Introduction

Bias is prevalent in machine learning models and can lead to unfair outcomes and perpetuate existing inequalities. When a model systematically misrepresents certain groups due to biased training data or flawed assumptions in algorithm design, it can result in decisions that disadvantage those groups. For example, biased recruitment algorithms may favor certain demographics over others, or facial recognition systems might perform poorly with people of certain races or genders. Such biases not only compromise the effectiveness and fairness of automated systems but also have real-world impacts, potentially affecting job prospects, legal outcomes, and access to services. Addressing bias is crucial to developing ethical, equitable, and reliable technological solutions.

In the field of deep learning, bias often arises from spurious correlations—situations where variables appear causally linked despite there being no such relationship. To address this, researchers continuously explore methods to minimize bias, with techniques like upweighting, oversampling, and undersampling being commonly employed. However, each of these methods comes with its own challenges. While upweighting increases the influence of underrepresented classes during training, it can inadvertently lead to overfitting as the model becomes too tuned to the smaller samples.Similarly, oversampling replicates data from minority classes and risks introducing significant redundancy in the training data which can also lead to overfitting. Undersampling reduces the size of overrepresented classes and can lead to a loss of valuable information.

A novel approach called "Bias Mimicking" has been proposed by researchers from Boston University and the MIT-IBM Watson AI Lab. This method is designed to mitigate bias by ensuring that the distribution of bias-related groups across each class is uniform, thereby rendering the bias statistically independent from the target variable. Bias Mimicking achieves this by subsampling each class to mirror the bias distribution seen in other classes. This technique allows the model to learn from every data point within the dataset while maintaining a balanced representation of groups that might introduce bias, such as gender or race.

Analysis of Methods

Bias mimicking simplifies the process of bias mitigation as it does not require additional hyperparameters or intricate adjustments, which typically require extensive testing and optimization. By focusing on straightforward subsampling and the training of binary classifiers, The subsampling strategy employed in Bias Mimicking utilizes linear programming to efficiently identify optimal subsampling solutions, ensuring quick and effective bias mitigation. Linear programming is a mathematical method used to determine the best possible outcome in a given mathematical model with certain constraints. In the context of Bias Mimicking, linear

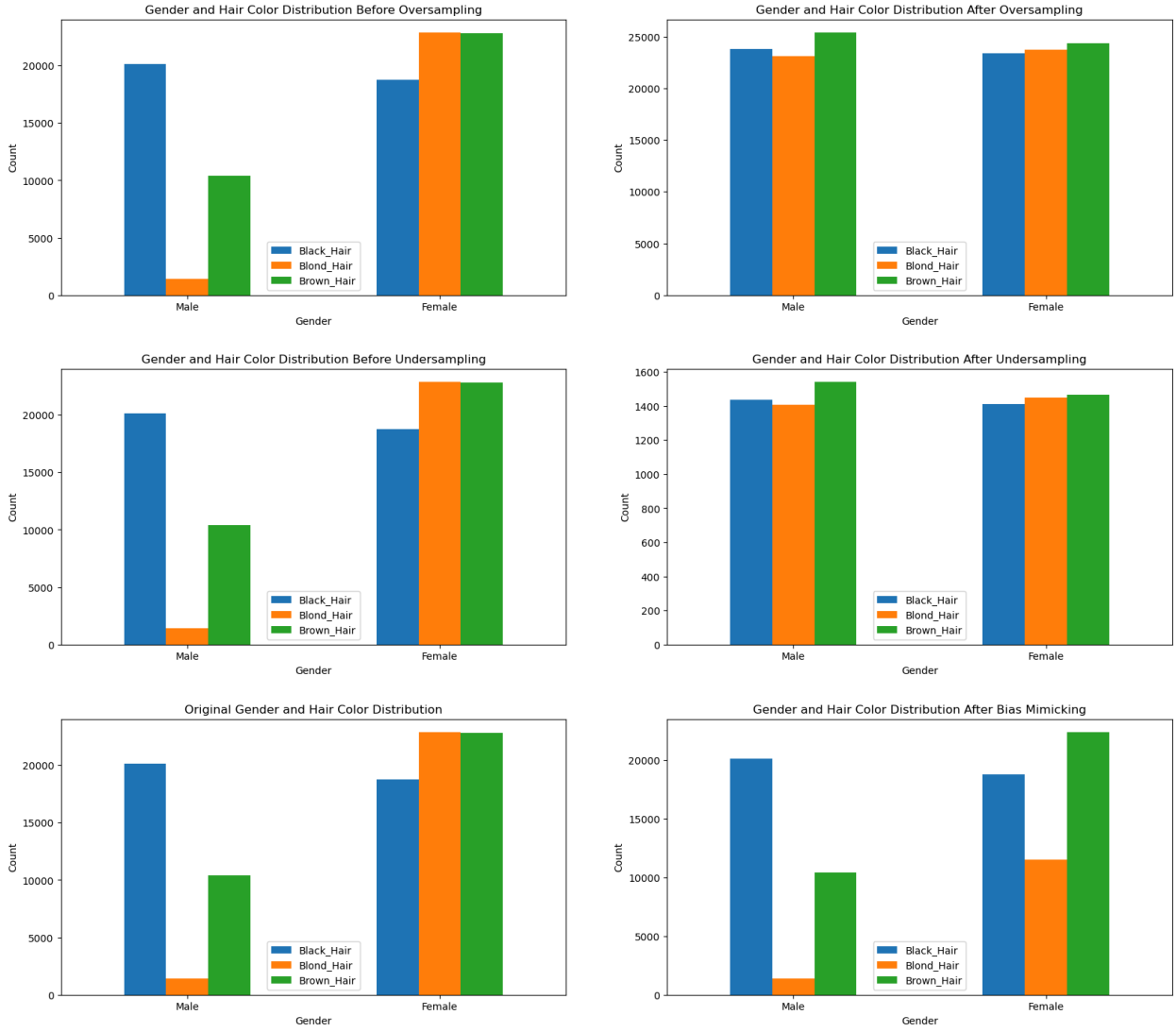
programming can be utilized to efficiently solve the problem of how to subsample the dataset to achieve the desired uniform bias distribution across all classes.

The core idea behind bias mimicking is to achieve statistical independence between the potential sources of bias (features) and the outcome variable (target). Statistical independence between two variables means that the occurrence of one variable does not affect the probability of occurrence of the other variable. If the distribution of bias groups with respect to a class c, $PD(B|Y = c)$, was the same across every $c \in C$, then B is statistically independent from Y. For bias mimicking, the goal is to adjust the sampling or weighting of the dataset so that the feature distribution for each category of the target variable is the same.

Researchers tested the proposed method on three datasets: CelebA, UTKFace, and CIFAR-S. Two main evaluation metrics were used: Unbiased Accuracy (UA) and Bias-Conflict (BC). UA measures the accuracy of each subgroup separately and then returns the mean of accuracies, while BC measures the accuracy on the minority subgroups. Several baselines from prior work were included for comparison. These baselines included non-sampling methods like Bias-Contrastive and Bias-Balanced Learning (BC + BB), Domain-Independent (DI), GroupDRO, and Adversarial Learning (Adv) with uniform confusion.

Bias Mimicking was the only technique that constantly performed well on the three datasets tested in the experiment: CelebA, UTK-Face, and CIFAR-S. Undersampling performed decently on the CelebA dataset, but struggled on other datasets. Oversampling consistently underperformed on all datasets, likely due to overfitting issues. Upweighting performed well on CelebA and UTK-Face but fell short on CIFAR-S. Even though Undersampling, Oversampling, and Upweighting were effective on some datasets, they failed to be as reliable as Bias Mimicking. Additionally, when compared to non-sampling methods, Bias Mimicking achieved similar results without the need for additional complex loss functions or model modifications.

Each of the three datasets contain one or more spurious correlations relating to hair color, gender and race. CelebA is a commonly used dataset containing over 200,000 images of celebrities with 40 attribute annotations, such as wavy hair, blue eyes, etc. Due to the nature of the data, models trained on CelebA often correlate blond hair with females so researchers used oversampling, undersampling, and bias mimicking to mitigate bias and improve the model's function. I also used the CelebA dataset to verify and validate the research study. I developed simple functions based on the principles of each of the three techniques and applied them to the CelebA dataset. The graphs below demonstrate the gender and hair color distribution before and after each of the techniques were used.



I further validated the study by measuring the difference in mean squared error (mse) between the original model and the model that used bias mimicking. The difference in mse between the two models was 21,557,308.5 indicating that bias mimicking introduces large-scale changes to attribute distributions in the dataset. I further tested for statistical significance by conducting a chi squared test on the two distributions. The test generated a chi squared statistic of 2720.95 and a p-value of less than 0.05 which means that bias mimicking significantly altered the gender and hair color distribution. The results of the mse and chi squared tests demonstrate the power bias mimicking has to reshape data and improve accuracy. However, my analysis is limited in its ability to grasp the accuracy and efficacy of bias mimicking as it did not compare the method to other sampling methods like oversampling and undersampling.

Normative Consideration

Biased machine learning models can perpetuate and even amplify existing societal biases and discrimination, making the mitigation of bias essential for developing responsible deep learning technologies. Users, particularly those who may not be well-versed in statistical intricacies, often place significant trust in these models, expecting them to operate fairly. This trust can lead to severe consequences when models that are presumed to be unbiased inadvertently reinforce biases, thus misleading users and affecting lives negatively. As AI and deep learning become increasingly prevalent across various industries, the responsibility on statisticians and developers to ensure these models are equitable and trustworthy grows.

Bias Mimicking presents a promising approach by simplifying, speeding up, and reducing the costs associated with mitigating bias, thereby supporting the development of fairer models. However, this method also has inherent limitations that must be critically examined from an ethical standpoint. The original study on Bias Mimicking primarily considered scenarios with mutually exclusive sensitive groups, implying that each sample belongs to only one sensitive group. This approach overlooks the concept of intersectionality, which is crucial for a comprehensive understanding of biases.

Intersectionality acknowledges that individuals often belong to multiple overlapping groups, such as race, gender, and class, and that these overlapping identities can compound experiences of discrimination. By ignoring intersectionality, Bias Mimicking oversimplifies the complexities of human identities, leading to solutions that fail to fully address how biases interplay and affect individuals at the intersections of these groups. From a moral and ethical perspective, this oversight challenges the core principles of justice and respect for persons, as it fails to recognize the full dignity and complexity of all individuals, particularly those who are most vulnerable to systemic inequalities. More research needs to be conducted on bias mimicking to ensure that models which use the method can address the complexities of intersectionality.

Incorporating ethical frameworks, such as utilitarianism or deontological ethics, into the development and evaluation of Bias Mimicking could provide more robust guidelines for ensuring fairness. Utilitarianism would advocate for modifications to Bias Mimicking that maximize overall happiness by reducing harm through more inclusive and representative training methods. Conversely, a deontological approach would emphasize the inherent duty to treat all individuals with fairness, respecting their diverse identities and avoiding any form of discrimination, regardless of the broader outcomes. To truly advance the fairness of AI, developers must engage deeply with these ethical considerations, ensuring that their models do not just superficially address bias but are genuinely designed to respect and reflect the rich, intersecting identities that characterize human societies.

Conclusion

As the utilization of artificial intelligence and machine learning technologies becomes increasingly widespread across various sectors, the imperative to address and mitigate biases within these systems cannot be overstated. Bias in machine learning, if left unchecked, can perpetuate and even exacerbate existing societal inequalities, leading to outcomes that unfairly disadvantage certain groups. The novel approach of Bias Mimicking, developed by researchers from Boston University and the MIT-IBM Watson AI Lab, presents a significant advancement in the field by offering a method to achieve statistical independence between bias sources and target outcomes. This approach not only simplifies the technical process of mitigating bias through its reliance on linear programming for efficient subsampling but also reduces the need for extensive hyperparameter tuning.

While Bias Mimicking contributes positively towards the creation of fairer and more equitable machine learning models, it also introduces new challenges, particularly in its current consideration of sensitive groups. The method's limitation in addressing the complexities introduced by intersectionality is a notable drawback. Intersectionality is vital in understanding the multifaceted nature of discrimination and ensuring that technologies do not exacerbate these disparities. Incorporating moral frameworks into the development and assessment of bias mitigation methods is essential. Techniques like Bias Mimicking should strive to align with ethical principles such as utilitarianism, which seeks the greatest good for the greatest number, and deontological ethics, which focuses on duties and rights. By adhering to these ethical standards, developers can ensure that their models are not only technically proficient but also morally sound, respecting the diverse identities and experiences of all individuals.

Citation

Qraitem, M., Saenko, K., & Plummer, B. A. (2023b). Bias mimicking: A simple sampling approach for bias mitigation. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https: //doi.org/10.1109/cvpr52729.2023.01945