Benchmarking Robustness to Natural Distribution Shifts for Facial Analysis

Anonymous Author(s)
Affiliation
Address
email

Abstract

During the deployment of machine learning models, performance degradation can occur compared to the training and validation data. This generalization gap can appear for a variety of reasons and be particularly critical in applications where certain groups of people are disadvantaged by the outcome, e.g. facial analysis. Literature provides a vast amount of methods to either perform robust classification under distribution shifts or at least to express the uncertainty caused by the shifts. However, there is still a need for data that exhibit different natural distribution shifts considering specific subgroups to test these methods. We use a balanced dataset for facial analysis and introduce subpopulation shifts, spurious correlations, and subpopulation-specific label noise. This forms our basis to investigate to what extent known approaches for calibrating neural networks remain reliable under these specified shifts. Each of the modifications leads to performance degradation, but the combination of ensembles and temperature scaling is particularly useful to stabilize the calibration over the shifts.

1 Introduction

Machine Learning is a central tool for many tasks in the area of computer vision and facial analysis. However, most approaches are evaluated on data that are identically and independently distributed (i.i.d.). This i.i.d. assumption often cannot be guaranteed during deployment. Thus, a relatively wide range of possible shifts between training and inference can occur that have a detrimental effect on the generalization performance. This can be particularly important in areas where predictions of machine learning systems have a direct effect on humans, as it is the case with facial analysis. Subpopulations may be underrepresented in the training data compared to the overall population [8, 14], or there may be spurious correlations between subpopulations and labels during training that have no real causal relationship [1, 13]. Since this poses a significant challenge to the development of real-world machine learning and pervades all areas and applications of machine learning, efforts have been made to mitigate this problem [1, 13, 3, 11, 10]. Furthermore, it is also desirable to get well-calibrated uncertainty estimates, that are robust across shifts such that we can at least determine when to trust the predictions of the model [12]. In general, however, we need data sets that go beyond the classic train/validation splits from i.i.d data to evaluate these methods. Thus, several benchmarks have been presented to evaluate several shifts across different domains [8, 14]. For facial analysis Sagawa et al. [13] generate spurious correlations using annotated attributes in CelebA. However, to the best of our knowledge, no structured benchmark currently exists to evaluate different types of naturally occurring shifts in the domain of facial analysis.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Consequently, our first contribution is a framework to systematically create various subpopulation-specific shifts on facial data: underrepresented to missing subpopulations, spurious correlations, and group-specific label noise. For this, we use FairFace [6] (a highly balanced dataset in terms of age, race and gender) as a base dataset. Furthermore, we want to address the following research question: What impact do these shifts have on the subpopulation-specific accuracy and to what extent do they affect the calibration? To answer this question, we compare a baseline model with several known approaches for building well-calibrated models. For reproducibility, we provide our code for creating the shifts and training [1].

2 Distribution shifts in facial data

In the following we describe the underlying data and modifications.

Baseline Data As a basis for our work we use the FairFace dataset [6] with 97,698 subjects (86,744 for training and 10,954 for testing). This dataset contains information on age, gender, and race, with special attention paid to the balance of these three attributes. *Gender* is categorized into two classes, *race* into seven, and *age* into nine. We attempt to estimate the latter by classification. Our goal is to examine the effects of several types of shifts without the influence of the number of training samples. Therefore, we ensure that although we reduce the number of subjects from certain groups, the total number of training samples per shift remains constant. We achieve this through a general random downsampling, also on the baseline data.

Subpopulation shift with one under-represented attribute This change of the dataset denotes a shift in the data distribution with respect to each attribute value respectively, e.g. *male* for *gender* and *East Asian* for *race*, which occurs with decreasing relative frequency p compared to the others. We use $p = \{1.0, 0.75, 0.5, 0.25, 0.0\}$ for the attribute *race* and $p = \{1.0, 0.5, 0.0\}$ for *gender* to investigate different shift intensities for the considered attribute value. Note that the selected group does not occur at the strongest shift.

Label noise In addition to normal subpopulation shifts, there may also be specific biases in the labels for individual groups. For example, certain races are estimated to be younger than they really are. This can have isolated effects for these specific groups or all others during the deployment. To investigate this, we introduce attribute-specific label noise in the training data: for each race group separately, we reduce the age class by subtracting samples from a left-hand side of a Gaussian distribution with mean 0 and increasing standard deviation of $p = \{0.5, 1.0, \ldots, 4.0\}$ from the ground truth labels.

Spurious correlations Our last shift consists of spurious correlations between *race* and the label *age*. Therefore we combine the lowest three age groups into the attribute value *young* and systematically correlate it with each race group in the training data. We consider two variants for this correlation: Sufficiency (e.g. *"if young then Indian"*) and equivalence (e.g. *"iff young then Indian"*, depicted in Appendix A). We create these variants by omitting subjects with certain *race-age* combinations: For example, for the spurious correlation *"if young then Indian"* we omit all young subjects that are not Indian. In the case of *"iff young then Indian"* we additionally remove all non-young Indians. Also, we created a baseline with a reduced number of young subjects over all races to match the overall number of young subjects to isolate the effect of spurious correlations and exclude that of data set size (for young).

3 Experimental Setup

We compare approaches that are known to improve the calibration, such as deep ensembles (3x) [9], mixup [15] and compare it to the simple baseline, referred to as vanilla. Furthermore, we combine

[1] https://anonymous.4open.science/r/DistributionShiftsOnFacialData

each of them with temperature scaling [4], a post-hoc recalibration method. We use a clean and balanced validation set for learning the scaling parameter.

We train a ResNet34 [5] model from five random seeds for each shift and each approach separately with the same parameter settings: Adam [7] optimizer with weight decay of 0.01, learning rate of 0.0001 and 30 epochs. We apply a random crop with of size 224 and RandAugment [2] for data augmentation. In addition to the classification accuracy, we are particularly interested in the reliability of the confidence outputs, which we measure with the ECE (Expected Calibration Error) [4]. It measures the deviation of accuracy and confidence, where a small value indicates reliability.

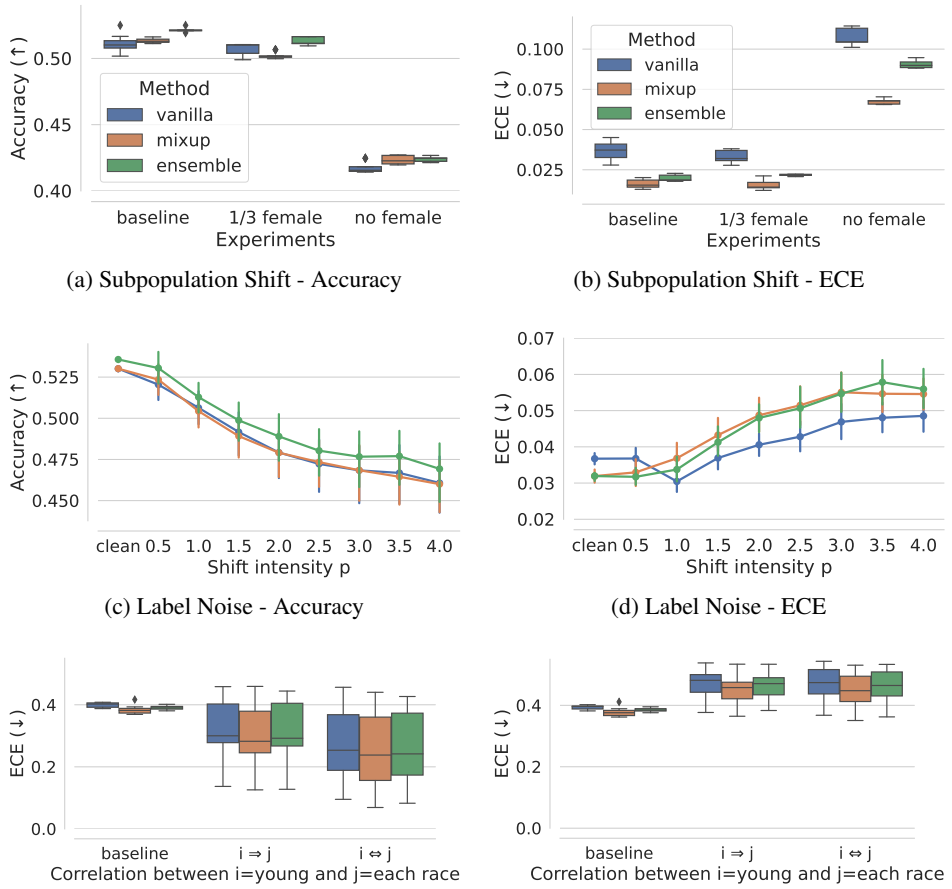
4 Results

In the following, we try to answer the research question for each of the defined shifts. First, we want to address the question of what impact the previously defined changes in the distributions have on the accuracy. We contrast the changes in the accuracy with the stability of the calibration. This allows us to answer the question of whether our model remains at least reliable even when the accuracy deteriorates. For this purpose, we focus on the relevant subpopulations that are affected by the changes, e.g. label noise, and if not explicitly mentioned, the observations refer to the mean performance of the relevant subpopulation. Thus, for each attribute value j, e.g. *Middle Eastern*, that is affected by a shift, we consider the accuracy and the ECE of all subjects of the relevant race j.

Subpopulation Shift For subpopulation shifts we considered both attributes: gender (see Figure 1) and race (see B.1). We can observe that in both cases the accuracy of all methods deteriorates with increasing shift. This effect is more pronounced for the attribute *gender* than for *race* when we compare the strongest shifts. This poses the question of whether the existing race diversity in the dataset is sufficient to learn race-invariant features, making the removal of a single race not very significant. The calibration error increases with stronger shift. Similarly as for the accuracy this effect is larger for the shift in *gender*. Among the compared approaches, ensembles and mixup result in an improvement of the calibration over vanilla; with mixup this effect is larger in the gender case. A recalibration causes a slight stabilization of the baseline, but leads to a deterioration of mixup.

Label Noise In the case of group-specific *label noise* we see a clear drop of accuracy (by 7.5 percentage points) with higher label noise. However, the calibration seems to be relatively stable (max. 1.0 percentage points drop for vanilla). Interestingly, the calibration improves even with slight label noise. An assumption would be that the otherwise overconfident networks become more uncertain, which leads to an improved calibration. We can observe that ensembles as well as mixup have better calibration without label noise, however under label noise the vanilla network outperforms both of them. On the other hand, ensembles have the better accuracy under noise. A recalibration (Appendix B.2) reduces the gap between all methods and leads to a stabilization over the noise intensities. In this case, ensembles show the lowest ECE. We also evaluated the subpopulations that did not encounter noise during training and observed that the accuracy worsens by only 1.5 percentage points for the vanilla network. Therefore, we can conclude that the label noise applied to a single group does not exhibit a similar strong performance degradation for all other groups.

Spurious Correlations We investigate the effect of spurious correlations between *young* and each value of *race*. For the metrics only *young* is considered. Also, we distinguish between the race groups that include the attribute *young* in the training data (young-correlated) and the remaining ones (non-young-correlated). The accuracy increases for the young-correlated groups (by 14 percentage points) and decreases for the non-young-correlated groups (by 7.5 percentage points) with stronger spurious correlations (see Appendix B.3), while the spread between the different young-correlated observations increases. In terms of calibration we observe a significant improvement for the young-correlated groups and a deterioration for non-young-correlated groups. However, we notice that the models are poorly calibrated, thus the uncertainty estimate of the methods are not reliable. As before,



(a) Subpopulation Shift - Accuracy

(b) Subpopulation Shift - ECE

(c) Label Noise - Accuracy

(d) Label Noise - ECE

(e) Spurious Correlation *young-correlated* - ECE for young

(f) Spurious Correlation *non-young-correlated* - ECE for young

Figure 1: Accuracy and ECE for subpopulation shift, label noise and spurious correlations.

temperature scaling improves the calibration. However, unlike for the other shifts, neither ensembles nor mixup can improve the calibration compared to vanilla and no clear ranking can be observed.

5 Conclusion & Discussion

In this paper, we introduced several shifts for a given facial analysis dataset and investigated the impact on the accuracy and calibration. In summary, each type of distributional change has an increasingly negative impact on the accuracy with higher shift intensity. Also the calibration of the methods continues to decline as the shift increases such that the models become less reliable. Ensembles and mixup outperformed vanilla for the subpopulation shift, but for spurious correlations there has not been a clear ranking and for label noise vanilla even performed better than the other methods. For the case that a balanced validation dataset is available, a simple post-hoc recalibration can improve the ECE significantly for all shifts. It stabilizes the calibration for all methods and leads to an advantage of ensembles over the other methods. However, getting a balanced data set poses difficulties in the wild. We did not examine this effect with respect to different modifications of the validation set, thus it remains unclear what significance the specific composition of the validation set might play. We leave this investigation for future work.

References

[1] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2019.

[2] Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3008–3017, 2020.

[3] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *ArXiv*, abs/2007.01434, 2021.

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[5] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[6] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[8] P. W. Koh, Shiori Sagawa, H. Marklund, Sang Michael Xie, Marvin Zhang, A. Balsubramani, Wei hua Hu, Michihiro Yasunaga, Richard L. Phillips, Sara Beery, J. Leskovec, A. Kundaje, E. Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. *ArXiv*, abs/2012.07421, 2020.

[9] Balaji Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, 2017.

[10] Zachary Chase Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors. *ArXiv*, abs/1802.03916, 2018.

[11] Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *ICML*, 2021.

[12] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019.

[13] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *ICLR*, 2020.

[14] Shibani Santurkar, D. Tsipras, and A. Madry. Breeds: Benchmarks for subpopulation shift. *arXiv: Computer Vision and Pattern Recognition*, 2021.

[15] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2018.

A Overview of the introduced Distribution Shifts

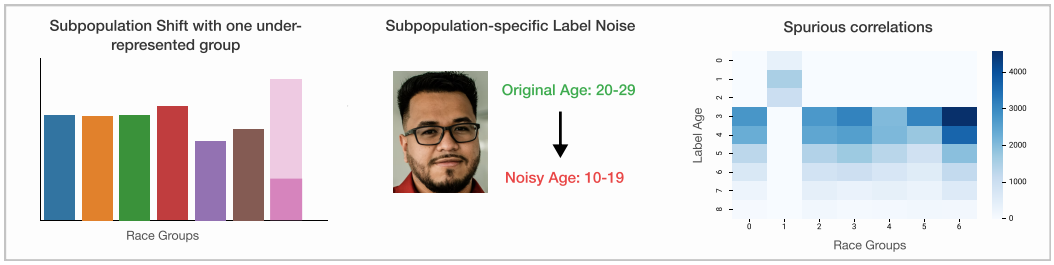
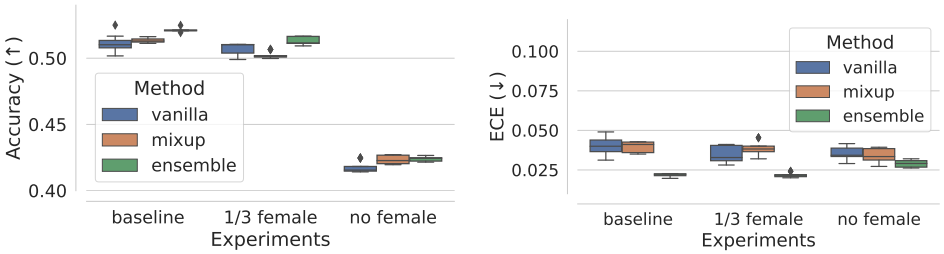


Figure 2: Overview of the introduced shifts, from left to right: Subpopulation shift with one under-represented race value, group specific label making the subjects jounger, and spurious correlations between *young* and one race group.

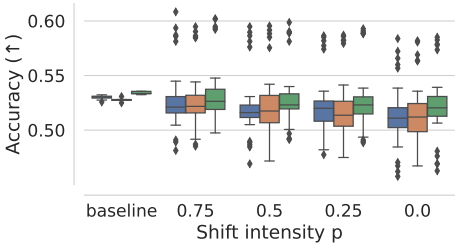
B Additional Results

B.1 Subpopulation Shift

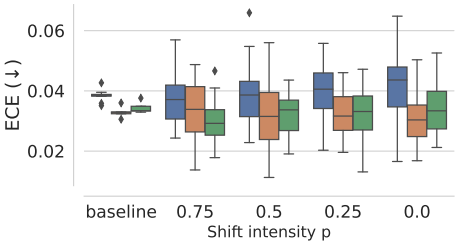


(a) Gender - Accuracy

(b) Gender - ECE



(c) Race - Accuracy



(d) Race - ECE

Figure 3: ECE and accuracy of subpopulation shift of race.

B.2 Label Noise

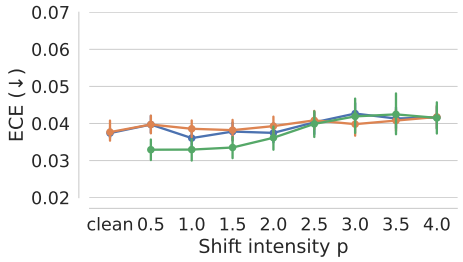
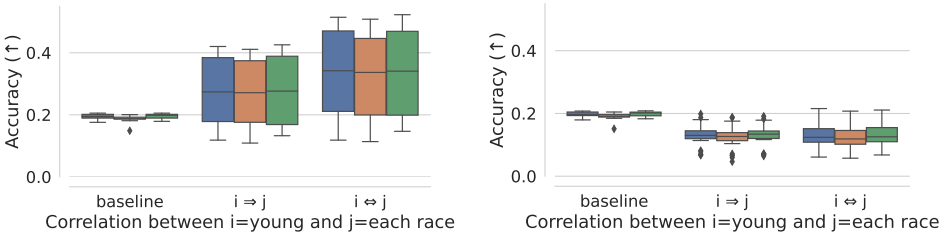


Figure 4: ECE of label noise with temperature scaling.

B.3 Spurious Correlations



(a) *Young-correlated*- Accuracy (b) *Non-young-correlated* - Accuracy

Figure 5: ECE accuracy of spurious correlations of race and young.