaffeNet [4]	54.3%	90.6%
GoogleNet [4]	58.5%	91.7%
VGG-16 [4]	62.8%	**92.6%**
CNN-ELM [5]	52.3% ± 5.7%	88.2% ± 1.7%
Our Model	**66.8% ± 1.0%**	87.1% ± 0.7%

V. CONCLUSION

In this paper, we have proposed a novel approach to combine Convolutional Neural Network and Depthwise Separable Convolution towards the construction of a lightweight multi-task model. The new model requires fewer parameters, yet it provides better performance than existing models. The proposed model facilitates age and gender identification in real time and on mobile devices.

In the near future, we plan to improve the accuracy of the model, especially in gender identification. In addition, we will

TABLE III
COMPARE NUMBER OF PARAMETERS WITH STATE-OF-THE-ART

Method	No. of parameters	Time to run on desktop	Time to run on mobile
AdienceNet [3]	≈9M	0.64	0.12
Best from [18]	≈6.1M	0.48	0.08
CaffeNet [4]	≈61M	-	-
GoogleNet [4]	≈4M	-	-
VGG-16 [4]	≈138M	-	-
CNNELM [5]	≈11M	-	-
Our Model	**≈0.9M**	**0.08**	**0.02**

also apply our model to other problems related to computer vision and image processing.

ACKNOWLEDGMENT

This research is partially supported by the Ministry of Science and Technology under Grant Number 108-2634-F-008 -004 through Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan.

REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[3] Gil Levi and Tal Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 34–42.

[4] Sebastian Lapuschkin, Alexander Binder, Klaus-Robert Müller, and Wojciech Samek, "Understanding and comparing deep neural networks for age and gender classification," in *of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1629–1638.

[5] Mingxing Duan, Kenli Li, Canqun Yang, and Keqin Li, "A hybrid deep learning cnn–elm for age and gender classification," *Neurocomputing*, vol. 275, pp. 448–461, 2018.

[6] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[7] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[8] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 6, pp. 681–685, 2001.

[9] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 29, no. 12, pp. 2234–2240, 2007.

[10] Feng Gao and Haizhou Ai, "Face age classification on consumer images with gabor feature and fuzzy lda method," in *International Conference on Biometrics*. Springer, 2009, pp. 132–141.

[11] Asuman Gunay and Vasif V Nabiyev, "Automatic age classification with lbp," in *2008 23rd International Symposium on Computer and Information Sciences*. IEEE, 2008, pp. 1–4.

[12] Shuicheng Yan, Ming Liu, and Thomas S Huang, "Extracting age information from local spatially flexible patches," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 737–740.

[13] Guodong Guo, Guowang Mu, Yun Fu, and Thomas S Huang, "Human age estimation using bio-inspired features," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 112–119.

[14] Eran Eidinger, Roee Enbar, and Tal Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[15] Guodong Guo, Yun Fu, Charles R Dyer, and Thomas S Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1178–1188, 2008.

[16] Guodong Guo and Guowang Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *CVPR 2011*. IEEE, 2011, pp. 657–664.

[17] Guodong Guo and Guowang Mu, "Joint estimation of age, gender and ethnicity: Cca vs. pls," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–6.

[18] Linnan Zhu, Keze Wang, Liang Lin, and Lei Zhang, "Learning a lightweight deep convolutional network for joint age and gender recognition," in *Pattern Recognition (ICPR), 23rd International Conference on*. IEEE, 2016, pp. 3282–3287.

[19] Beatrice A Golomb, David T Lawrence, and Terrence J Sejnowski, "Sexnet: A neural network identifies sex from human faces.," in *NIPS*, 1990, vol. 1, p. 2.

[20] Michael J Lyons, Julien Budynek, Andre Plante, and Shigeru Akamatsu, "Classifying facial attributes using a 2-d gabor wavelet representation and discriminant analysis," in *Automatic Face and Gesture Recognition, Proceedings. Fourth IEEE International Conference on*. IEEE, 2000, pp. 202–207.

[21] Baback Moghaddam and Ming-Hsuan Yang, "Learning gender with support faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 707–711, 2002.

[22] Shumeet Baluja and Henry A Rowley, "Boosting sex identification performance," *International Journal of computer vision*, vol. 71, no. 1, pp. 111–119, 2007.

[23] Rich Caruana, "Multitask learning: A knowledge-based source of inductive bias," in *ICML*, 1993.

[24] Jonathan Baxter, "A bayesian/information theoretic model of learning to learn via multiple task sampling," *Machine learning*, vol. 28, no. 1, pp. 7–39, 1997.