Deeply End to End Learning for Robust Apparent Face Age Estimation

September 15, 2015

1 Team details

- Team name:
  ICT-VIPL
- Team leader name:
  Xin Liu
- Team leader address, phone number and email
  No.6 Kexueyuan South Road Zhongguancun,Haidian District Beijing,China, +86 18810460195, xin.liu@vipl.ict.ac.cn
- Rest of the team members:
  Shaoxin Li, shaoxin.li@vipl.ict.ac.cn
  Meina Kan, meina.kan@vipl.ict.ac.cn
  Jie Zhang, jie.zhang@vipl.ict.ac.cn
  Shuzhe Wu, shuzhe.wu@vipl.ict.ac.cn
  Hu Han, hu.han@vipl.ict.ac.cn
  Shiguang shan, shiguang.shan@vipl.ict.ac.cn
- Affiliation
  Visual Information Processing and Learning Group, Institute of Computing Technology, Chinese Academy of Science.

2 Contribution details

- Title of the contribution
  Deeply End to End Learning for Robust Apparent Face Age Estimation.
- Final score
  The best performance of our method on the validation set is 0.2873, and the performance on the test set is still unknown yet.

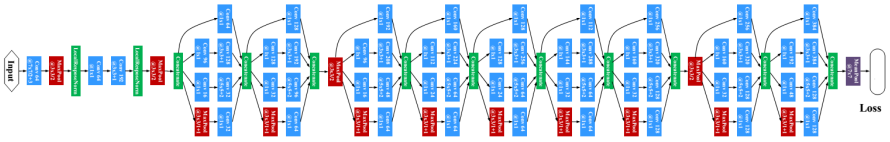

Figure 1: The diagram of the 22 layer deep convolution network

- General method description
  Our approach is a totally end to end learning framework based on general to specific deep transfer learning, and the main steps are:

  1) Pre-train 22 layer large-scale deep convolutional neural network for multi-class face classification using the CASIA-WebFace database [1].

  2) Fine-tune 22 layer large-scale deep convolutional neural network for age estimation on large outside age dataset. In this work, two kind of loss are involved. We adopt Euclidean loss for singe dimension age encoding and cross-entropy loss of label distribution learning [2] based age encoding.

  3) Fine-tune 22 layer large-scale deep convolutional neural network on the final apparent age training set.

  4) Ensemble Learning, the final age estimation output is the fusion of 10 deep neural network.

  In Figuge1, we demonstrate the architecture of the 22 layer deep convolutional neural network, which is modified from the well-known GoogleLet [3]. The loss layer is depended on the task. For multi-class face classification, we adopt the softmax loss, and for the age estimation task, we adopt Euclidean loss and cross-entropy loss.

- Representative image / diagram of the method
  In Figuge2, we present the deeply end to end learning for robust apparent age estimation. The proposed approach adopts very large scale deep neural network. To reduce the risk of overfitting on the small scale apparent age training set, we introduce a general to specific deep transfer learning strategy. Firstly, multi-class face classification models are pre-trained, then real-age data are fine-tuned. Finally, the apparent age training data are fine-tuned in the real-age network to get the end-to-end apparent age estimation results.

- Preprocess of the face images
  The face images process pipeline includes three steps: face detection, face landmarks detection and face normalization. After the preprocess, each face images is normalized to 256x256 pixels.

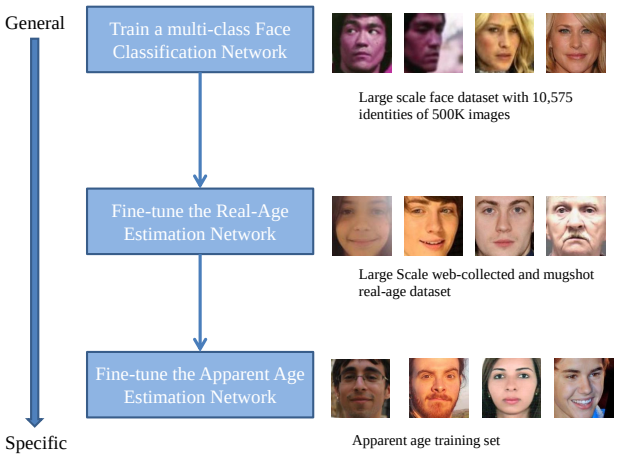

Figure 2: Overview of the deeply end to end learning for robust apparent age estimation

3 Face Detection Stage

In face detection stage, we adopt a commercial face detection toolkit developed by ICT-VIPL. The general process will be: Given an input image, face detection is performed in the sliding window paradigm and the image is resized to various scales in order to detect faces of different sizes. Candidate windows are first evaluated by traditional boosting classifiers so that the majority of non-face regions are removed, while retaining high recall of faces. The surviving windows are then classified by the more complex and more powerful neural networks, giving more accurate face predictions. The final detections are obtained by merging all windows classified as face, ensuring that each face corresponds to only one bounding box.

4 Face Landmarks Detection Stage

We apply the Coarse-to-Fine Auto-Encoder Networks (CFAN) [4] to detect the five facial landmarks in the face: the left and right center of the eyes, the nose tip, the left and right corner of mouth.

5 Global Method Description

- Total method complexity: all stages
  Our method is based on very deep convolutional neural network, so it is very

Table 1: Outside training data for the proposed approach.

Train Set	Short description
CASIA-WebFace[1]	500,000 images of 11,575 identities
CACD [5]	163,446 images of age 14 to 62
WebFaceAge(Private)	60,000 images of age 1 to 85
Morph [6]	55,135 images of age 16 to 77

computation expensive and must run on latest NVidia GPUs, such as Titan X 12G.

- Which pre-trained or external methods have been used
  We have two pre-train steps in the proposed approaches. The first pre-train step is the multi-class face classification deep network and the second pre-train step is the real-age estimation deep network.

- Which additional data has been used in addition to the provided ChaLearn training and validation data (at any stage, if any)

  In Table 1, we present the four outside training sets in the proposed approach. The CASIA-WebFace dataset is adopted in the pre-train stage of the multi-class face classification. CACD, Morph and a private web-collected age dataset are deployed to finetune the real-age network.

- Qualitative advantages of the proposed solution
  Our approach achieves 0.2873 mean error rate in the evaluation set.

- Novelty degree of the solution and if is has been previously published
  The novelty of our method can be summarized as:

  1) Our approach is totally end to end, and a general to specific deep adaptation learning strategy is adopted.

  2) We adopt very large scale deep convolutional neural network for face age estimation and proves it efficiency in apparent age estimation problem.

  Our method has not been published yet.

6 Other details

- Language and implementation details (including platform, memory, parallelization requirements)
  Face Detection: Windows executable binary fie.

  Face Landmarks Detection Sage: Windows Matlab scripts.

  Deep network for age estimation: Linux C++ code.

  Mode ensemble stage: Windows Matlab scripts.

- Human effort required for implementation, training and validation?
  No human effort need for the implementation except coding. No human labeling is involved in any step of the proposed approach.

- Training/testing expended time?
  Pretrain-1: It takes 3 days in Titan-X 12G per model, and we have four models of two face normalization and two different crop size.

  Finetune-1: It takes 1.5 days in Titan-X 12G per model, and we have 10 models in total.

  Finetune-2: It takes 3.5 hours in Titan-X 12G per model, and we have 10 models in total.

  Totally, it takes 28.5 days to train all the models in a single Titan-X 12G device. In this work, we apply four Titan-X GPUs in total, so all the experiments can run in one week.

- General comments and impressions of the challenge? what do you expect from a new challenge in face and looking at people analysis?
  We expect to see the final score on the final test set when we submit the results.

References

[1] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[2] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 10, pp. 2401–2412, 2013.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.

[4] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *ECCV2014*, pp. 1–16, Springer, 2014.

[5] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Computer Vision–ECCV 2014*, pp. 768–783, Springer, 2014.

[6] K. Ricanek Jr and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pp. 341–345, IEEE, 2006.