Unsupervised Image Recognition De-Biasing via Meta-Learning

Wei Zhenyu
byte.miner.g@gmail.com

Abstract

Due to unbalanced datasets and lack of representation for certain populations, existing image recognition and classification algorithms have varying accuracy rates for people of different races, genders, skin colors, and even ages, which is unfair to them. This proposal researches on how to mitigate the algorithm's dependence on non-essential attributes in images (e.g., race, gender, skin color, age, etc.) via the model-agnostic meta-learning(MAML) based de-biasing method for unsupervised image recognition to improve the fairness and reliability of models. Specifically, we investigate how to solve the problem of the lack of full sample distributions with unbalanced datasets via few-shot learning by MAML, which motivates models to learn features that are generic rather than biased in recognizing humans, enabling arbitrary models with the ability of fairness adaptation, thus reducing biases and promoting social fairness and justice.

1. Introduction

The common vision tasks for modern robots are face recognition, autonomous driving, medical imaging, as well as recognizing, analyzing and predicting human activities. However, existing computer vision techniques are not always fair and inclusive, the recognition rate for humans is often affected by the non-essential features such as race, gender, religion, and age, causing the occurrence and even amplification of the pre-existing biases and discrimination in the datasets, such as the lower accuracy of recognition for darker-skinned races as well as females, which brings trouble to their lives, potentially places them at risk (e.g., autonomous vehicles miss detection of pedestrians) and may put innocent people on the list of criminals [1]. And the recognition of objects could also be affected by non-essential features, e.g. color schemes for vehicles, leading to bias. These situations result in discriminatory treatment for different groups in robots, which undermines social fairness and equality eventually [2].

Vision has become an essential capability for most robots, and the biases contained in vision algorithms have affected many minorities and special populations, hence a growing number of researchers are dedicated to mitigating or eliminating such biases. And with the advances in computer vision, vision classification and recognition algorithms are progressively shifting from supervised learning to semi-supervised and unsupervised learning. These emerging learning methods are typically trained with large amounts of unlabeled data to generalize features for the vision tasks, and these data and features usually contain biased information. For example, the existing datasets used for training human classification and recognition are underrepresented by minorities due to the imbalance in the quantity and quality of images in the vast number of pictures it collects indiscriminately from the Internet [3], [4]. Models trained with these unbalanced features are prone to inherit these biases, resulting in algorithms that are less inclusive and therefore less likely to recognize some certain groups in the society.

This proposal dives into the reasons why robots have biases against human populations in visual decision making and how to avoid such biases in the future. Our research is based on the observation as follows: visual bias and discrimination in advanced robots arise from the imbalance in datasets and the lack of inclusive feature extraction in algorithms, especially in the field of unsupervised learning. We argue that responsible robots should perform visual tasks fairly for all populations and we propose a method to mitigate biases via meta-learning, specifically, it applies model-agnostic meta-learning (MAML) [5] to the problem of fairness adaptivity in unsupervised image recognition. Our objectives are as follows:

1. Design a MAML-based de-biasing method for unsupervised image recognition that minimizes the dependence on non-essential attributes (e.g., race, gender, skin color, age, etc.) in the image data, so as to improve the fairness and reliability of the model.

2. Find a suitable unsupervised image recognition model combined with MAML and fairness loss function to achieve the ability to extract the necessary attributes.

3. Evaluate the robustness of this fairness optimization approach and the optimization ratio in visual tasks to demonstrate the effectiveness of the proposed algorithm.

2. Related Work

2.1. Fair computer vision

Computer vision-induced bias became known after ”Gender Shades” [6], first with gender bias due to racial differences, and later with factors such as skin color or gender differences. This may arise because of the social biases that influence the selection of datasets, the process of data labeling, the programming of engineers, and the scenario of algorithm application. Fairness evaluation and optimization in computer vision aims to research the potential bias and discrimination of algorithms in the performance of different populations, scenarios, and tasks, etc., and to propose methods and metrics for fairness evaluation and optimization to reduce the impact of computer vision on social inequities. Common approaches improve the algorithm or make it more transparent through the following three perspectives:

Pre-processing methods mitigate bias by using more diverse and balanced datasets. One approach is to generate datasets with more balanced samples via data generation methods like GAN [4], [7], [8]. Another way is to manipulate the dataset with data augmentation before the model training to attenuate the biased information in it [3], [9]–[11].

In-processing methods introduce fairness constraints or fairness adversarial training to algorithms to build models with higher interpretability. Fairness constraint training treats the fairness constraint as a regularization term during training, limiting the model's learning ability and mitigating the model's bias by narrowing the differences in non-essential features [12]–[17]. For example, Jung, Lee, Park, *et al.* [18] designs regularized terms for fair feature distillation to make student group conditions more consistent with average differences, and Liu, Deng, Zhong, *et al.* [19] adjusts the boundaries of each face feature category, thus balancing the issue of category imbalance and improving the fairness of face recognition. Whereas, in the fairness adversarial training, an adversary is introduced during the training process to perturb the input data so that the classifier cannot be reliant with non-essential features [20]–[22]. For example, Wang, Zhao, Yatskar, *et al.* [3] uses an additional classifier to remove gender-sensitive information from the image representation.

Post-processing methods evaluate and constrain the results of model inference after the model is trained. Usually an unfair algorithm is used to obtain classification predictions and then revise the results [23]. For example, Hendricks, Burns, Saenko, *et al.* [24] makes the model output equal probabilities for different genders by adjusting the gender probabilities in the generated descriptions, and Du, Mukherjee, Wang, *et al.* [25] adjusts the weights of the classifiers after they were trained so that the influence of non-essential attributes on the classifiers was minimized.

2.2. Meta-learning

Meta Learning, which can be referred to as ”learning to learn”, aims to quickly adapt new knowledge by learning generalized prior knowledge, such as recognizing the digits by ignoring non-essential features like brush angle and thickness. The purpose of traditional classification and recognition is to make final parameters achieve the best accuracy and the least loss. However, meta-learning is oriented towards the learning process rather than the result, which is to learn the experiences of classification and recognition [26], [27]. Existing meta-learning based methods perform object detection [28], [29], semantic segmentation [30], point cloud analysis [31] and multi-task learning [32] with a few labeled data and learn how to adapt quickly to new object classes from them. Some optimizations of the meta-learning fundamental methods have also been studied to improve accuracy and efficiency [33]–[35].

With the unbalanced datasets and lack of representation for certain populations, models can suffer from the lack of full sample distributions as in the case of few-shot learning, which in turn leads to bias. While MAML does not focus only on learning the distribution of individual categories over the sample space, but rather on the distribution of each category over the entire task space [36]. This can motivate models to learn features that are generic rather than biased in recognizing humans, enabling arbitrary models with the ability of fairness adaptation. MAML performs alternating optimizations on multiple tasks to find a universal model initialization parameter that makes the parameter available on any new task [37].

3. Method

Our research explores the problem of fairness adaptation by applying MAML to the unsupervised image recognition, where we need to prepare a meta training set that contains several datasets and tasks for image recognition of different people, and perform initial data preprocessing on them, each dataset represents a task $n \in (1, N)$. Then, we need to experiment to find a suitable model for unsupervised image recognition that can extract features from unlabeled images and perform clustering or other forms of recognition.

We also need to define a loss function based on image recognition fairness that measures the model's fairness performance on visual tasks. This loss function could be based on meta-learning

$$L(\hat{\theta}) = \sum_{n=1}^{N} l^n(\hat{\theta}^n), \tag{1}$$

where θ is a parameter of the network and l is the loss for each task. Then, we perform meta-training on the parameters of the models with MAML. Specifically, for each task, we need to randomly sample some images from its dataset as a meta-training set and some other images as a meta-test set. Then, we need to update the model with one or more gradient updates on the meta-training set and compute the fairness loss of the model on the meta-testing set, perform backpropagation and update the initial parameters of the model based on the loss, and we need to repeat this process on all tasks until the model's initial parameters converge

$$\hat{\theta} = \phi - \eta \nabla_{\phi} l(\phi). \tag{2}$$

Eventually, we get the image recognition model that extracts the necessary features which do not contain biased and discriminatory views such as race, gender, color, age, etc. When we need to expand it on a new dataset and task, we can use the parameters of the meta-trained model as the initial parameters of the new model, and with only a small amount of gradient updating of the model on the new dataset, we can obtain a fairness unsupervised image recognition model that is quickly adapted to the new task.

References

[1] K. H. Patrick Grother Mei Ngan, *Face recognition vendor test (frvt) part 3: Demographic effects*, Website, https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf, 2019.

[2] R. Steed and A. Caliskan, "Image representations learned with unsupervised pre-training contain human-like biases," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21, Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 701–713, ISBN: 9781450383097. DOI: 10.1145/3442188.3445932. [Online]. Available: https://doi.org/10.1145/3442188.3445932.

[3] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.

[4] N. Quadrianto, V. Sharmanska, and O. Thomas, "Discovering fair representations in the data domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.

[5] C. Finn, P. Abbeel, and S. Levine, *Model-agnostic meta-learning for fast adaptation of deep networks*, 2017. arXiv: 1703.03400 [cs.LG].

[6] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, S. A. Friedler and C. Wilson, Eds., ser. Proceedings of Machine Learning Research, vol. 81, PMLR, 23–24 Feb 2018, pp. 77–91. [Online]. Available: https://proceedings.mlr.press/v81/buolamwini18a.html.

[7] V. V. Ramaswamy, S. S. Y. Kim, and O. Russakovsky, "Fair attribute classification through latent space de-biasing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 9301–9310.

[8] K. Choi, A. Grover, T. Singh, R. Shu, and S. Ermon, "Fair generative modeling via weak supervision," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 13–18 Jul 2020, pp. 1887–1898.

[9] C.-Y. Chuang and Y. Mroueh, "Fair mixup: Fairness via interpolation," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=DNl5s5BXeBn.

[10] M. H. Sarhan, N. Navab, A. Eslami, and S. Albarqouni, "Fairness by learning orthogonal disentangled representations," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX*, Glasgow, United Kingdom: Springer-Verlag, 2020, pp. 746–761, ISBN: 978-3-030-58525-9. DOI: 10.1007/978-3-030-58526-6_44. [Online]. Available: https://doi.org/10.1007/978-3-030-58526-6_44.

[11] Y. Zhang and J. Sang, *Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing*, 2020. arXiv: 2007.13632 [cs.CV].

[12] X. Xu, Y. Huang, P. Shen, *et al.*, "Consistent instance false positive improves fairness in face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 578–586.

[13] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.

[14] V. S. Lokhande, A. K. Akash, S. N. Ravi, and V. Singh, "Fairalm: Augmented lagrangian method for training fair models with little regret," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*, Glasgow, United Kingdom: Springer-Verlag, 2020, pp. 365–381, ISBN: 978-3-030-58609-6. DOI: 10.1007/978-3-030-58610-2_22. [Online]. Available: https://doi.org/10.1007/978-3-030-58610-2_22.

[15] Z. Wang, K. Qinami, I. C. Karakozis, *et al.*, "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8916–8925. DOI: 10.1109/CVPR42600.2020.00894.

[16] Y. Savani, C. White, and N. S. Govindarajulu, "Intra-processing methods for debiasing neural networks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 2798–2810.

[17] B. Zhao, X. Xiao, G. Gan, B. Zhang, and S.-T. Xia, "Maintaining discrimination and fairness in class incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.

[18] S. Jung, D. Lee, T. Park, and T. Moon, "Fair feature distillation for visual recognition," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 110–12 119. DOI: 10.1109/CVPR46437.2021.01194.

[19] B. Liu, W. Deng, Y. Zhong, *et al.*, "Fair loss: Margin-aware reinforcement learning for deep face recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.

[20] S. Gong, X. Liu, and A. Jain, "Jointly de-biasing face recognition and demographic attribute estimation," in Oct. 2020, pp. 330–347, ISBN: 978-3-030-58525-9. DOI: 10.1007/978-3-030-58526-6_20.

[21] M. Wang and W. Deng, "Mitigating bias in face recognition using skewness-aware reinforcement learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.

[22] P. Li, H. Zhao, and H. Liu, "Deep fair clustering for visual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.

[23] G. Balakrishnan, Y. Xiong, W. Xia, and P. Perona, "Towards causal benchmarking of bias in face analysis algorithms," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII*, Glasgow, United Kingdom: Springer-Verlag, 2020, pp. 547–563, ISBN: 978-3-030-58522-8. DOI: 10.1007/978-3-030-58523-5_32. [Online]. Available: https://doi.org/10.1007/978-3-030-58523-5_32.

[24] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sep. 2018.

[25] M. Du, S. Mukherjee, G. Wang, R. Tang, A. Awadallah, and X. Hu, "Fairness via representation neutralization," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 12 091–12 103.

[26] C. Zhao, F. Chen, and B. Thuraisingham, "Fairness-aware online meta-learning," ser. KDD '21, Virtual Event, Singapore: Association for Computing Machinery, 2021, pp. 2294–2304, ISBN: 9781450383325. DOI: 10.1145/3447548.3467389. [Online]. Available: https://doi.org/10.1145/3447548.3467389.

[27] T. Wei and J. He, "Comprehensive fair meta-learned recommender system," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '22, Washington DC, USA: Association for Computing Machinery, 2022, pp. 1989–1999, ISBN: 9781450393850. DOI: 10.1145/3534678.3539269. [Online]. Available: https://doi.org/10.1145/3534678.3539269.

[28] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5149–5169, 2022. DOI: 10.1109/TPAMI.2021.3079209.

[29] B. Demirel, O. B. Baran, and R. G. Cinbis, "Meta-tuning loss functions and data augmentation for few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 7339–7349.

[30] X. Zhang, H. Zhang, J. Lu, L. Shao, and J. Yang, "Target-targeted domain adaptation for unsupervised semantic segmentation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 560–13 566. DOI: 10.1109/ICRA48506.2021.9560785.

[31] H. Lin, X. Zheng, L. Li, *et al.*, "Meta architecture for point cloud analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 17 682–17 691.

[32] R. Upadhyay, P. C. Chhipa, R. Phlypo, R. Saini, and M. Liwicki, *Multi-task meta learning: Learn how to adapt to unseen tasks*, 2023. arXiv: 2210.06989 [cs.CV].

[33] Y. Tu, B. Zhang, Y. Li, *et al.*, "Learning from noisy labels with decoupled meta label purifier," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 19 934–19 943.

[34] L. Wang, S. Zhou, S. Zhang, X. Chu, H. Chang, and W. Zhu, "Improving generalization of meta-learning with inverted regularization at inner-level," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 7826–7835.

[35] X. Qin, X. Song, and S. Jiang, "Bi-level meta-learning for few-shot domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 15 900–15 910.

[36] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17, Sydney, NSW, Australia: JMLR.org, 2017, pp. 1126–1135.

[37] H. Wang, Y. Wang, R. Sun, and B. Li, "Global convergence of maml and theory-inspired neural architecture search for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 9797–9808.