임베디드 시스템을 위한 멀티태스킹 딥러닝 학습 기반 경량화 성별/연령별 추정

Huy-Tran Quoc Bao*, 정선태**
*Dept. of Information and Telecommunication, Graduate School, Soongsil University
**Dept. of Smart System Software, Soongsil University
*Huy.tsusak@gmail.com, **cst@ssu.ac.kr

A light-weight Gender/Age Estimation model based on Multi-taking Deep Learning for an Embedded System

Huy-Tran Quoc Bao*, Chung Sun-Tae**
*Dept. of Information and Telecommunication, Soongsil University
**Dept. of Smart System Software, Soongsil University

ABSTRACT

Age estimation and gender classification for human is a classic problem in computer vision. Almost research focus just only one task and the models are too heavy to run on low-cost system. In our research, we aim to apply multi-tasking learning to perform both task on a lightweight model which can achieve good precision on embedded system in the real time.

1. Introduction

The studying of age estimation and gender classification issue has many useful applications in products recommendation or human-computer interaction. Estimating physical ages for facial images are such a challenge not only for computer but also for humans because more often than not, physical ages can be very different from apparent ages. Gender classification seems to be a lighter task compared to age estimation because the classes to be classified are just two (male or female) while estimation of the age can range from 0 up to 100.

Multi-task learning [1] is a technology to learn several tasks at the same time. In deep neural network, Multi-task learning model can do several predictions simultaneously from the input data. Multi-task learning was inspired from human's learning process. For example, a person who studied playing piano can also use that knowledge about music notes and chords to for studying how to play the guitar. Thus, when a multi-task neural networks learn to perform multi tasks, tasks should have some relations among them. Since the feature extraction layer of multi-task networks can learn much more information, they could perform better than single task ones,

In a simultaneous age estimation and gender classification, features extracted from facial image can support each other to perform the task better. In research of B.Yoo[2], they create a CNN which can apply gender information to utilize age prediction. They also show that the gender information significantly improves robust age estimation accuracy.

Inspired by that result and Multi-task learning. we create a light-weight model which do both tasks of age estimation and gender classification. Through experiments, our model can run in real-time on an embedded system such as Nvidia Jetson Nano board[3] with high accuracy compared to. previous models for gender/age estimation

2. Related work

Many approaches are proposed to solve age estimation problem. The age estimation task is more complicated than gender classification because the number of age value is much more than gender.

There are proposed many models showing great results on age estimation. DEX[4], MV[5], ARN[6] do the task by using VGG-16[7] architecture and applying multi-class classification. DEX also ensembles the outputs of many network to produce a robust result. AgeNet[8] approaches aging labeling as a regression problem using deep neural net based on GoogleNet[9]. For making it run on embedded systems or low-cost devices, many lightweight neural network architectures have been designed so as to reduce the model size such as

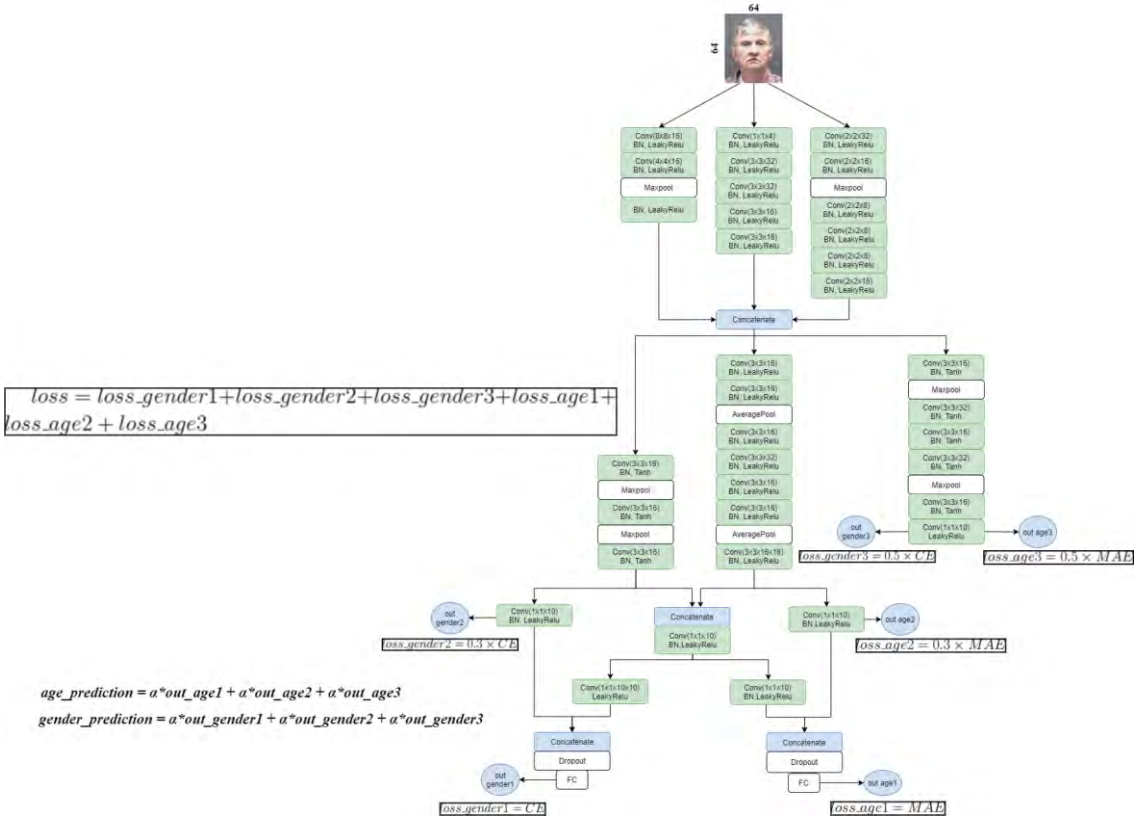

Figure 1: Proposed Network architecture

MobileNet[10], DenseNet[11]. SSR-Net[12] proposed by Tsun-Yi Yang using Soft-Stagewise regression network is a very compact model but also achieves competitive results.

[2] proposed a model which predicts both age and gender label using conditional Multi-task learning. The model performs age estimation job with combined gender information so as to improve accuracy. The research shows that gender feature really useful for age estimation task.

We propose a light-weight CNN neural network with multi-task learning to predict age and gender simultaneously from facial image.

3. Proposed CNN network with multi-tasking

The overall proposed network architecture is showed in Figure 1. At the beginning of the network, we use 3 branches of convolution layers followed by Batch Normalization and Leaky-Relu activation function to extract some general features. Each branch describes information of images differently to get useful information for both age and gender estimation.

After getting some general features, we instruct our network to extract gender and age feature separately. To do that, we construct the network with two branches added; the first branch aims to detect gender features and the second branch aims to detect age features. Since age estimation is harder task, the second branch contain more convolution layers. In order to force the second branch to learns about ages features, we add a virtual output with a loss function at the end of this branch. The loss function here is aimed to instruct this branch to learn about age features. We only make the age prediction at this output. Similarly, for the gender task is easier, we also add a virtual output at the end of this branch to force this branch to learn useful features for gender classification. At the top of our network, we combine both age and gender features with a few simpler convolution layers, dropout layers and fully connected layers to output proper age and gender. Beside the age and gender branch, we also design another branch to predict both age and gender. This branch uses low level features to make prediction which use only 6 convolution layers followed by fully connected layers. These low-level outputs are used to ensemble with virtual outputs and the final outputs to make prediction. The input images are used with small size (64x64) to reduce the number of computation.

Ensemble is a well-known and efficient technique to improve accuracy by aggregating prediction of many machine learning models. Inspired by this idea, we design our network to have 3 pairs outputs of gender and age. After getting 3 outputs we ensemble them to get the final prediction. By doing this way, we can utilize the advantage of ensemble technique without sacrificing processing speed too much and make our model run successfully on our embedded board – Jetson Nano[3].

4. Training process

4.1 Dataset

MORPH-II[13] is a large dataset for real age estimation, which contains 55,134 color images of 13,617 subjects with age and gender information. The ages in MORPH-II ranges from 16 to 77 years old.

FGNET[14] data set contain 1002 image of 82 people; each image is labeled with physical age. The ages range from 0 to 69. This is a small dataset; the method usually adopted to evaluate is 'leave one-person-out (LOPO)'[15].

We aim to detect age and gender for Asian people, however the face dataset which consist of both age and gender information for Asian people is limited. To the best of our knowledge, MegaAge-Asian dataset [16] is the largest face dataset for Asian people which includes about 44k face images. However, this dataset only provides age information. Therefore, we need to annotate gender information for this dataset. To do that, we train 4 well-known image recognition models on another Asian dataset - AAF[17] dataset and annotate mega-Asian dataset semi-automatically using the trained models. Recognition. If the 4 networks produce the same results, we keep that result as ground-truth, if not we manually annotate them. The four networks we use for this purpose are Inceptionv3[18], Xception[19], Resnet50[20] and Densenet201[21].

4.2 Loss function

We define 6 losses for 6 outputs. For gender recognition, we use binary cross entropy loss function which is defined as in (1). For age estimation, we use mean absolute error which is defined as formula (2). We use a small weight loss for 2 virtual outputs, slightly larger weights for low-level output since our final output is more important.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)) \quad (1)$$

where y is the label (1 for man and 0 for woman) and $p(y_i)$ is the predicted probability of i class for all N images.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \quad (2)$$

Where y_i is the ground truth age and $\hat{y}_i$ the predicted age, N is the total images used for calculating loss.

The final loss is the weighted loss of 4 losses as in formula (3).

$$\begin{aligned} total\ \ loss = {} & lossGender1 + lossAge1 \\ & + (lossGender2 + lossAge2) \times 0.3 \\ & + (lossGender3 + lossAge3) \times 0.5 \end{aligned} \quad (3)$$

5. Experiment

We evaluate our model on Morph-II, MegaAge-Asian, and FGNET dataset and compare the evaluations with results of some state-of-art age-gender recognition methods.

Table1. result on Morph II dataset Age+Gender task

Network	**Params**	**MAE**	**Acc (Gender)**	**Protocol**
CMT	5M	2.91	99.2	5-fold
Compact CNN	56.9M	3.23	98.8	2-fold
Ours	99.2k	2.88	99.34	5-fold

Protocol: k-fold means splitting the dataset equally into k parts, and chosen (k-1) parts are used for training and the remaining 1 part is used for evaluating. Then, take average of evaluation of all choices.

From Table 1, CMT model has over 5-million parameters which is 50 times more than ours. Furthermore, they use higher resolution images (128x128) for their application. Even though our model has less parameters and uses smaller input images, it produces better result (2.88/99.34) than CMT (2.91/99.2) respectively for age estimation and gender classification. Ours method also outperforms the performance of Compact CNN.

Table 2. Results on FG-NET dataset

Network	Params	MAE	Protocol
CMT	5M	3.43	LOPO
Ours	99.2K	3.41	LOPO

Our model also performs better than CMT among FGNET dataset with the same evaluation method – leave-one-person-out (LOPO).

In addition to comparison with Multi-task model, we also compare with some models which focus in just only single task – age estimation.

Table3: compare with single task network on Morph-II:

Network	**Parameter**	**MAE**
Ranking CNN	500M	2.96
DEX	138M	3.25
ARN	138M	3.00
MV	**138M**	**2.41**
Dense-Net	242K	5.05
MobileNet-V1	226.3K	6.50
SSR	40.9K	3.16
Ours	**99.2K**	**2.88**

Table 4: compare with single task network on FGNET:

Network	**MAE**
DEX	4.63
MV	4.10
Ours	**3.41**

Compared with the MV network, our network predicts with a lower accuracy in Morph-II dataset. However, the result in the Table 4 shows that we achieve better result in the FGNET dataset. Ours network is also much more lightweight than the MV network.

Cumulative accuracy (CA) measurement is calculated when evaluate on MegaAge-Asian dataset

$$CA(n) = \frac{K_n}{K} \quad (4)$$

K is total number of test images, K_n is number of test images whose absolute error is less than *n*.

Table 5. Results on MegaAge-Asian dataset

Network	Parameters	CA(3)	CA(5)
MobileNet-V1	226.3K	0.440	0.606
DenseNet	242K	0.517	0.694
SSR	40.9K	0.549	0.741
Ours	99.2K	0.585	0.788

Even though our network has a larger volume of parameters, it is still a light-weight model and has more functionality and produce much better result compared to SSR net.

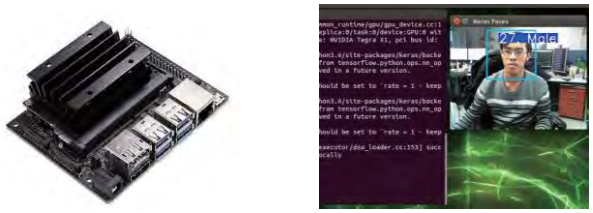

Figure 2: Jetson Nano board and executing age/gender estimation

6. Conclusion

In this paper, we proposed a network model which can perform age estimation and gender classification at the same time and achieve good precision. In addition, this model is light-weight enough to be deployed on embedded system-Jetson Nano. In the future work, we are going to continue to apply Multi-tasking learning for learning more tasks and improve the results.

References

[1] Ruder, Sebastian. "An overview of multi-task learning in deep neural networks.", arXiv preprint arXiv:1706.05098 (2017).

[2] Yoo, ByungIn, et al. "Deep facial age estimation using conditional Multi-task learning with weak label expansion.", IEEE Signal Processing Letters 25.6 (2018): 808-812.

[3] https://elinux.org/Jetson_Nano

[4] Rothe, Rasmus, Radu Timofte, and Luc Van Gool, "Dex: Deep expectation of apparent age from a single image." Proceedings of the IEEE international conference on computer vision workshops. 2015.

[5] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face.", CVPR, 2018.

[6] E. Agustsson, R. Timofte, and L. Van Gool. "Anchored regression networks applied to age estimation and super resolution.", ICCV, 2017

[7] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." CoRR, abs/1409.1556, 2014.

[8] G. Levi and T. Hassner. "Age and gender classification using convolutional neural networks.", CVPRW, 2015.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions.", CVPR, 2015.

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications.", arXiv preprint arXiv:1704.04861, 2017.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. "Densely connected convolutional networks.", CVPR, 2017.

[12] Yang, Tsun-Yi, et al. "SSR-Net: A Compact Soft Stagewise Regression Network for Age Estimation.", IJCAI. Vol. 5. No. 6. 2018.

[13] K. Ricanek and T. Tesafaye. "Morph: A longitudinal image database of normal adult age-progression.", FGR, 2006.

[14] The FG-NET Aging Database, https://yanweifu.github.io/FG_NET_data/, 2014.

[15] J. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi and R. Chellappa, "A cascaded convolutional neural network for age estimation of unconstrained faces," 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), Niagara Falls, NY, 2016, pp. 1-8.

[16] Y. Zhang, L. Liu, C. Li, et al., "Quantifying facial age by posterior of age comparisons", Proceedings of the British Machine Vision Conference, London, UK, 2017.

[17] Cheng, Jingchun, et al. "Exploiting effective facial patches for robust gender recognition.", Tsinghua Science and Technology 24.3 (2019): 333-345.

[18] Xia, Xiaoling, Cui Xu, and Bing Nan. "Inception-v3 for flower classification." 2017 2nd International Conference on Image, Vision and Computing (ICIVC). IEEE, 2017.

[19] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[20] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition.", CVPR, pages 770–778, 2016.

[21] Huang, Gao, et al. "Densely connected convolutional networks.", Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.