Midterm

2024-03-22

Introduction

Bias in deep learning algorithms is often associated with spurious correlation, a phenomenon where two or more variables appear to be causally related when in actuality they are not related in that way. Inaccurate algorithmic prediction as a result of bias can lead to real harm depending on how the algorithm is used so it's imperative that developers take steps to mitigate bias as much as possible. There are several ways to mitigate bias including sampling techniques oversampling, undersampling, and upweighting but each of these have their own set of problems. The paper Bias Mimicking: A Simple Sampling Approach for Bias Mitigation from researchers at Boston University and MIT-IBM Watson AI Lab proposes bias mimicking as a new way to reduce bias caused by spurious correlation.

Bias Mimicking

Bias Mimicking is based on the observation that if the distribution of the bias groups with respect to each class is the same across all classes, then bias is statistically independent of the target variable. Bias mimicking involves subsampling the class so that each class's bias distribution is mimicked across all other classes which ensures that the model learns from every observation in the dataset while maintaining its desired bias distribution. Essentially, the model identifies groups in the data that might introduce bias, like gender or race, and then divides the data into smaller sets, ensuring each set contains a fair representation of different groups. The computer is trained separately on each of these balanced sets.

Bias Mimicking automatically addresses bias in the dataset without requiring additional hyperparameters or complex tuning. This means there's less need for extensive experimentation and optimization, reducing time and resources spent on hyperparameter search. Instead of introducing new layers or complexities, Bias Mimicking uses a straightforward approach of subsampling and training binary classifiers, which leads to faster training times compared to more elaborate methods. The process of subsampling in Bias Mimicking is optimized using linear programming techniques to quickly find the best solution for subsampling while meeting the desired constraints.

Summary of Methods

The study tested the performance of Bias Mimicking, as well as Undersampling, Oversampling, and Upweighting in mitigating bias in deep learning models across different datasets. The researchers selected three main datasets for evaluation: CelebA dataset, UTKFace dataset, and CIFAR-S benchmark. Each dataset represented different classification tasks with biased attributes, such as gender, race, age, and hair color. CelebA focused on predicting BlondHair while considering Gender bias. UTKFace focused on predicting Gender while considering Race/Age bias. And CIFAR-S focused on introducing bias synthetically into by converting some images to grayscale.

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.

Summary of Results

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. And Upweighting performed well on CelebA and UTK-Face but fell short on CIFAR-S. Even though Undersample, 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.

Normative Consideration

Biased models can perpetuate and exacerbate societal bias and discrimination so mitigating bias is crucial to developing deep learning models. Users, especially users who are unfamiliar with the nuances of statistics, trust models to be fair and it can be dangerous if something that is expected to be fair proves to perpetuate bias. AI and deep learning models are becoming common in every industry and the more these models are used the more statisticians and developers have to be careful to create models that are fair and trustworthy. Bias Mimicking makes it easier, faster, and cheaper for developers to mitigate bias in their model which facilitates the creation of fairer models.

However, Bias Mimicking comes with its own set of limitations. The bias scenarios that were explored in the study only considered mutually exclusive sensitive groups, where a sample can belong to only one sensitive group at a time. This doesn't account for intersectionality, which is an important consideration in mitigating bias because intersectionality reflects the reality of the human experience. Ignoring intersectionality can lead to oversimplified solutions that fail to address the complex ways in which bias operates.

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