Deeply Label Distribution Learning for Age Estimation

September 15, 2015

1 Team details

- Team name: SEU_NJU
- Team leader name: Zhou Ying
- Team leader address, phone number and email:

  Address: 439 Room, Computer Science and Engineer Academy, Southeast University (Jiu Longhu Campus), Jiangning District, Nanjing 211189, China.

  Phone number: +86-181-1596-7520

  Email: zhouying1@seu.edu.cn
- Rest of the team members (In Alphabetical Order):

  Gao Binbin, Geng Xin, Huo Zengwei, Wei Xiushen, Wu Jianxin, Xing Chao and Yang Xu
- Team website URL:

  `http://palm.seu.edu.cn/`

  `http://lamda.nju.edu.cn/`
- Affiliation: Southeast University and Nanjing University

2 Contribution details

- Title of the contribution: Deeply Label Distribution Learning for Age Estimation
- Final score: We can achieve 0.3377 performance on the validation set.

- General method description:

  In short, our method had two streams to get its corresponding age predictions, and then we ensembled these predictions as the final predictions.

  Firstly, we pre-processed the images of both the competition and other facial image data sources (e.g., the MORPH data set, etc.), including face detection [1], key point detection [2] and face alignment.

  Secondly, we used two streams to train different deep models seperately for age estimation. *In the first stream*, we fine-tuned the popular pre-trained CNN model, i.e., VGG16, on the MORPH [3] data set. And then, we used this fine-tuned model to generate multiple different deep net models by fine-tuning again on about 58,000 Internet facial images which were collected by us. For these deep models, we changed the loss function into KL-divergence based on Label Distribution Learning [4], and finally fine-tuned the models on the competition images, and then got the age predictions. *In the other stream*, we designed a new deep convolutional neural network architecture called *Age-Net*, which is shown in Fig. 3. For Age-Net, we downloaded 16,146 facial images from 1-years-old to 85-years-old by search engines including Google, Bing and Baidu. In addition, FG-Net which contains 1,002 facial images is also used in our model. After that, we trained Age-Net on the 22,100 images including 16,146 Internet images, 1,002 FG-Net images and $2,476 \times 2$ ChaLearn Age images (with/without alignment). Similarly to the first stream, we finally fine-tuned Age-Net on the competition images and predicted the age estimations.

  Because each stream had multiple deep models, we did the low-level ensemble in each stream firstly, and finally ensembled the predictions of both two streams to get the final age estimations (i.e., high-level ensemble). The details of our method can be found in Section 5 and Section 6.

- References

[1] M. Mathias, R. Benenson, M. Pedersoli and L. V. Gool. Face detection without bells and whistles. ECCV, 2014.

[2] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. CVPR, 2013.

[3] K. R. Jr and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. FGR, 2006.

[4] X. Geng, Q. Wang and Y. Xia. Facial age estimation by adaptive label distribution learning. ICPR, 2014.

- Representative image / diagram of the method

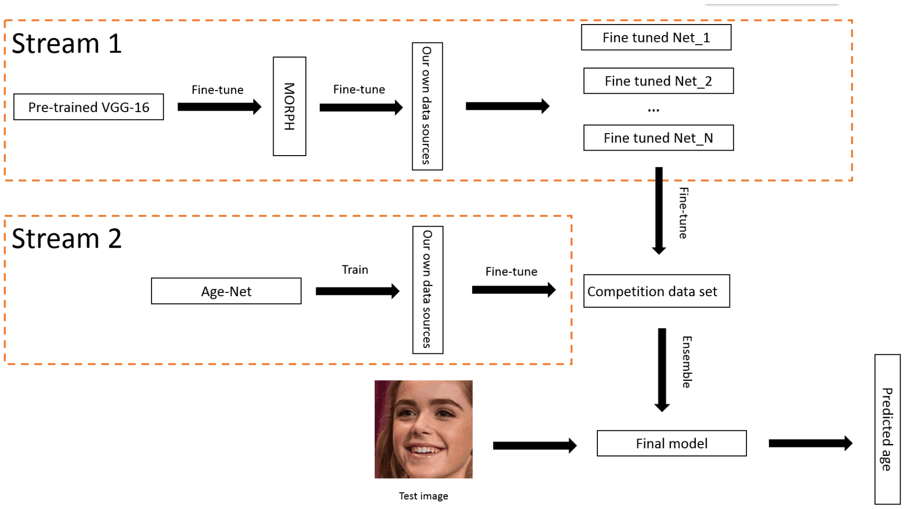

Figure 1: Deeply label distribution learning framework: Two streams for age estimations.

3 Data preprocession

- Face Detection

  We used Voc-DPM for face detection, which is available at `https://bitbucket.org/rodrigob/doppia/src/tip/src/applications/dpm_face_detector/?at=preparing_v2`.

- Key point detection

  We employed Code_point as facial point detectors. Please note that, this facial point detector is ONLY implemented in MATLAB2014a on Windows system platform. Its code is available at `http://mmlab.ie.cuhk.edu.hk/archive/CNN/data/code_point.zip`.

- Face alignemnt

  We would like to extend our sincere gratitude to Key Lab of Intelligent Information Processing (IIP), Chinese Academy of Sciences for the help of face alignment. Its code is available at `http://vipl.ict.ac.cn/resources`.

- Representative image / diagram of the method

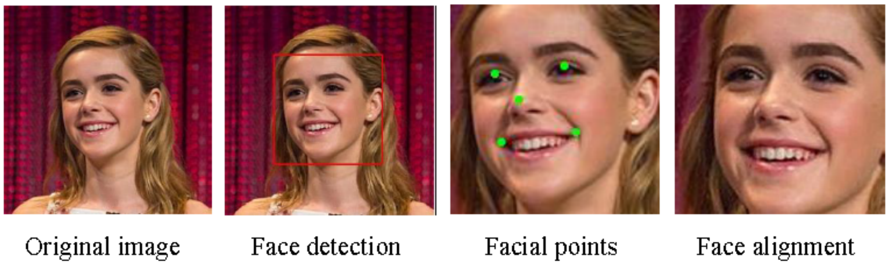

Figure 2: The face region of each image was detected by the DPM model described in [1]. Then the detected face was feed into a public available facial point detector software, i.e., [2], to detect five facial key points, including the left/right eye centers, nose tip and left/right mouth corners. Finally, based on these facial points, we employed face alignment for these facial images.

4 Age distribution learning

The age distribution is composed of description degree of each age. The discription degree is between 0 and 1, and all of which is sum to 1. The KL-divergence is used to measure the difference between the ground truth distribution (generated by the provided ages and variances) and the predicted age distribution. Our learning is to minimize the difference and we call it *age distribution learning*, which can predict the whole distribution with description degrees of all the ages.

5 The first stream for age estimation

The first stream which shown in Fig. 1 mainly used the popular pre-trained CNN model, i.e., VGG-16, as the deep model for age estimation. Here we will present some details about the procedure in this stream.

- First fine-tuning on MORPH

  As aforementioned, we firstly fine-tuned VGG-16 on the pre-processed MORPH data set. In this step, the loss function of VGG-16 was still the softmax loss function, while its number of outputs was changed into the number of classes of MORPH.

- Fine-tuning again on our own data sources

  In this step, we downloaded 28,000 facial images from Google and 30,000 facial images from Bing and Baidu. After that, we fine-tuned the deep model obtained in the first step on these two facial image data sources, respectively. Then we can get two deep models and the loss functions of them were still softmax loss.

- Last fine-tuning on the competition images

  Finally, we used the competition images to fine-tune these two models. For each model, on one side, we still employed the softmax loss; on the other side, we changed the loss function into KL-divergence based on Label Distribution Learning. Therefore, in this stream, we can obtain four deep models for age estimations.

- Multiple VGG-16 nets ensemble

  In this step, we concatenated the features extracted from the output layers of these four deep models as the final facial image representations. After that, we did the low-level ensemble by using the following approach:

  $$t = \sum_{n=1}^{N} t_n \times K(x', x_n), \tag{1}$$

  where

  $$K(x', x_n) = \exp(-\alpha \times ||x' - x_n||_2^2). \tag{2}$$

  By given a training set $S = \{(x_1, t_1), (x_2, t_2), \ldots, (x_n, t_n)\}$, where N is the number of the training images, x_n is the concatenated representations of the n^{th} facial image and t_n is the corresponding ground truth, and x' is representations of the testing image. In addition, $\alpha > 0$ can be chosen via the validation set. Finally, we can get the low-level ensemble prediction results for the testing images, i.e., t in the formula (1).

6 The second stream for age estimation

The second stream in our method was based on a new CNN architecture shown in Fig. 3. Here we present some details about the architecture and procedures as follows.

- The architecture of Age-Net

  As shown in Fig. 3, a 256 × 256 alignment facial image is presented as input. The convolutional filter is $11 \times 11 \times 32$ with stride 2, followed by a 3 × 3 max-pooling layers with stride 2. Moreover, batch normalization is applied to the resulting feature maps. Then, they are passed through a parameter ReLU (not shown) and pooled. Similar operations are repeated in layers 2, 3 and 4. The last two layers of Age-Net are fully connected layers. The final layers is an 85-dimention KL-divergence loss function.

- Pre-training and fine-tuning

  For Age-Net, we downloaded 16,146 facial images from 1-years-old to 85-years-old by search engines including Google, Bing and Baidu. In addition, FG-Net which contains 1,002 facial images is also used in our

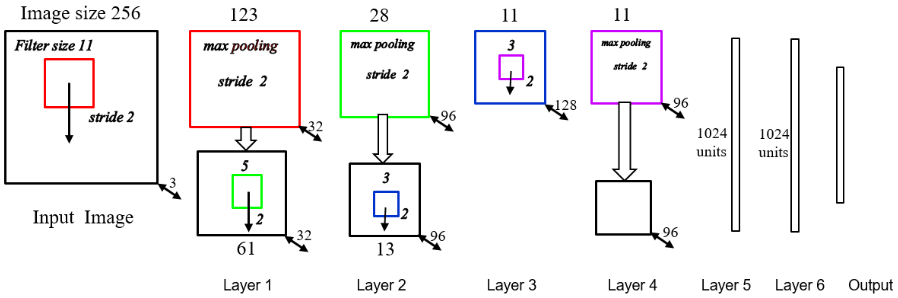

Figure 3: Architecture of the seven layers Age-Net model.

model. After that, we trained Age-Net on the 22,100 images including 16,146 Internet images, 1,002 FG-Net images and $2,476 \times 2$ ChaLearn Age images (with/without alignment). At the fine-tuning stage, we only fine-tuned Age-Net by utilizing 2,476 training and 1,136 validation images from the competition, and then obtained the age predictions of the final testing images.

- Multiple Age-Nets ensemble

  We combined the power of multiple networks trained by feeding different types of inputs to the Age-Net: 1) The first type of inputs were the aligned facial images in RGB colorspace; 2) The gray-level images with images gradient magnitudes and orientations; 3) The single channel gray-level images; 4) The HSV colorspace images; 5) The three channels images obtained by sobel filter and 6) The sobel-level images with the ones in RGB colorspace. In testing time, an alignment face image was operated by a random horizontal scale variants and flipped with 50 times. Thus 300 results were obtained from the above six type Age-Nets. Finally, we gave age estimations by averaging these 300 predicted results.

7 Other details

- Final ensemble

  After the low-level ensemble on each stream, we can get the high-level ensemble results from these two. If the predictions of the two streams are within 11 years difference, we averaged their predictions as the final predicted results; if not, then we take the predictions of the first stream as the final results.