Uncovering Bias: Exploring Gender Dynamics in Distance-Aware Mixup Techniques

Anonymous authors
Paper under double-blind review

Abstract

Bias is a pervasive issue in machine learning and has implications in multiple AI applications, encompassing dimensions like gender, age, demographics, and social aspects. Complex models, including deep neural networks, transformers etc., often inherit biases and stereotypes during training, attributable to selection bias within training data and algorithmic creation processes. Augmentation techniques like Mixup exhibit promising potential as debiasing frameworks, leveraging specialized sampling strategies and spatial information for bias mitigation. In this study, we evaluate gender bias within the distance-aware mixing frameworks, while exploring diverse sampling strategies for mixup. Using the Trustpilot corpus, we conduct experiments[1] quantitatively analyzing bias as error disparity, investigating the impact of distance thresholds and various gender-based criteria on mixup operations. Our quantitative analysis indicates that employing a cross-gender mixup strategy yields the most effective bias reduction. We also release the code for our work.

1 Introduction and Related Work

Deep neural networks have shown to be effective in learning intricate data patterns enabling complex decision-making in several applications ranging from natural language and speech processing, computer vision, reinforcement learning etc. These models, trained under the paradigm of Empirical Risk Minimization (ERM) principle, tend to memorize the nuances of training data, resulting in poor generalization ability (Zhang et al., 2017). They are prone to learning simple patterns and spurious correlations in the data, thus inducing bias. Bias in machine learning can manifest in various forms including selection bias, label bias, algorithmic bias and be based on different dimensions like gender, age, demographics etc. The prevalence of spurious correlations within datasets exacerbates this issue, primarily attributable to biases present in the data. Mixup (Zhang et al., 2017) was proposed to augment ERM by synthesizing diverse training examples, thus enhancing model generalization and mapping complex data patterns. Selective mixup, employing bias-label aware sampling strategy (Hwang et al., 2022), also shows promise as a debiasing framework. Incorporating (dis/)similarity information between samples while sampling minimizes reinforcement of biased associations, promotes diversity, and fosters fairer representations across groups. In this work, we conduct a bias analysis for a learnable distance-aware mixup methodology to gauge the impact of gender-based mixup sampling strategy on the overall model bias. Based on quantitative experiments, we show the efficiency of cross-gender, distance-aware mixup strategy and the inverse relationship between the distance threshold and the model bias.

Recent works have investigated multiple dimensions of biases in machine learning including gender (Sawhney et al., 2021; Costa-jussà et al., 2022), age (Chu et al., 2023), demographics and political inclinations (Rozado, 2023), social aspects (Olteanu et al., 2019) etc. These often stem from inherent selection biases within training data (Tommasi et al., 2015) and the model bias due to factors like biased algorithmic creation (Forum, 2018). Bias also stems out of underrepresentation of specific groups, such as females in this field (Leavy, 2018). Language models such as BERT (Bidirectional Encoder Representations from Transformers), GPT2, and contextual embeddings like word2vec also tend to inherit gender stereotypes and exhibit biases (Nadeem et al., 2020; Bolukbasi et al., 2016).

[1] Code available at https://anonymous.4open.science/r/dmix-bias-0BC9/

Mixup (Zhang et al., 2017) is a popular augmentation technique that interpolates two examples along with their corresponding labels. Applying mixup on latent input representations results in additional improvements (Chen et al., 2020a). Utilizing data-level spatial similarity information for sample selection also leads to performance enhancements (Chen et al., 2020b).Chhabra et al. (2023) demonstrate effectiveness and improved generalization of mixup in different geometric spaces. Sawhney et al. (2022) propose a learnable distance-aware mixup methodology based on similarity between latent representations in hyperbolic space.

2 Methodology and Experiments

DMix (Sawhney et al., 2022) defines distance aware sample-level mixup i.e. for each element x_i

$$\text{DMIX}(x_i) = \text{DMixup}(x_i, x_j), x_j \sim S_i \tag{1}$$

where DMixup is interpolative mixup with mixing ratio $\mathbf{M}_{ij}$ - a learnable parameter initialized with the distance between x_i and x_j. S_i is the set of all samples such that distance between x_i and x_j is greater than a threshold $\tau = T \cdot \max(M_i)$, where T is a hyperparameter $\in (0, 1)$. The sample x_j for mixup is sampled randomly from S_i. To analyse model-induced gender bias, we constrain the construction of the set S_i. We induce two types of constraints - for any sample x_i, one of the following holds:

$$S_i = \{x_k \mid \mathbf{M}_{ik} \geq \tau \wedge g_i = g_k\} \text{ OR } S_i = \{x_k \mid \mathbf{M}_{ik} \geq \tau \wedge g_i \neq g_k\}$$

Here, g_i and g_k are genders of x_i and x_k respectively. The first scenario is same-gender mixup and the second is cross-gender mixup. We also vary T over a range of values to study its effect on bias.

We follow Saunders & Byrne (2020); Sawhney et al. (2021) to define bias as the performance error disparity $\Delta G = BCE_f - BCE_m$, where BCE stands for binary cross-entropy loss, and f and m stand for female and male respectively. We use the Trustpilot corpus (Hovy et al., 2015), which contains text-based user reviews from the Trustpilot website, rating companies and services on a 1 to 5 star scale. In particular, we use all English reviews from the US. We perform sentiment analysis task on this dataset by binarizing the ratings as 0 for ratings less than 3, 1 for ratings greater than 3, and ignore the rest. We choose this dataset due to the prevalent gender bias in sentiment analysis (Thelwall, 2018; Young et al., 2009). The dataset also has a balanced gender distribution allowing us to capture model-induced instead of dataset-induced bias.

3 Results and Analysis

Model	$\Delta G = BCE_f - BCE_m$			
	$T = 0.1$	$T = 0.3$	$T = 0.5$	$T = 0.7$
DMix (no constraint)	0.48	0.48	0.25	0.16
DMix (same gender)	0.40	0.39	0.26	0.07
DMix (different gender)	0.27	0.18	0.26	0.01

Table 1: Bias Analysis: Performance error disparity for different constraints applied to DMix

On increasing the value of T, which controls the diversity of the samples selected for mixup, samples that are more far apart (dissimilar) are mixed which leads to better generalizability for female samples, thus reducing model bias. Constraining mixup to only same-gender samples reduces intra-gender. It fails to mitigate inter-gender bias since same-gender mixup fails to capture inter-gender relations between the two distributions. Different gender mixup can capture these relations which leads to the least bias.

4 Conclusion

In this work, we study the gender bias inherent in distance-aware mixup techniques. Building on existing works, we consider two types of sampling strategies during mixup and quantitatively evaluate their effect on model bias. Experimental results on Trustpilot dataset show that cross-gender mixup strategy and increasing the thresholding distance achieves best bias mitigation.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

Ethical Consideration

Although our study focuses on gender bias, we acknowledge that there could exist other dimensions that lead to different kinds of bias which can include age, race, demographics, socio-economic and cultural factors. Our study is limited to the gender labels available in the data (male and female) but we acknowledge the presence of other forms of gender as sensitive attributes and the biases inherent to those.

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. Local additivity based data augmentation for semi-supervised NER. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1241–1251, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.95. URL https://aclanthology.org/2020.emnlp-main.95.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020b.

Parth Chhabra, Atula Tejaswi Neerkaje, Shivam Agarwal, Ramit Sawhney, Megh Thakkar, Preslav Nakov, and Sudheer Chava. Learning through interpolative augmentation of dynamic curvature spaces. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 2108–2112, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3592008. URL https://doi.org/10.1145/3539618.3592008.

Charlene H Chu, Simon Donato-Woodger, Shehroz S Khan, Rune Nyrup, Kathleen Leslie, Alexandra Lyn, Tianyu Shi, Andria Bianchi, Samira Abbasgholizadeh Rahimi, and Amanda Grenier. Age-related bias and artificial intelligence: a scoping review. *Humanit. Soc. Sci. Commun.*, 10(1), August 2023.

Marta R. Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. Interpreting gender bias in neural machine translation: Multilingual architecture matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):11855–11863, Jun. 2022. doi: 10.1609/aaai.v36i11.21442. URL https://ojs.aaai.org/index.php/AAAI/article/view/21442.

World Economic Forum. https://www3.weforum.org/docs/WEF_GGGR_2018.pdf, 2018. [Accessed 09-12-2023].

Dirk Hovy, Anders Johannsen, and Anders Søgaard. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pp. 452–461, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. ISBN 9781450334693. doi: 10.1145/2736277.2741141. URL https://doi.org/10.1145/2736277.2741141.

Inwoo Hwang, Sangjun Lee, Yunhyeok Kwak, Seong Joon Oh, Damien Teney, Jin-Hwa Kim, and Byoung-Tak Zhang. Selecmix: Debiased learning by contradicting-pair sampling. *Advances in Neural Information Processing Systems*, 35:14345–14357, 2022.

Susan Leavy. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, GE '18, pp. 14–16, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357388. doi: 10.1145/3195570.3195580. URL https://doi.org/10.1145/3195570.3195580.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models, 2020.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2:13, 2019.

David Rozado. Danger in the machine: The perils of political and demographic biases embedded in ai systems. *Manhattan Institute*, 2023.

Danielle Saunders and Bill Byrne. Reducing gender bias in neural machine translation as a domain adaptation problem. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7724–7736, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.690. URL https://aclanthology.org/2020.acl-main.690.

Ramit Sawhney, Arshiya Aggarwal, and Rajiv Ratn Shah. An empirical investigation of bias in the multimodal analysis of financial earnings calls. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3751–3757, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.294. URL https://aclanthology.org/2021.naacl-main.294.

Ramit Sawhney, Megh Thakkar, Shrey Pandit, Ritesh Soun, Di Jin, Diyi Yang, and Lucie Flek. DMix: Adaptive distance-aware interpolative mixup. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 606–612, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.67. URL https://aclanthology.org/2022.acl-short.67.

Mike Thelwall. Gender bias in sentiment analysis. *Online Information Review*, 42(1):45–57, 2018.

Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. *ArXiv*, abs/1505.01257, 2015. URL https://api.semanticscholar.org/CorpusID:5665048.

Suzanne Young, Leslie Rush, and Dale Shaw. Evaluating gender bias in ratings of university instructors' teaching effectiveness. *International Journal for the Scholarship of Teaching and Learning*, 3(2):n2, 2009.

Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2017. URL https://api.semanticscholar.org/CorpusID:3162051.

A Appendix

A.1 Overview: DMIX

DMIX (Sawhney et al., 2022) is a an adaptive distance aware interpolative data augmentation methodology that makes use of similarity information in the data distribution to optimally select mixup samples. Since language representations possess complex geometry and hierarchical structures which are not effectively captured in Euclidean space, DMIX makes use of hyperbolic space to compute latent similarities. Given two samples $x_i, x_j \in X$, where X is the set of data samples and $i, j \in [1, N]$ (N is number of samples), we firstly initialize a learnable matrix M_{NxN} where

$$M_{ij} = D_h(x_i, x_j) \tag{2}$$

where $D_h(.)$ refers to the hyperbolic distance. The matrix M is also row normalized for scaling purposes. Using the mixing ratio M_{ij}, DMIX is then defined as

$$\text{DM}\textsc{ix}(x_i) = \text{DMixup}(x_i, x_j) = (1 - M_{ij}) * x_i + M_{ij} * x_j \tag{3}$$

The sample x_j is randomly sampled from the set S_i which is the set of all samples such that distance between x_i and x_j is greater than a threshold $\tau = T \cdot \max(M_i)$, where T is a hyperparameter $\in (0, 1)$.