Fostering Fairness in Image Classification through Awareness of Sensitive Data

Ivona Colakovic
ivona.colakovic@um.si
Faculty of Electrical Engineering and Computer Science,
University of Maribor
Maribor, Slovenia

Sašo Karakatič
Faculty of Electrical Engineering and Computer Science,
University of Maribor
Maribor, Slovenia
saso.karakatic@um.si

ABSTRACT

The wide use of machine learning (ML) attests to its power to unveil patterns hidden in data. However, biases and discrimination against certain groups of individuals that can appear in data or, consequently, in ML outcomes have been of great concern recently. In the world of increasing production of unstructured data, such as images, the questions about fairness are of great relevance, as the patterns in images may nudge the neural network (NN) models to make unfair decisions (based on the learned patterns that stem from societal and historical biases). To tackle this issue, we propose a new algorithmic improvement in learning of NNs that balances the treatment quality among the sensitive groups, resulting in fairer treatment when classifying images while maintaining good classification performance. The proposed approach is evaluated on the FairFace dataset, and the results indicate that the fair approach improves fairness while maintaining comparable overall quality. Furthermore, it closes the gap between the model's quality of different sensitive feature groups.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Supervised learning by classification**.

KEYWORDS

fairness, machine learning, image classification, bias

1 INTRODUCTION

Machine learning (ML) significantly impacts decisions in different fields, like hiring and credit eligibility. Due to past biases, regulations like the EU's AI Act aim to ensure fairer and more trustworthy ML models. Certain human traits like gender, age, race and so on, should not be deal-breakers in decision-making. Those human traits are called sensitive features.

Sensitive features should not influence ML decisions, yet issues persist, like facial recognition technologies favoring white individuals [2] or Google Photos misclassifying black people as gorillas [1]. Creating fair ML models is challenging due to the lack of a standard fairness definition, leading to over 20 distinct metrics [9]. Furthermore, achieving perfect fairness can reduce overall model quality and exacerbate existing biases.

We focus on group fairness, aiming for equal quality of service among sensitive feature groups, ensuring no group is disadvantaged. Historical data often contains biases that impact ML models, which can persist even after removing sensitive features. Unfairness can arise from various biases like from underrepresentation or difficulty in extracting patterns [7], necessitating corrections during the training phase.

Our contribution is a fair approach to image classification that enhances fairness while maintaining overall performance.

2 RELATED WORK

Algorithmic fairness has been a subject of study since the mid-1990s, with substantial research commencing around 2015. Approaches to mitigate bias can be categorized based on the stage of fairness consideration: pre-processing, in-processing, and post-processing [7]. While pre-processing methods address data imbalances and post-processing address post hoc unfairness while observing the model as a black box, in-processing methods directly incorporate fairness into the learning phase. This study focuses on in-processing methods, specifically on neural networks used for image classification.

Existing work has primarily addressed tabular data, with limited exploration of fairness in image classification using CNNs [4, 5, 8, 10]. Using in-processing methods for image classification, we believe it is essential to address the backpropagation process, which iteratively adjusts the weights of a neural network based on the loss function, especially with the improved fairness in ensemble methods, where also sample weights are iteratively adapted [3].

3 METHODOLOGY

Convolutional Neural Networks (CNNs) are widely used for image classification due to their effective feature extraction capabilities. The learning process involves a forward pass where the input data is processed through the network to produce an output, and a backward pass where the error is propagated back to adjust the network weights using gradient descent. The common loss function for classification tasks is cross-entropy (CE), which measures the dissimilarity between predicted and actual distributions.

To address fairness, this study modifies the cross-entropy loss function to account for disparities in treatment quality among sensitive feature groups. Instead of computing a single loss value for all data points, the proposed method calculates the loss separately for each sensitive feature group. The final loss used for backpropagation is the maximum loss among all groups, ensuring that the

KDD '24X, August 25–29, 2024, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

worst-performing group's error drives the model updates. This approach aims to equalize the model's performance across different sensitive feature groups.

Convolutional neural networks (CNNs) are employed due to their proficiency in handling image classification tasks. CNNs are composed of multiple layers that apply various filters to input images, creating a hierarchical representation of features that the network can use for classification decisions. The training process of a CNN involves two main phases: the forward pass and the backward pass.

During the forward pass, input images are processed through the network, where each layer applies convolutional filters, pooling operations, and non-linear activations. This results in a set of output probabilities, one for each class. These output probabilities are then compared to the ground truth labels using a loss function to compute the error. Traditionally, the cross-entropy loss function is used, which measures the dissimilarity between the predicted probability distribution and the true distribution. The formula for cross-entropy loss is:

$$CE(P_x, P_y) = -\sum_i P_x(i) log(P_y(i)) \tag{1}$$

The backward pass follows, where the goal is to minimize the loss by adjusting the network weights. This is achieved using gradient descent, where the gradient of the loss with respect to each weight is calculated and the weights are updated accordingly. This process involves propagating the gradient backward through the network layers and making incremental updates to the weights to reduce the error.

To incorporate fairness into the training process, this methodology introduces a novel adaptation of the cross-entropy loss function. Instead of calculating a single loss value for the entire dataset, the loss is computed separately for each sensitive feature group. For instance, if race is the sensitive feature, separate losses are calculated for each racial group.

The key innovation lies in defining the final loss as the maximum of the losses calculated for each sensitive feature group. This approach ensures that the optimization process focuses on improving the performance of the worst-performing group, thereby promoting fairness. The modified loss function is given by:

$$loss = \max_{s \in S} CE(P_{x,s}, P_{y,s}) \tag{2}$$

4 EXPERIMENTS

The proposed approach was evaluated using the FairFace dataset [6], which is balanced across various demographic groups. The target variable for classification was gender, with race as the sensitive feature. A pre-trained ResNet-18 model was fine-tuned on the FairFace dataset using both the conventional cross-entropy loss and the proposed fairness-aware approach. The model's performance was assessed for overall classification performance using metrics such as accuracy, f-score, and from fairness perspective using metrics like accuracy difference (acc-diff), f-score difference (f1-diff).

4.1 Results

The results shown in 1 indicate that the fair approach achieved comparable overall classification quality to the conventional approach but notably improved fairness. Specifically, the fair approach demonstrated a substantial reduction in performance disparities among sensitive feature groups.

Despite a slight drop in overall accuracy and f-score, the improvement in fairness metrics underscored the trade-off between classification quality and fairness. This trade-off is deemed acceptable, given the societal importance of fair treatment in ML models.

Table 1: Results achieved by conventional and fair approach with the results of statistical test

Metric type	Metric	Conventional approach	Fair approach
Standard	Accuracy ↑	**0.745**	0.730
	F-score ↑	**0.765**	0.749
Fair	Acc-diff ↓	0.207	**0.157**
	F1-diff ↓	0.253	**0.198**

5 CONCLUSION

The study presents a novel fairness-aware approach to image classification that modifies the cross-entropy loss function to prioritize the performance of the worst-performing sensitive feature group. The experimental evaluation on the FairFace dataset demonstrates that the proposed method improves fairness while maintaining comparable classification quality. To additionally validate the proposed method, an evaluation of different fairness perspectives could be included, alongside a statistical comparison of the results should be done.

Although the fair approach demonstrated an improvement in fairness, it relies on instances being labelled with sensitive features. Future research could focus on exploring methods to acquire or infer sensitive features when they are not readily available. With various fairness aspects and a significant number of network parameters, this work allows the incorporation of different fairness metrics and leaves space for additional fine-tuning. To better understand the impact of the proposed approach, it could be evaluated on different datasets and compared to *state-of-the-art* methods in future. Finally, the power of fair approach evaluated on pretrained network shows the potential for use in knowledge distillation.

ACKNOWLEDGMENTS

The authors acknowledge the financial support from the Slovenian Research and Innovation Agency (Research Core Funding No. P2-0057).

REFERENCES

[1] [n. d.]. Google Photos labeled black people 'gorillas'. https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/

[2] Joy Buolamwini. 2018. Opinion | When the Robot Doesn't See Dark Skin. *The New York Times* (June 2018). https://www.nytimes.com/2018/06/21/opinion/facial-analysis-technology-bias.html

[3] Ivona Colakovic and Sašo Karakatič. 2023. FairBoost: Boosting supervised learning for learning on multiple sensitive features. *Knowledge-Based Systems* 280 (Nov. 2023), 110999. https://doi.org/10.1016/j.knosys.2023.110999

[4] Tongxin Hu, Vasileios Iosifidis, Wentong Liao, Hang Zhang, Michael Ying Yang, Eirini Ntoutsi, and Bodo Rosenhahn. 2020. FairNN - Conjoint Learning of Fair Representations for Fair Decisions. In *Discovery Science*, Annalisa Appice, Grigorios Tsoumakas, Yannis Manolopoulos, and Stan Matwin (Eds.). Springer International Publishing, Cham, 581–595.

[5] Jian Kang, Tiankai Xie, Xintao Wu, Ross Maciejewski, and Hanghang Tong. 2022. InfoFair: Information-Theoretic Intersectional Fairness. In *2022 IEEE International Conference on Big Data (Big Data)*. 1455–1464. https://doi.org/10.1109/BigData55660.2022.10020588

[6] Kimmo Karkkainen and Jungseock Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Waikoloa, HI, USA, 1547–1557. https://doi.org/10.1109/WACV48630.2021.00159

[7] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (July 2021), 1–35. https://doi.org/10.1145/3457607

[8] Andrija Petrović, Mladen Nikolić, Sandro Radovanović, Boris Delibašić, and Miloš Jovanović. 2022. FAIR: Fair adversarial instance re-weighting. *Neurocomputing* 476 (March 2022), 14–37. https://doi.org/10.1016/j.neucom.2021.12.082

[9] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*. ACM, Gothenburg Sweden, 1–7. https://doi.org/10.1145/3194770.3194776

[10] Yanfu Zhang, Shangqian Gao, and Heng Huang. 2022. Recover Fair Deep Classification Models via Altering Pre-trained Structure. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 481–498.