FairFusion: Distributionally Robust Fair-Multimodal Learning for College Admissions

Multimodal learning techniques have shown considerable promise in improving predictive performance across varying data modalities. This perfectly aligns in the context of college admissions, where we deal with both text data such as essays, Personal Insight Questions (PIQs), and tabular data such as GPA, test scores, and demographics. In multimodal fusion, empirical risk minimization encourages the model to exploit the correlations across modalities whilst inadvertently encoding the inherent bias. Recent analysis [1] demonstrate the presence of spurious correlations in training data driving the cross-attention module to learn them and effectively amplify the existing biases. Bias amplification in the fusion layer, in turn, makes the predictions highly sensitive to group-specific features. Moreover, as cross-attention treats one modality as the query and another as the key/value, any misalignment between the latent spaces of these modalities could result in unstable attention weights. This poses the threat for the model to over-rely on a single modality.

These challenges in the decision-making might inadvertently affect the socioeconomically challenged applicants. Given that college admissions is a high-impact, sensitive problem, learning models could perpetuate the historical bias in the training data into the model predictions. In addition, the applicant demographics changes annually, posing a high possibility of distributional shifts in data. This renders the task highly critical, where it is imperative to ensure model robustness to such shifts while maintaining equitable performance across sensitive groups.

To circumvent these issues, we propose a fairness-aware multimodal learning framework that integrates a gradient-norm-based Wasserstein Distributional Robust Optimization (W-DRO) regularizer. Our architecture employs two heads- a classifier head, for predicting the admission decisions, and an adversary head, to prevent the sensitive attribute leakage into the classifier's predictions. Next, we apply the W-DRO penalty in the fusion layer to regularize the joint learning process. The architectural choice to add the penalty term at the fusion layer complements the adversarial learning setup. Specifically, it facilitates the fairness penalties to operate on a well-aligned representation. It also ensures the model's robustness to both spurious correlations and distributional shifts under the fairness constraints. We exploit the theoretical motivations from distributionally robust optimization literature [2], and extend it to the fair multimodal setting with cross attention. Unlike the prior works that focus on fairness-aware fusion modeling or assume uni-modal robustness, our framework achieves fair-multimodal learning robust to distributional shifts. We provide performance guarantees to substantiate our claims under bounded worst-case data perturbations. This penalty constraints the Lipschitz behavior of the model, ensuring that the fairness constraints enforced during the training generalize reliably under test-time distribution shifts.

For our experimental setup, we chose demographic parity and $p\%$-rule as the fairness definitions. We conduct experiments on a set of two real-world applicants datasets acquired from the Dept. of Computer Science at an $R1$ University. It includes both one year, and a seven year worth applicants data, to demonstrate the model's scalability. The initial experiments performed with respect to socioeconomic sensitive attributes showcase the demographic parity gap to reduce from 0.3049 to 0.0759 with $p\%$-rule over 80%. This is achieved with an almost negligible drop in predictive accuracy. Along with that we also conduct an extensive ablation study and compare our model performance to baseline models that further affirm our claims.

[1] You, C., et al. "Calibrating multi-modal representations: A pursuit of group robustness without annotations" *CVPR*, 2024, https://arxiv.org/abs/2403.07241.

[2] Sinha, A., et al. "Certifying some distributional robustness with principled adversarial training" *ICLR*, 2018, https://arxiv.org/abs/1710.10571.