CERIAS

The Center for Education and Research in Information Assurance and Security

Unlearning Machine Learning Bias using Task Vector Arithmetic

Omkar Pote, Romila Pradhan
{opote, rpradhan}@purdue.edu
Department of Computer and Information Technology, Purdue University, West Lafayette, IN 47907, United States

Introduction

Machine learning models have become integral to decision-making in fields such as criminal justice, finance, and healthcare. However, these models often inherit biases present in their training data, leading to **unfair and unethical outcomes**, particularly for marginalized groups. Recent work in the area of natural language processing has hypothesized bias to be a linear subspace in word embeddings. We evaluate the applicability of this concept to model weights of structured data, and introduce a novel task arithmetic based approach to unlearn bias in tabular datasets.

Our method selectively **unlearns biases** introduced during training by fine-tuning a model on high-bias data, computing a **bias task vector**, and subtracting it from the original model to mitigate unwanted biases. Our evaluations show that this approach is competitive with state-of-the-art bias mitigation techniques, significantly improving fairness on several metrics with minimal accuracy loss.

This technique is evaluated on well-known datasets with documented biases:

- **COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)**: Used for recidivism prediction but has been shown to disproportionately classify Black defendants as high risk.
- **Adult Income Dataset**: Predicts income levels but exhibits **gender biases**, often underestimating female earnings potential.

Evaluation

- **Datasets**
  - COMPAS (7214 rows x 12 columns)
  - Adult Income (45221 rows x 15 columns)
- **Fairness Metrics**: For a sensitive attribute A, we use the following established fairness metrics:
  - Demographic Parity Difference (DPD) measures the difference in favorable outcome across groups.
    $DPD = P(\hat{Y}=1 \mid A=0) - P(\hat{Y}=1 \mid A=1)$
  - Equal Opportunity Difference (EOD) compares true positive rates across demographics.
    $EOD = TPR(A=0) - TPR(A=1)$
  - False Positive Rate Difference (FPRD) assesses discrepancies in false positive rates.
    $FPRD = FPR(A=0) - FPR(A=1)$
  - Predictive Parity Difference (PPD) evaluates the consistency of precision across groups.
    $PPD = Precision(A=0) - Precision(A=1)$

(values closer to 0 represent higher fairness)

References

- Ilharco, G.; Ribeiro, M. T.; Wortsman, M.; Gururangan, S.; Schmidt, L.; Hajishirzi, H.; and Farhadi, A. (2023). Editing Models with Task Arithmetic. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*.
- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems 29 (NIPS 2016)*

Methodology

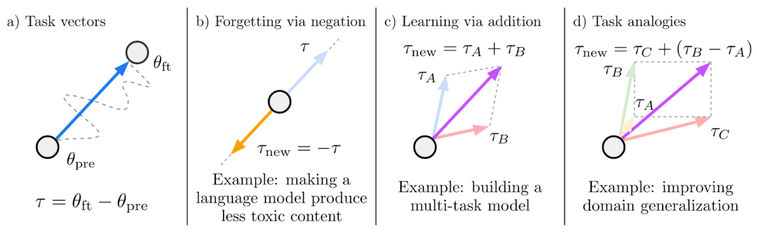

Figure 1: An illustration of task vectors and the arithmetic operations (Ilharco et al. 2023)

Our approach builds on task arithmetic, which involves editing model weights to add or remove specific learned concepts. The key steps are:

1. Baseline Model Training: A TabNet model is trained on the dataset, establishing a starting point for predictions.

2. Fine-Tuning on High-Bias Data: The model is exposed to a subset of highly biased data, causing it to amplify existing biases.

3. Bias Task Vector Computation: Let the parameters of the original pre-trained model be denoted as θ_0, and the parameters of the biased fine-tuned model be denoted as θ_{biased}. The task vector $\Delta\theta_{bias}$ is given by:

$$\Delta\theta_{bias} = \theta_{biased} - \theta_0$$

4. Task Vector Subtraction: To "unlearn" bias, we subtract the task vector and apply the obtained weights to a new instance of the original model:

$$\theta_{unbiased} = \theta_0 - \Delta\theta_{bias}$$

We further extend the use of task arithmetic by introducing a scaling factor (λ) to regulate the impact of the task vector negation:

$$\theta_{unbiased} = \theta_0 - \lambda\Delta\theta_{bias}$$

We conduct an additional experiment to observe the impact on fairness and accuracy, of high-bias negation against direct low-bias fine-tuning (Figure below).

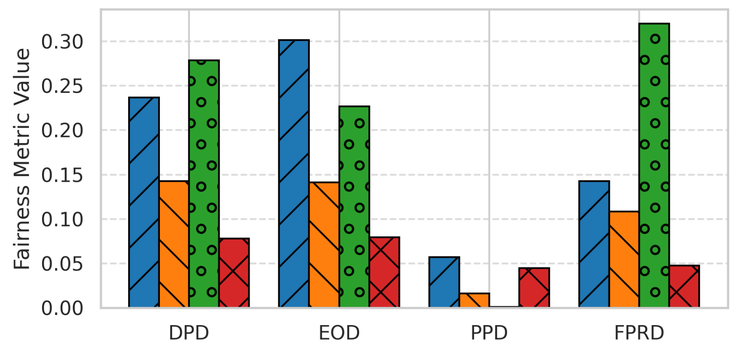

Figure 2: Fairness metrics of different models on the COMPAS dataset (Values between ± 0.1 signify high fairness)

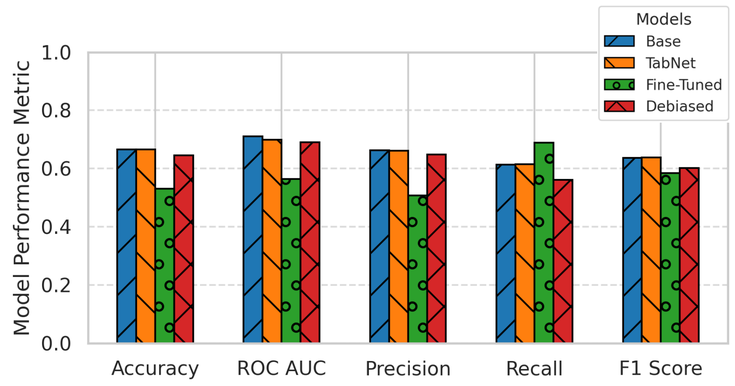

Figure 3: Accuracy metrics of different models on the COMPAS dataset

Conclusion

- The task arithmetic approach demonstrates **significantly lower bias**, with fairness metrics improving across both datasets.
- A **small reduction in accuracy (2-3%)** is observed, but the fairness gains are substantial, making the trade-off justifiable, particularly in high-stakes applications like criminal justice.
- Fine-tuning on low-bias data results in overfitting, confirming that **bias subtraction is a superior approach** for accuracy retention.
- **Task Arithmetic is a scalable and adaptable** to different datasets and models, requiring minimal retraining.

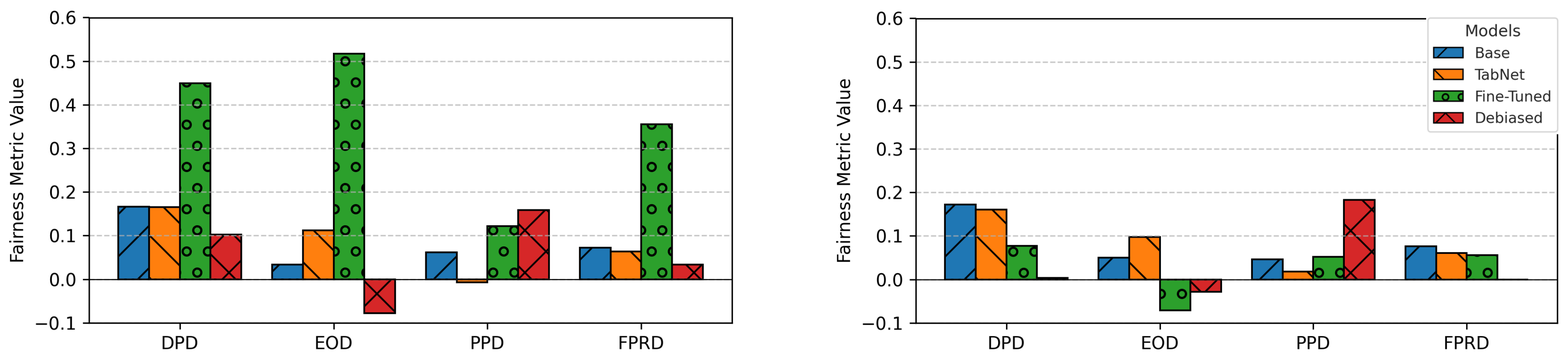

Figure 4: Fairness metrics of models on the Adult Income dataset
Part A (left) depicts high-bias fine-tuning followed by subtraction, Part B (right) depicts low-bias fine-tuning followed by further extrapolation (λ = 0.5 in both cases).
(Values between ± 0.1 signify high fairness)

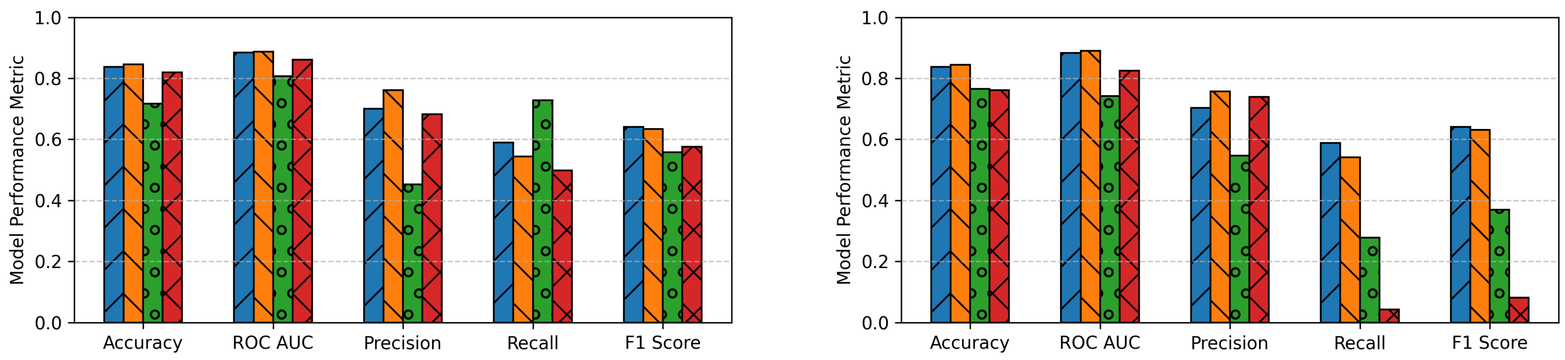

Fig. 5: Accuracy metrics of different models on the Adult Income dataset
Part A (left) depicts high-bias fine-tuning followed by subtraction, Part B (right) depicts low-bias fine-tuning followed by further extrapolation (λ = 0.5 in both cases).
We see a significant drop in accuracy in Part B.

PURDUE UNIVERSITY

We are the Responsible Data Science Lab

To find out more about our work

Scan this QR or go to

https://romilapradhan.github.io/

CERIAS