GroupMixNorm Layer for Learning Fair Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent research has focused on proposing algorithms for bias mitigation from automated prediction algorithms. Most of the techniques include convex surrogates of fairness metrics such as demographic parity or equalized odds in the loss function, which are not easy to estimate. Further, these fairness constraints are mostly data-dependent and aim to minimize the disparity among the protected groups during the training. However, they may not achieve similar performance on the test set. In order to address the above limitations, this research proposes a novel *GroupMixNorm* layer for bias mitigation from deep learning models. As an alternative to solving constraint optimization separately for each fairness metric, we have formulated bias mitigation as a problem of distribution alignment of several groups identified through the protected attributes. To this effect, the GroupMixNorm layer probabilistically mixes group-level feature statistics of samples across different groups based on the protected attribute. The proposed method improves upon several fairness metrics with minimal impact on accuracy. Experimental evaluation and extensive analysis on benchmark tabular and image datasets demonstrate the efficacy of the proposed method to achieve state-of-the-art performance.

1 Introduction

Adoption of AI systems is increasing at a rapid rate to deliver seamless consumer experience across various domains such as travel time prediction, health monitoring, authentication etc Mehrabi et al. (2021); Du et al. (2021b); Bellamy et al. (2018). Such pipelines are mostly automated in nature without any human intervention, owing to the large data processing, high efficiency, and high accuracy. Despite the benefits of automated processing, current AI systems are marred with the challenge of biased predictions resulting in unfavourable outcomes. In order to rectify such biases and advance in the society, we need models that generate fair results without any discrimination or favouritism towards certain individual or groups of society. To this effect, this research proposes a novel *GroupMixNorm* layer for learning an unbiased model from the given data ensuring fair outcomes across different groups.

Existing in-processing techniques mostly aim to solve a constraint optimization problem to ensure fairness Woodworth et al. (2017); Zafar et al. (2017); Kamishima et al. (2011); Wu et al. (2018); Wilson et al. (2019); Zhang et al. (2018) by introducing a penalty term in the loss function corresponding to the convex surrogates of the fairness objective like *demographic parity* or *equalized odds*. However, as has been observed in literature, it is challenging to formulate surrogates for different fairness constraints that is a reasonable estimate of the original constraint Manisha and Gujar (2020). The proposed *GroupMixNorm* layer is applied at the in-processing stage which promotes the model to learn bias invariant features for classification by modeling the data distribution during training.

In this research, we formulate the problem of bias mitigation as distribution alignment of several groups identified through the protected attributes (Figure 1). Our formulation is motivated by the observation that Deep Learning based algorithms tend to explore the difference in the distribution

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

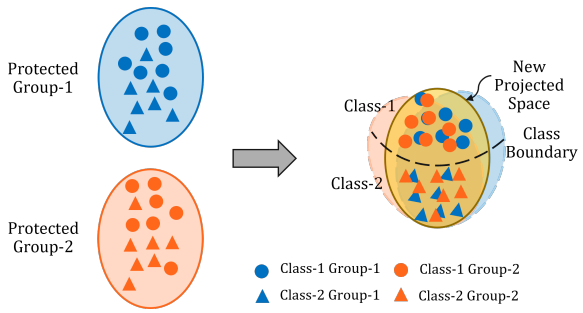


Figure 1: The proposed GroupMixNorm layer projects the representations of different classes and protected attributes onto a space which minimizes the distinction between the protected attributes. Lower protected attributed-based distinction enables the generation of a fairer classification model.

among the groups of the protected attributes (for example, male and female with similar features like age and education may have different salaries, thus resulting in different distributions) to lift the overall performance and in the process may result in bias being exhibited towards certain groups. The proposed method, termed GroupMixNorm, mixes the group-level feature statistics and further transforms all the features in a training batch based on the interpolated group statistics. This enables the classifier to learn invariant features pertaining to protected attributes during training. Key highlights of this research are as follows:

- This research proposes a novel *GroupMixNorm* layer for learning fairer classification models. The proposed layer is applied at the architectural level, and is an in-processing technique which focuses on distribution alignment of different groups during model training.
- GroupMixNorm operates at the feature level, thus making it flexible to be placed across various layers of a neural network and fits well into the mini-batch gradient-based training.
- The efficacy of the proposed approach has been demonstrated on different datasets (structured and unstructured), where it achieves improved performance while enforcing multiple fairness constraints such as demographic parity, equal opportunity and equalized odds.

2 Proposed GroupMixNorm Layer

The proposed *GroupMixNorm* layer focuses on achieving group fairness across different sensitive groups within a protected attribute by eliminating the difference between the group features/statistics during training. As part of the GroupMixNorm layer, we normalize each group of a protected attribute in a batch separately to collect group specific statistics (i.e. for the gender attribute, normalize all male samples and female samples in a batch separately) and further take a probabilistic convex combination between the group-level statistics and apply across all the samples in a batch. This process ensures that any protected group related diversity is removed from the internal representation of a neural network and doesn't allow the network to explore this information to lift the overall performance. The introduction of additional inductive bias in the network structure enforces it to learn invariant features pertaining to protected attributes while training the network.

The GroupMixNorm layer is implemented as a plug-and-play module. It can be inserted between fully connected layers of a neural network-based classifier during training. Let X, Y and S be the input features, class labels and protected attribute labels in a training batch respectively. As illustrated in Figure 2 consider Z be an n dimensional hidden representation obtained from the previous layer and A_i represent feature along dimension i. We identify the groups G_j in a batch based on the protected attribute labels S, and calculate their respective mean ($\mu_{G_j,i}$) and variance ($\sigma_{G_j,i}$) along each dimension, as shown in the step-1 of Figure 2. Next we calculate weighted average of mean $\gamma_{mix,i}$ and variance $\beta_{mix,i}$ along each dimension (Equation 1). We mix group level statistics by calculating a weighted average of mean γ_{mix_i} and variance β_{mix_i} along each dimension (Equation 1). As we mix statistics of two groups, the mixing coefficient λ is sampled from a symmetric Beta distribution $Beta(\alpha, \alpha)$, for $\alpha \in (0, \infty)$. The hyper-parameter α controls the interpolation strength.

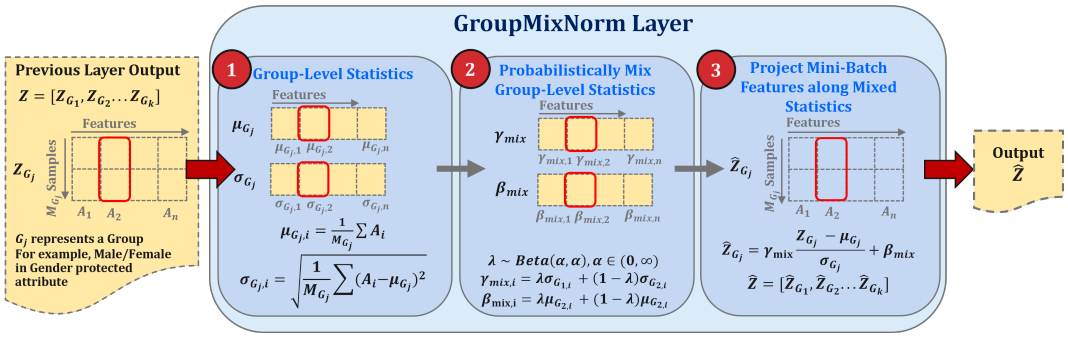
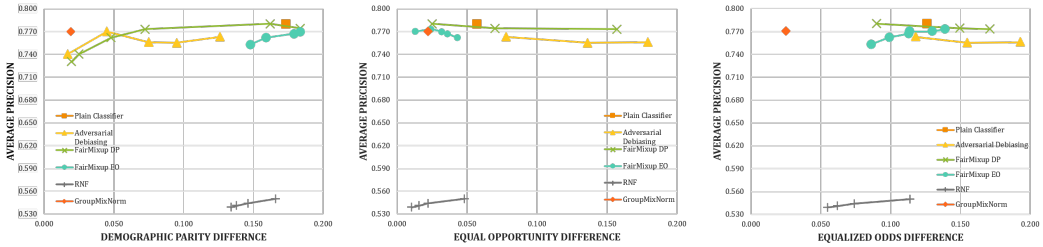


Figure 2: Diagrammatic representation of the workings of the GroupMixNorm layer. The previous layer output is provided as input to the GroupMixNorm layer along with the protected attribute information for each sample. The GroupMixNorm layer computes the group-level statistics followed by the probabilistic mixing and projection of the mini-batch features along the mixed statistics.



(a) Demographic Parity (b) Equal Opportunity (c) Equalized Odds

Figure 3: The fairness-AP trade-off curve comparison of the proposed GroupMixNorm with other techniques on the Adult dataset. Results are obtained by varying the trade-off parameter as suggested in their respective publications: (i) Adversarial Debiasing: [0.01 $\sim$ 1.0], (ii) Fair Mixup DP: [0.1 $\sim$ 0.7], (iii) Fair Mixup EO: [0.5 $\sim$ 5.0], and (iv) RNF: [0.05, 0.015, 0.025, 0.035].

Finally, we normalize all the samples by applying the calculated γ_{mix} and β_{mix} to each sample as shown in Equation 3. For the ease of notation, we have considered two groups i.e. binary protected attributes. Nevertheless, the proposed solution can be easily applied to non-binary protected attributes. The three step approach is shown in Figure 2.

$$\gamma_{mix,i} = \lambda \sigma_{G_1,i} + (1-\lambda)\sigma_{G_2,i}; \beta_{mix,i} = \lambda \mu_{G_1,i} + (1-\lambda)\mu_{G_2,i} \tag{1}$$

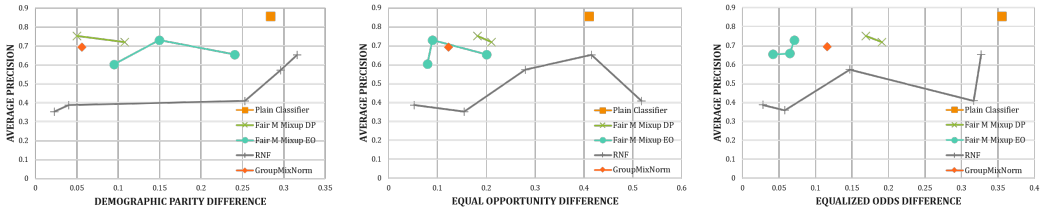
$$\gamma_{mix} = [\gamma_{mix,1}, \ldots \gamma_{mix,n}]; \beta_{mix} = [\beta_{mix,1}, \ldots \beta_{mix,n}] \tag{2}$$

$$\hat{Z}_{G_j} = \gamma_{mix} \frac{(Z_{G_j} - \mu_{G_j})}{\sigma_{G_j}} + \beta_{mix} \tag{3}$$

The updated features $\hat{Z} = [\hat{Z}_{G_1}, \hat{Z}_{G_2}]$ are then provided as input to the following layer of the neural network for further processing. The process of mixing group level statistics in a GroupMixNorm layer occurs in the feature space and has no learnable parameters. The GroupMixNorm layer is easy to implement and fits perfectly into mini-batch training. We train the entire neural network end-to-end for the classification task on Binary cross-entropy loss. However, GroupMixNorm layer is turned off during inference, hence we don't need protected attribute information during inference time.

3 Results

Since the proposed technique focuses on mitigating bias during the training process, comparisons have been performed with algorithms that optimize fairness constraints during training. In particular, we have compared with a (i) Plain Classifier, (ii) Adversarial Debiasing Zhang et al. (2018), (iii) Fair Mixup: Fairness via Interpolation (Fair Mixup) Chuang and Mroueh (2021), and (iv) Fairness via Representation Neutralization (RNF) Du et al. (2021a). The Fair Mixup technique uses two separate regularizing terms for optimizing the fairness metrics of Demographic Parity(DP) and Equal



(a) Demographic Parity (b) Equal Opportunity (c) Equalized Odds

Figure 4: The fairness-AP trade-off comparison of GroupMixNorm with other techniques on the CelebA dataset. The trade-off parameters are varied as suggested in their respective publications: (i) Fair M Mixup DP: [25, 50], (ii) Fair M Mixup EO: [1, 10, 50], (iii) RNF: [0.1, 0.5, 1, 5, 10].

Opportunity (EO), and thus can solve for either DP or EO at a time. In this paper, we refer to these two variants of Fair Mixup as Fair Mixup DP and Fair Mixup EO.

For a fair comparison, we evaluate all the models under the same setting. We have reported the classification performance on Average Precision (AP) and fairness on standard metrics like Demographic Parity Difference (DPD), Equalized Odds Difference (EOD), and Equal Opportunity Difference (EOD). For the Adult dataset, we follow the defined pre-processing Chuang and Mroueh (2021) for all the experiments. We report the average results of 10 independent runs with different seeds. Further, techniques such as Adversarial Debiasing, Fair Mixup and RNF introduce a regularization term in the loss function to improve fairness, via a hyper-parameter α that controls the trade-off between the average precision (AP) and fairness metrics (DPD, EOD, and EOP). We have reported the results on varying values of α as suggested in their respective papers. In our case, the GroupMixNorm layer is proposed towards architecture design and not the loss function, thus there is no such trade-off. Performance analysis on different datasets is as follows:

Comparison on the UCI Adult Income Dataset: Figure 3 shows the performance comparison on the UCI Adult dataset, where we observe that the proposed GroupMixNorm method produces fairer results as compared to other techniques across all the three fairness metrics (DPD, EOP, EOD) with minimal impact on average precision (AP). Fair Mixup solves separate constraint optimizations to achieve lower Demographic Parity and Equalized Odds, thus minimizing either DP or EO at a time. This is evident from Figure 3(a), where Fair Mixup EO doesn't work well for DP and the DPD values are greater than 0.1 throughout. In terms of fairness metrics, RNF produces fair results however the average precision is relatively much lower, thus making it unsuitable for the classification task.

Comparison on the CelebA Dataset: Figure 4 presents the experimental results on the CelebA dataset. For Fair Mixup, comparison has been performed with the combination of manifold mixup Verma et al. (2019) which has shown to achieve improved results in the published manuscript Chuang and Mroueh (2021) (Fair M Mixup DP and Fair M Mixup EO). Similar to the previous experiments, it is observed that either Fair M Mixup DP or Fair M Mixup EO achieves optimal performance at a time for a fairness metric. Further, the RNF model produces fair results across all the fairness metrics, however achieves low average precision (classification performance). The proposed GroupMixNorm achieves comparable performance to the best performing model across all the metrics, while obtaining high average precision as well, thus suggesting high applicability in real-world setups.

4 Conclusion

Learning bias-invariant models are the need of the hour for the research community. While existing research has focused on proposing novel solutions for learning unbiased classifiers, most of the techniques incorporate an additional term in the loss function for modeling the model fairness. We believe that it is often difficult to extrapolate the learnings of such an optimization function to the test set. To this effect, this research proposes a novel *GroupMixNorm* layer, which promotes learning fairer models at the architectural level. GroupMixNorm is a distribution alignment strategy operating across the different groups identified in a protected attribute, enabling attribute-invariant feature learning. Experimental evaluation has been performed on two datasets containing tabular and image data, respectively for the classification task. Across multiple experiments, GroupMixNorm demonstrates improved fairness metrics while maintaining the highest average precision level.

References

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *CoRR* abs/1810.01943 (2018).

Ching-Yao Chuang and Youssef Mroueh. 2021. Fair Mixup: Fairness via Interpolation. In *International Conference on Learning Representations*. Virtual Event, Austria, May 3-7, 2021.

Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. 2021a. Fairness via Representation Neutralization. *Advances in Neural Information Processing Systems* 34 (2021).

Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2021b. Fairness in Deep Learning: A Computational Perspective. *IEEE Intell. Syst.* 36, 4 (2021), 25–34.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware Learning through Regularization Approach. In *IEEE International Conference on Data Mining Workshops*, Myra Spiliopoulou, Haixun Wang, Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu (Eds.). 643–650.

Padala Manisha and Sujit Gujar. 2020. FNNC: Achieving Fairness through Neural Networks. In *International Joint Conference on Artificial Intelligence*, Christian Bessiere (Ed.). 2277–2283.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6 (2021), 115:1–115:35.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold Mixup: Better Representations by Interpolating Hidden States. In *International Conference on Machine Learning*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. 6438–6447.

Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive Inequity in Object Detection. *CoRR* abs/1902.11097 (2019).

Blake E. Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. 2017. Learning Non-Discriminatory Predictors. In *Conference on Learning Theory*, Satyen Kale and Ohad Shamir (Eds.), Vol. 65. 1920–1953.

Yongkai Wu, Lu Zhang, and Xintao Wu. 2018. Fairness-aware Classification: Criterion, Convexity, and Bounds. *CoRR* abs/1809.04737 (2018).

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *International Conference on Artificial Intelligence and Statistics*, Aarti Singh and Xiaojin (Jerry) Zhu (Eds.), Vol. 54. 962–970.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *AAAI/ACM Conference on AI, Ethics, and Society*, Jason Furman, Gary E. Marchant, Huw Price, and Francesca Rossi (Eds.). 335–340.