Fair and accurate age prediction using distribution aware data curation and augmentation

Yushi Cao[1,a*], David Berend[1,a], Palina Tolmach[1], Guy Amit[2], Moshe Levy[2]
Yang Liu[1], Asaf Shabtai[2], Yuval Elovici[2]
[1]Nanyang Technological University, [2]Ben-Gurion University of the Negev
{yushi002, bere0003, palina0001}@e.ntu.edu.sg, {guy5, moshe5}@post.bgu.ac.il
yangliu@ntu.edu.sg, {shabtaia, elovici}@bgu.ac.il

A. Appendix

A.1. Calculation of OOD Score

To extract an OOD score, FOOD creates a copy of a trained DNN model and replace the last fully-connected layer with a Gaussian likelihood layer. Usually, the DNN model is trained for a few more iterations to optimize the weights of the final layer [2]. To make it more lightweight and enable its integration, we adjust the technique such that it can be integrated in any workflow of an age prediction system without requiring additional training.

The final Gaussian likelihood layer receives the output of the penultimate DNN model layer as input. The penultimate layer is commonly used for analysis, as it contains the most processed information without limiting the feature space. With the help of the Gaussian layer, the data is represented as a multivariate Gaussian with two parameters: a center vector and a co-variance matrix. Given our adjustment, those two parameters can be directly calculated based on the training data for each class. For the class c and penultimate representations of the dataset X, we calculate the center μ_c and the co-variance Σ_c as follows:

$$\mu_c = \frac{1}{|c|} \sum_{x_i \in c} x_i \tag{1}$$

$$\Sigma_c = \frac{1}{|c|} \sum_{x_i \in c} (x_i - \mu_c)(x_i - \mu_c)^T \tag{2}$$

with the d-dimensional penultimate representation, where $\mathcal{N}$ stands for the multivariate Gaussian distribution, as shown in Equation 3.

$$f(x|\Sigma_c;\mu_c) = \log\left(\mathcal{N}(x|\mu_c;\Sigma_c)\right) = -\frac{d}{2}log(2\pi) - \frac{1}{2}log(|\Sigma_c|) - \frac{1}{2}(x-\mu_c)^T\Sigma_c^{-1}(x-\mu_c) \tag{3}$$

The closer a sample is to the class center, the higher the confidence that the input belongs to a certain class and to the trained distribution.

We calculate the OOD scores using a log-likelihood ratio ($\mathcal{LLR}$) test on the subtraction of two log-likelihood scores (Equation 4). The test takes the probability ratio between the log of the predicted class and the logs of the unpredicted classes, where K represents the group of the k class indices which do not belong to the ground truth and have the top likelihood scores $\hat{y}$.

$$\mathcal{LLR} = \max_{c \in \{1,\ldots,C\}} f(x|\mu_c;\Sigma_c) - \frac{1}{k}\sum_{i \in K} f(x|\mu_k;\Sigma_k) \tag{4}$$

The test provides an estimate that measures how far away the sample is from its predicted class in the penultimate representation. Samples that are too far away from their predicted class relative to other classes are given a low LLR, which translates into a high OOD score.

A.2. Generalization Datasets

Neither the CACD [4] or the AFAD dataset[13] are integral to the training or testing set; both show difference in style to the training and testing set and are collected by different sources. The CACD contains $163,446$ facial images of $2,000$ celebrities; for the AFAD, we opt to use the light version which contains $60,000$ facial images collected from various Internet sources. Table A1 shows the summary of all the datasets we used. The preprocessing workflow was applied to both datasets before testing (Section 4.1).

A.3. Calculation of Fairness Score

Function K (Equation 7) is an indicator function that indicates fairness for one sensitive feature pair s_j and s_k ($k \neq j$) when the average predicted ages $P(s_j|y_i)$ and $P(s_k|y_i)$ are close enough to each other defined by threshold t divided by 2 given the absolute value. $P(s_j|y_i)$ represents the average

Table A1: Summary of datasets.

Name	Purpose	Size	Related Work
IMDB-WIKI	Pretraining	636,022	[14, 20, 21]
MORPH-2	Curation&Augmentation	55,000	[20, 21, 19, 10]
APPA-REAL	Curation&Augmentation	7,591	[1, 7, 10]
UTKFace	Curation&Augmentation	20,000	[23, 9]
Mega Asian	Curation&Augmentation	40,000	[22, 19]
AFAD	Validation	164,432	[13, 5]
CACD	Validation	163,446	[4, 17, 14]

predicted age at sensitive feature s_j, given actual age y_i. Function F is another indicator function that indicates for age y_j, if the distance of average predicted age of every pair of sensitive features are close enough to each other. Therefore, F represents the overall distribution of how often the DL system performs fairly one age. Finally, p summarised all ages by taking the ratio of those ages which were considered fair by F and all ages together.

$$p = \frac{1}{n}\sum_{i=1}^{n} F(y_i|\mathbf{s}) \tag{5}$$

$$F(y_i|\mathbf{s}) = \mathbb{1}\left(\left(\sum_{j\neq k} K(s_k, s_j|y_i)\right) = C_m^2\right) \tag{6}$$

$$K(s_k, s_j|y_i) = \mathbb{1}\left(|P(s_k|y_i) - P(s_j|y_i)| < \frac{t}{2}\right) \tag{7}$$

A.4. Comparison of Augmentation Approaches

In prior research, data augmentation has been assessed by identifying which augmentation types produce sufficiently diverse data for a DL system. [8] studied different sets of augmentation combinations to maximize diversity [8], named *AutoAugment*, which is used in various prior research [16, 15, 12]. In the field of contrastive learning, it was found that some augmentations are beneficial when combined while others are not [6]. As a result, the authors propose an augmentation setting used in contrastive learning to minimize the distance among augmentations from the same images while maximizing the distance among different images to determine best augmentation practices. This augmentation type is named *SimCLR*. The third augmentation type utilizes both affine and color augmentations and follows prior research [3, 18, 11] by empirically assessing the boundaries of individual augmentations to control the diversity and realism, named *Fine-grained*. Figure 1 shows the differences among different augmentation types. Table A2 shows that Fine-grained method performs the best among all the settings and we mainly opt for this augmentation techniques.

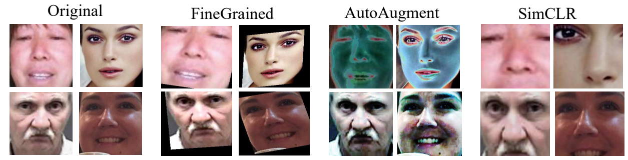

Figure 1: Example augmentations depending on augmentation type.

Table A2: Augmentation results comparing no augmentation setting to the presented augmentation types in Section 4.4.

Type	CACD↓	AFAD↓	Ethnicity↑	Gender↑
None	4.77	7.11	70.50	**81.00**
Fine-grained	**4.53**	**7.01**	**73.50**	80.00
AutoAugment	5.04	7.30	69.50	74.00
SimCLR	4.58	7.04	72.00	69.00

A.5. The pesudocode of data curation

Algorithm 1: Curating a diverse and sensitive feature balanced dataset

Result: Curated dataset
num_sample ←Sum of number of samples from all datasets by class C and state S;
sort(num_sample by s);
$max\_sample \leftarrow min\{quantile(num\_sample_{c,s}, 0.8)|s \in S\}$;
$min\_sample \leftarrow max\{quantile(num\_sample_{c,s}, 0.2)|s \in S\}$;
for *all* $c \in C$ **do**
    $threshold \leftarrow min\{num\_sample_{c,s}|s \in S\}$;
    $threshold \leftarrow min(max\_sample, max(min\_sample, threshold))$;

    $ds\_num \leftarrow$ number of datasets;
    $select\_size \leftarrow threshold/ds\_num$;
    for *all* $s \in S$ **do**
        sort(D,c,s);
        for *all* $d \in D$ **do**
            num ←length of $d_{c,s}$;
            if $num < select\_size$ **then**
                select all data in $d_{c,s}$;
                $remain \leftarrow select\_size - num$;
                update($select\_size, remain$) ;
            else
                random_select($d_{c,s}, select\_size$);
            end
        end
    end
end

A.6. Cross-analysis results

Table A3: Individual cross-analysis results retrieved on prior research DL age prediction system approaches (Section 4.2).

DNN	Train	Test	MAE↓
AlexNet	APPA	APPA	7.6
		Megagsian	11.8
		MORPH	6.5
		UTKFace	7.7
		Average (others)	**8.6**
	Megagsian	Megagsian	3.6
		APPA	11.5
		MORPH	8.3
		UTKFace	9.4
		Average (others)	**9.7**
	MORPH	MORPH	2.9
		APPA	11.7
		Megagsian	9.4
		UTKFace	10.6
		Average (others)	**10.5**
	UTKFace	UTKFace	5.3
		APPA	9.6
		Megagsian	8.3
		MORPH	7.9
		Average (others)	**8.6**
DEX VGG	APPA	APPA	7.0
		Megagsian	12.2
		MORPH	6.4
		UTKFace	7.9
		Average (others)	**8.8**
	Megagsian	Megagsian	6.5
		APPA	11.4
		MORPH	7.3
		UTKFace	10.7
		Average (others)	**9.8**
	MORPH	MORPH	2.5
		APPA	10.6
		Megagsian	8.4
		UTKFace	10.0
		Average (others)	**9.7**
	UTKFace	UTKFace	5.2
		APPA	8.4
		Megagsian	7.8
		MORPH	6.4
		Average (others)	**7.5**

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] Guy Amit, Moshe Levy, Ishai Rosenberg, Asaf Shabtai, and Y. Elovici. Food: Fast out-of-distribution detector. 2020.

[3] David Berend, Xiaofei Xie, Lei Ma, Lingjun Zhou, Yang Liu, Chi Xu, and Jianjun Zhao. Cats are not fish: Deep learning testing calls for out-of-distribution awareness. In *The 35th IEEE/ACM International Conference on Automated Software Engineering*, New York, NY, USA, 2020. Association for Computing Machinery.

[4] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[5] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5183–5192, 2017.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] A. Clapés, G. Anbarjafari, O. Bilici, D. Temirova, E. Avots, and S. Escalera. From apparent to real age: Gender, age, ethnic, makeup, and expression bias analysis in real age estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2436–243609, 2018.

[8] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *CoRR*, abs/1805.09501, 2018.

[9] Abhijit Das and Antitza Dantcheva. Mitigating bias in gender, age, and ethnicity classification: a multi-task convolution neural network approach. 10 2018.

[10] Fadi Dornaika, Ignacio Arganda-Carreras, and C Belver. Age estimation in facial images through transfer learning. *Machine Vision and Applications*, 30(1):177–187, 2019.

[11] Jianmin Guo, Yu Jiang, Yue Zhao, Quan Chen, and Jiaguang Sun. Dlfuzz: Differential fuzzing testing of deep learning systems. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2018, page 739–743, New York, NY, USA, 2018. Association for Computing Machinery.

[12] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4182–4192. PMLR, 13–18 Jul 2020.

[13] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016.

[14] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015.

[15] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[16] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[17] J. Wan, Z. Tan, Z. Lei, G. Guo, and S. Z. Li. Auxiliary demographic information assisted age estimation with cascaded structure. *IEEE Transactions on Cybernetics*, 48(9):2531–2541, 2018.

[18] Xiaofei Xie, Lei Ma, Felix Juefei-Xu, Minhui Xue, Hongxu Chen, Yang Liu, Jianjun Zhao, Bo Li, Jianxiong Yin, and Simon See. Deephunter: A coverage-guided fuzz testing framework for deep neural networks. In *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA 2019, page 146–157, New York, NY, USA, 2019. Association for Computing Machinery.

[19] Tsun-Yi Yang, Yi-Hsuan Huang, Yen-Yu Lin, Pi-Cheng Hsiu, and Yung-Yu Chuang. Ssr-net: A compact soft stagewise regression network for age estimation. In *IJCAI*, volume 5, page 7, 2018.

[20] Chao Zhang, Shuaicheng Liu, Xun Xu, and Ce Zhu. C3ae: Exploring the limits of compact model for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12587–12596, 2019.

[21] Ke Zhang, Na Liu, Xingfang Yuan, Xinyao Guo, Ce Gao, Zhenbing Zhao, and Zhanyu Ma. Fine-grained age estimation in the wild with attention lstm networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[22] Yunxuan Zhang, Li Liu, Cheng Li, and Chen Change Loy. Quantifying facial age by posterior of age comparisons. In *British Machine Vision Conference (BMVC)*, 2017.

[23] 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.