Model	Regularizer Intensity	Err (%↓)	Bias (↓)	
			Δ_{EO}	Δ_{DP}
Retrain+EO	0.1	$16.763_{\pm 0.223}$ (+0.598)	$0.214_{\pm 0.027}$ (-0.009)	$0.196_{\pm 0.027}$ (+0.004)
	0.5	$17.440_{\pm 0.676}$ (+1.275)	$0.150_{\pm 0.040}$ (-0.073)	$0.183_{\pm 0.011}$ (-0.009)
	0.9	$18.810_{\pm 0.588}$ (+2.645)	$0.147_{\pm 0.031}$ (-0.076)	$0.184_{\pm 0.032}$ (-0.008)
Retrain+DP	0.1	$16.752_{\pm 0.263}$ (+0.587)	$0.206_{\pm 0.035}$ (-0.017)	$0.174_{\pm 0.026}$ (-0.018)
	0.5	$17.659_{\pm 0.375}$ (+1.494)	$0.201_{\pm 0.024}$ (-0.022)	$0.168_{\pm 0.023}$ (-0.024)
	0.9	$18.472_{\pm 0.228}$ (+2.307)	$0.204_{\pm 0.023}$ (-0.019)	$0.143_{\pm 0.021}$ (-0.049)
OURS	—	$15.829_{\pm 0.340}$ (-0.336)	$0.197_{\pm 0.011}$ (-0.026)	$0.180_{\pm 0.022}$ (-0.012)

Table 3: Experiment results on the Adult dataset using in-processing fairness methods during fine-tuning w.r.t. the test errors (F1 score) and fairness violations (Δ_{DP} and Δ_{EO}). Parentheses show changes from the original pre-trained model to the one after fine-tuning, with negatives indicating improvements. Our method achieves lower Δ_{DP} than Retrain+EO and lower Δ_{EO} than Retrain+DP. It enhances accuracy and fairness simultaneously.

(46 versus 742, 93.8% reduction), suggesting improved efficiency without the need for adjusting objective functions for new tasks. Furthermore, our approach provides greater flexibility by eliminating the need to predefine fairness criteria.

Model	Err (%↓)	Bias (↓)	
		Δ_{EO}	Δ_{DP}
Dataset: Adult			
MLP-3	$16.678_{\pm 0.681}$	$0.185_{\pm 0.023}$	$0.162_{\pm 0.027}$
MLP-4	$16.237_{\pm 0.456}$	$0.173_{\pm 0.029}$	$0.170_{\pm 0.022}$
MLP-5	$16.166_{\pm 0.337}$	$0.179_{\pm 0.012}$	$0.166_{\pm 0.014}$
Dataset: LFW+a			
ResNet-50	$10.634_{\pm 0.241}$	$0.218_{\pm 0.081}$	$0.080_{\pm 0.032}$
DenseNet	$9.775_{\pm 0.131}$	$0.232_{\pm 0.054}$	$0.073_{\pm 0.009}$
Dataset: CelebA			
ResNet-50	$17.582_{\pm 0.226}$	$0.243_{\pm 0.061}$	$0.182_{\pm 0.005}$
DenseNet	$18.438_{\pm 0.174}$	$0.201_{\pm 0.028}$	$0.169_{\pm 0.009}$

Table 4: Ablation study: The test errors (%) measured using F1 score and fairness violations using different model architectures across different datasets.

4.5 Ablation Study

In this section, we conduct an ablation study of our method, which we divide into two parts for discussion. The first part pertains to the model architectures. For the Adult dataset, we experiment with adjusting the number of layers in different MLP structures. For the image datasets, we compare the performance of ResNet-50 [He *et al.*, 2016] and DenseNet-121 [Huang *et al.*, 2017]. The results are listed in Table 4. For the Adult dataset, increasing model complexity leads to marginally improved predictions, with little variation in fairness metrics (Δ_{EO} and Δ_{DP}). For image datasets (LFW+a and CelebA), enhancing model complexity generally improves both predictive performance and fairness, though these improvements are not significant. Overall, the impact of the model's structure on its performance is not obvious. In the second part, we adjust each demographic group's impact as detailed in Eq. (2), diverging from our initial approach which uniformly considers each group's contribution through averaging. Here, we vary the ratio: $\hat{I}_N = \alpha \hat{I}_{S=1} + (1-\alpha)\hat{I}_{S=2}$, $\alpha \in [\frac{1}{2}, 1)$ and the result is presented in Fig. 5. From the graph, we can see there is a slight decrease in prediction errors, while in the meantime, both Δ_{DP} and Δ_{EO} generally have an increasing pattern as the value of α increases.

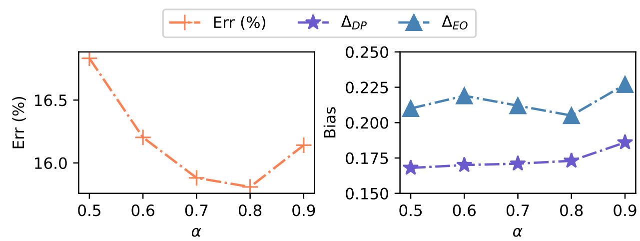

Figure 5: Ablation study: The impact of varying the contribution of different demographic groups in neutralized weight importance on prediction errors and fairness violations using the Adult dataset.

5 Conclusion

In this work, we addressed the crucial issue of fairness in fine-tuning. We tackle the limitations of constructing new models from scratch on new tasks, which is computationally intensive in many real-world scenarios. To this end, we proposed a novel weight importance neutralization strategy during fine-tuning to mitigate bias. Our approach involves assessing the weight importance using Fisher information and then incorporating this into SVD for low-rank approximation. By doing this, our method not only mitigates bias effectively but also enhances the efficiency of fine-tuning large pre-trained models. Our empirical analysis shows the effectiveness of our proposed method and demonstrates that even with a fair pre-trained model, it can still exhibit biases when fine-tuning on new tasks. Future research could further refine this technique, and investigate the applicability of this method across diverse domains with even larger models.

References

[Acharya *et al.*, 2018] Anish Acharya, Rahul Goel, Angeliki Metallinou, and Inderjit Dhillon. Online embedding compression for text classification using low rank matrix factorization, 2018.

[Agarwal *et al.*, 2018] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. *CoRR*, abs/1803.02453, 2018.

[Aghajanyan *et al.*, 2021] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online, August 2021. Association for Computational Linguistics.

[Ben Noach and Goldberg, 2020] Matan Ben Noach and Yoav Goldberg. Compressing pre-trained language models by matrix decomposition. In Kam-Fai Wong, Kevin Knight, and Hua Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 884–889, Suzhou, China, December 2020. Association for Computational Linguistics.

[Bilal Zafar *et al.*, 2015] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *arXiv e-prints*, page arXiv:1507.05259, Jul 2015.

[Bilal Zafar *et al.*, 2016] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *arXiv e-prints*, page arXiv:1610.08452, Oct 2016.

[Bird *et al.*, 2016] Sarah Bird, Solon Barocas, Kate Crawford, and Hanna Wallach. Exploring or exploiting? social and ethical implications of autonomous experimentation in ai. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT-ML), New York University*, page 4, October 2016.

[Buolamwini and Gebru, 2018] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

[Calmon *et al.*, 2017] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3992–4001. Curran Associates, Inc., 2017.

[Creager *et al.*, 2019] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi, and Richard S. Zemel. Flexibly fair representation learning by disentanglement. *CoRR*, abs/1906.02589, 2019.

[Dressel and Farid, 2018] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018.

[Dwork *et al.*, 2011] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness Through Awareness. *arXiv e-prints*, page arXiv:1104.3913, Apr 2011.

[Fu *et al.*, 2023] Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

[Golub and Reinsch, 1971] Gene H. Golub and Christian Reinsch. Singular value decomposition and least squares solutions. *Linear Algebra*, pages 134–151, 1971.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[He *et al.*, 2022] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2022.

[Hsu *et al.*, 2022] Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. In *International Conference on Learning Representations*, 2022.

[Hu *et al.*, 2022] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269. IEEE Computer Society, 2017.

[Iosifidis and Ntoutsi, 2019] Vasileios Iosifidis and Eirini Ntoutsi. Adafair: Cumulative fairness adaptive boosting. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 781–790, New York, NY, USA, 2019. Association for Computing Machinery.

[Jaderberg *et al.*, 2014] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Speeding up convolutional neural networks with low rank expansions, 2014.

[Kamishima *et al.*, 2012] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 35–50, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[Khandani *et al.*, 2010] Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767 – 2787, 2010.

[Kim *et al.*, 2015] Sung-Eun Kim, Hee Young Paik, Hyuk Yoon, Jung Lee, Nayoung Kim, and Mi-Kyung Sung. Sex- and gender-specific disparities in colorectal cancer risk. *World journal of gastroenterology : WJG*, 21:5167–5175, 05 2015.

[Kohavi, 1996] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 202–207. AAAI Press, 1996.

[Lin *et al.*, 2020] Zhaojiang Lin, Andrea Madotto, and Pascale Fung. Exploring versatile generative language model via parameter-efficient transfer learning. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 441–459, Online, November 2020. Association for Computational Linguistics.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. cite arxiv:1907.11692.

[Oneto *et al.*, 2020] Luca Oneto, Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer, and Massimiliano Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15360–15370. Curran Associates, Inc., 2020.

[Pascanu and Bengio, 2014] Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[Roh *et al.*, 2020] Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. FR-train: A mutual information-based approach to fair and robust training. 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 8147–8157. PMLR, 13–18 Jul 2020.

[Tayaranian Hosseini *et al.*, 2023] Mohammadreza Tayaranian Hosseini, Alireza Ghaffari, Marzieh S. Tahaei, Mehdi Rezagholizadeh, Masoud Asgharian, and Vahid Partovi Nia. Towards fine-tuning pre-trained language models with integer forward and backward propagation. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1912–1921, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[Wolf *et al.*, 2011] Lior Wolf, Tal Hassner, and Yaniv Taigman. Effective Unconstrained Face Recognition by Combining Multiple Descriptors and Learned Background Statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1978–1990, 2011.

[Xu *et al.*, 2021] Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. Raise a child in large language model: Towards effective and generalizable fine-tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9514–9528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Zemel *et al.*, 2013] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[Zhang *et al.*, 2018] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA, 2018. Association for Computing Machinery.

[Zhang *et al.*, 2023] Yixuan Zhang, Feng Zhou, Zhidong Li, Yang Wang, and Fang Chen. Fair representation learning with unreliable labels. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 4655–4667. PMLR, 25–27 Apr 2023.