A Call to Reflect on Evaluation Practices for Age Estimation: Comparative Analysis of the State-of-the-Art and a Unified Benchmark

Supplementary Material

A. Literature Survey

To thoroughly survey the CVPR and ICCV, we adhered to the following procedure. We systematically searched for all papers presented at the CVPR or ICCV conferences from 2013 onwards that encompassed keywords such as "age", "aging", "face", "facial", and "ordinal" in their titles. We excluded papers that primarily focused on face detection, recognition, or editing, as our specific focus was on age estimation. However, we retained papers that addressed the learning of facial representations (such as unsupervised pretraining and clustering), as they could evaluate the quality of their representations on the age estimation task.

Subsequently, we meticulously reviewed the remaining papers to determine whether they centered around age estimation. We documented the datasets employed by these papers and their data partitioning strategies. Additionally, we extended our survey to include age estimation literature from other conferences that were referenced by the aforementioned papers.

Our estimates, based on the surveyed literature, indicate that approximately 70% of the papers that attempt to improve age estimation follow this approach. Only a minority of these papers adequately ablate the impact of the proposed modifications. Most of the remaining papers suggest modifications to the training procedure, backbone, or aspects of the data pipeline.

B. Compared Methods

This paper compares various recent age estimation methods utilizing feedforward neural networks which receive an image $x \in \mathcal{X}$ and output an age estimate $y \in \mathcal{Y}$. We focus solely on age estimation methods that modify the standard classification approach by changing the last few layers of the neural network or the loss function. Although this may appear restrictive, it is essential to note that a majority of the methods proposed in the field fall into this category. By comparing methods that modify only a small part of the network, we aim to ensure a fair evaluation, as the remaining setup can be kept identical. Some recent methods, such as Moving Window Regression proposed by Shin *et al*. [17], were therefore omitted from this study.

Traditionally, age estimation relied on classification and regression-based approaches. However, these methods often overlook the inherent ordinal nature of age. In multi-class classification, misclassifications are treated equally, even though some age predictions may be more accurate than others. On the other hand, regression approaches can predict nonsensical and even negative age values. Ordinal regression has therefore emerged as a well-motivated approach to address these limitations. Unlike classification, where the labels merely represent categories, ordinal regression utilizes labels that provide sufficient information to order the objects. Below, we provide a concise overview of recent age estimation and ordinal regression approaches.

Classification The conventional classification approach still remains popular in the literature. For instance, Rothe *et al*. [16] achieved victory in the ChaLearn LAP 2015 challenge on apparent age estimation [5] with a model that employed cross-entropy to learn the posterior age distribution.

Extended Binary Classification Niu *et al*. [13] (OR-CNN) follow the approach proposed by Li and Lin [9] and transform the ordinal regression task into multiple binary classification sub-problems. For each age value $y_k \in \mathcal{Y}$, they construct a binary classifier to predict whether the true age $y \in \mathcal{Y}$ of a sample $x \in \mathcal{X}$ is larger than y_k. Cao *et al*. [2] (CORAL) modify this approach by restricting the hypothesis class such that the binary classifier predictions are consistent, i.e., the predicted probabilities satisfy $p(y > y_k|x) \geq p(y > y_{k+1}|x)$; $\forall k$.

Fixed Distribution Learning Gao *et al*. [7] (DLDL) approach the task as multi-class classification. However, they encode the label distribution as a normal distribution centered at the true label. Díaz and Marathe [4] (SORD) approach the task similarly, but encode the label distribution as a double exponential distribution centered at the true label. In a follow up to their work [7], Gao *et al*. [8] (DLDL-v2) propose to also minimize the difference between (i) the true label $y \in \mathcal{Y}$, and (ii) the expectation $\mathbb{E}_{\hat{y}\sim f(x)}[\hat{y}]$ of the model output distribution $f(x)$.

Adaptive Distribution Learning An approach emerging in recent years is not to model a specific distribution, such as normal or double exponential distribution, but instead, to constrain the model by some statistical measure or a condition. Pan *et al*. [14] (Mean-Variance) approach the task as standard multi-class classification, but design a loss function that (i) minimizes the squared difference between the expectation $\mathbb{E}_{\hat{y}\sim f(x)}[\hat{y}]$ and the true label $y \in \mathcal{Y}$, and (ii) minimizes the variance $\mathbb{E}_{\bar{y}\sim f(x)}\left[(\bar{y} - \mathbb{E}_{\hat{y}\sim f(x)}[\hat{y}])^2\right]$ of

Evaluation Dataset	Annotations	
	EM-CNN [6]	FP-AGE [11]
AgeDB	6.44	**6.30**
AFAD	**6.86**	7.23
CACD2000	**5.81**	5.90
CLAP2016	6.24	**5.53**
FG-NET	10.32	**6.09**
MORPH	**4.94**	5.30
UTKFace	8.31	**6.26**
Overall	**6.28**	6.36
IMDB	4.90	5.15

Table 1. MAE ↓ of ResNet-50 trained on IMDB-WIKI with clean age labels from (i) EM-CNN [6], and (ii) FP-AGE [11]. Results on IMDB-WIKI are not included in the *Overall* result.

the model output distribution $f(x)$. Similarly, Li *et al*. [10] (Unimodal) design a loss function (i) which constrains the model to output unimodal distributions, and (ii) concentrates the output distribution around the true label $y \in \mathcal{Y}$.

Note on Prediction Strategy Note that for all methods which model the posterior distribution $p(y|x)$, namely (i) cross-entropy, (ii) DLDL [7], (iii) DLDL-v2 [8], (iv) SORD [4], (v) Mean-Variance loss [14], and (vi) Unimodal loss [10], we use the optimal plugin Bayes predictor for MAE loss, i.e., we predict $\arg\min_y \mathbb{E}_{\hat{y}\sim f(x)}[|y - \hat{y}|]$. For regression, we use the absolute error as the loss function.

C. Additional Comments

Pre-training For some experiments, we pre-train the models on IMDB-WIKI. However, it is important to note that the labels (bounding box, identity, age) in the IMDB-WIKI dataset are known to be noisy. To mitigate this problem, Lin *et al*. [11], and Franc and Čech [6] attempted to clean the labels. To assess the quality of these labels, we trained ResNet-50 models on the dataset using the labels proposed by Lin *et al*. [11] and Franc and Čech [6], and evaluated the models' performance on the other datasets [1, 3, 12, 13, 15, 18]. The results are presented in Tab. 1. Both models achieved similar results, so the choice of labels between [6, 11] is in our opinion arbitrary. Due to a slightly lower overall Mean Absolute Error (MAE), we decided to use the labels from Franc and Čech [6] for model pre-training in this paper.

Comment on Task Uncertainty The irreducible Bayes error of the age estimation task is contingent on the specific formulation. When estimating the *real age* from an observation x_i, the label y_i is a realization of the distribution $p(y \mid x)$, a random variable. The observation x_i does not contain all the necessary information about the person's genetics, lifestyle, etc., and the Bayes error is non-zero. Interestingly, when estimating the *apparent age*, the Bayes error can be 0. Specifically, when the label is defined as $y_i = \mathbb{E}_{\hat{y}_i\sim p(y \mid x)}[\hat{y}_i]$, which is *not* a random variable. Apparent age is by definition the expected annotation provided by observers. Therefore the observation x_i must necessarily contain all the information used to generate the label. The process of dataset collection provides further insight. For AgeDB, AFAD, and CACD2000, the age label is defined as the *year when the photo was taken* minus the *year of birth*. Two individuals born on December 31st and January 1st would thus have different age labels, even though they were born just a day apart. The *year taken* and the *year born* can also be noisy. For instance, AFAD collects data from a social network with a minimum age requirement of 15, without verifying the actual age of the users. CLAP2016, FG-NET, and MORPH use the legal age as the age label. Still, uncertainty arises from the discrete nature of the label. For UTKFace, the labels are estimated by DEX [16], and manually verified. Therefore, DEX should attain zero error on UTKFace.

Implementation Our implementation, including data annotations and splits, is available in the supplementary material. Due to file size limitations, annotations for IMDB-WIKI are not included. All materials mentioned are also accessible on our GitHub repository: https://github.com/paplhjak/Facial-Age-Estimation-Benchmark.

References

[1] E. Agustsson, R. Timofte, S. Escalera, X. Baro, I. Guyon, and R. Rothe. Apparent and real age estimation in still images with deep residual regressors on appa-real database. In *12th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2017*. IEEE, 2017. 2

[2] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140: 325–331, 2020. 1

[3] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *Computer Vision – ECCV 2014*, pages 768–783, Cham, 2014. Springer International Publishing. 2

[4] Raúl Díaz and Amit Marathe. Soft labels for ordinal regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4733–4742, 2019. 1, 2

[5] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzàlez, Hugo Jair Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. Chalearn looking at people

2015: Apparent age and cultural event recognition datasets and results. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 243–251, 2015. 1

[6] Vojtech Franc and Jan Cech. Learning cnns from weakly annotated facial images. *Image and Vision Computing*, 2018. 2

[7] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017. 1, 2

[8] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 712–718. International Joint Conferences on Artificial Intelligence Organization, 2018. 1, 2

[9] Ling Li and Hsuan-Tien Lin. Ordinal regression by extended binary classification. In *Advances in Neural Information Processing Systems*, page 865 – 872, 2007. Cited by: 195. 1

[10] Qiang Li, Jingjing Wang, Zhaoliang Yao, Yachun Li, Pengju Yang, Jingwei Yan, Chunmao Wang, and Shiliang Pu. Unimodal-concentrated loss: Fully adaptive label distribution learning for ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20513–20522, 2022. 2

[11] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Fp-age: Leveraging face parsing attention for facial age estimation in the wild. *IEEE Transactions on Image Processing*, 2022. 2

[12] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017. 2

[13] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4920–4928, 2016. 1, 2

[14] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5285–5294, 2018. 1, 2

[15] K. Ricanek and T. Tesafaye. Morph: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345, 2006. 2

[16] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 252–257, 2015. 1, 2

[17] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: A novel approach to ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18760–18769, 2022. 1

[18] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 2