Fairness Reprogramming

Guanhua Zhang[1*], Yihua Zhang[2*], Yang Zhang[3], Wenqi Fan[4], Qing Li[4], Sijia Liu[2, 3], Shiyu Chang[1]

[1]University of California, Santa Barbara, [2]Michigan State University, [3]MIT-IBM Watson AI Lab, [4]The Hong Kong Polytechnic University

([*]Equal Contribution)

Code Paper

NEURAL INFORMATION PROCESSING SYSTEMS

What is Model Reprogramming and Why?

- **Input-agnostic** perturbation to manipulate model behavior.
- Sharing similar ideas with text and visual prompting[1, 2].
- Model weights are sometimes not alterable or inaccessible.
  - large computational storage costs
  - low data efficiency
  - model privacy issues

Motivations

Can an unfair model be reprogrammed to fair one?
If so, why and how would it work?

Contributions

- FairReprogram: a novel generic fairness-enhancing paradigm in a min-max fashion through model reprogramming for both NLP and CV tasks.
- Theoretically and empirically show **why and how** fairness can be promoted using an input-agnostic fairness trigger.
- Compared to retraining-based methods[3], FairReprogram promotes model fairness with far less trade-off in accuracy across various NLP and CV datasets with in-the-wild biases.

Fairness Metrics

- Equalized Odds (EO): $\hat{Y} \perp Z|Y$, $\sum_{z\in\mathcal{Z}} (|\text{FPR} - \text{FPR}_z| + |\text{FNR} - \text{FNR}_z|)/2$,
- Demographic Parity (DP): $\hat{Y} \perp Z$, $\sum_{z\in\mathcal{Z}} |\text{p}(\hat{Y} = 1) - \text{p}(\hat{Y} = 1|Z = z)|$,

Related Work

[1] Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation.
[2] Bahng H, et al. Visual Prompting: Modifying Pixel Space to Adapt Pre-trained Models.
[3] Zhang B H, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning

Fairness Reprogramming

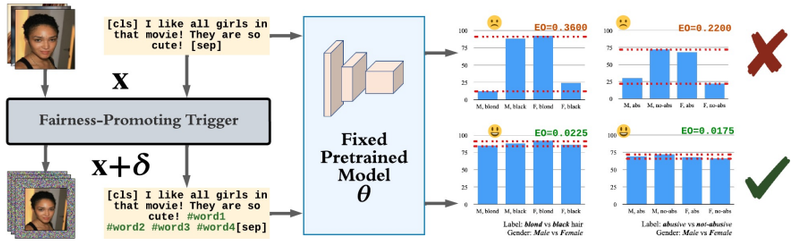


Figure 1. An example of fairness reprogramming in CV and NLP tasks. The input-agnostic trigger can promote fairness without altering the pretrained model.

- Optimization objective:
  - Optimizable variables: fairness trigger: $\boldsymbol{\delta}$ discriminator: $\boldsymbol{\phi}$
  - Overall optimization objective:

$$\min_{\boldsymbol{\delta},\boldsymbol{\theta}} \ \mathcal{L}_{util}(\mathcal{D}_{tune}, f^* \circ m) + \lambda \mathcal{L}_{fair}(\mathcal{D}_{tune}, f^* \circ m),$$

  - Task utility loss:

$$\mathcal{L}_{util}(\mathcal{D}_{tune}, f^* \circ m) = \mathbb{E}_{\boldsymbol{X},Y\sim\mathcal{D}_{tune}}[\text{CE}(Y, f^*(m(\boldsymbol{X})))]$$

  - Discriminator-based fairness promoting loss:

$$\mathcal{L}_{fair}(\mathcal{D}_{tune}, f^* \circ m) = \max_{\boldsymbol{\phi}} \mathbb{E}_{\boldsymbol{X},Y,Z\sim\mathcal{D}_{tune}}[-\text{CE}(Z, d(f^*(m(\boldsymbol{X})), Y; \boldsymbol{\phi}))]$$

- Why does fairness trigger work?

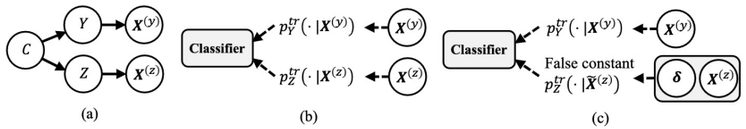
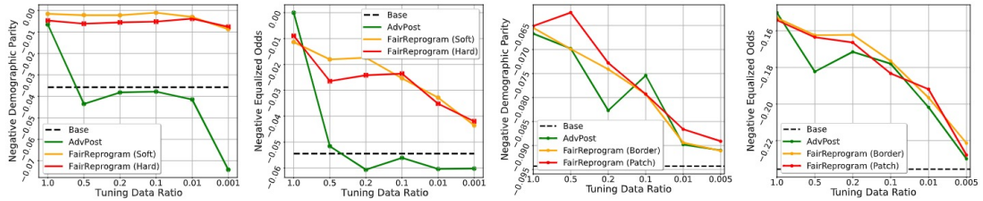


Figure 2. Illustration of why fairness trigger works. (a) Data generation. (b) Information flow from data to the classifier through the sufficient statistics. (c) Fairness trigger **strongly indicative of a demographic group** can confuse the classifier with a false demographic posterior, and thus preventing the classifier from using the correct demographic information.

Experiment Results Highlights



(a) `Civil Comments` (b) `CelebA`

Figure 4. (Main Results) Results on (a) Civil Comments and (b) CelebA. We report the negative DP (left) and the negative EO (right) scores. We vary the trade-off parameter λ to record the performance. The closer a dot to the upper-right corner, the better the model is.

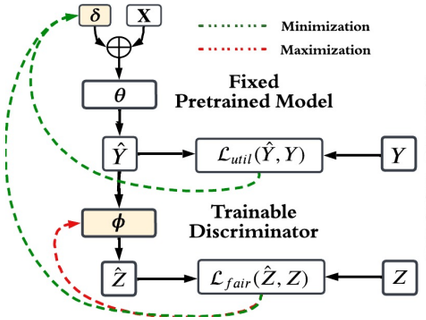


Figure 3. (Algorithm Illustration) An illustration of the FairReprogram algorithm pipelines formulized in a min-max fashion.

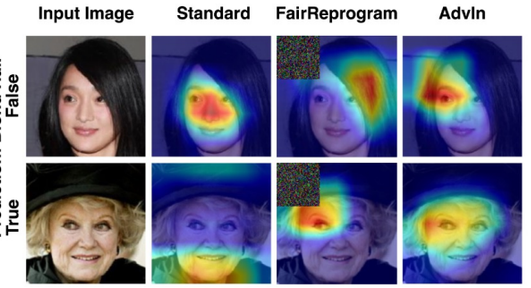
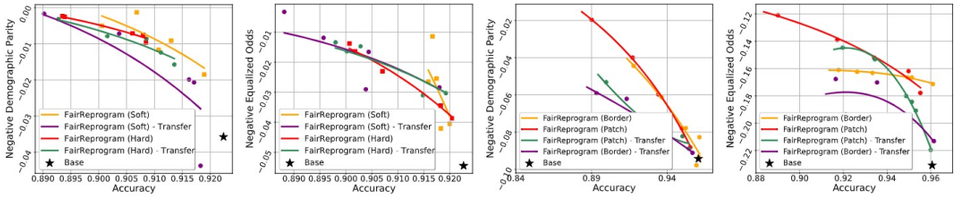


Figure 5. (Input Saliency Analysis) Gradient-based saliency map visualized with GradCAM. Highlighted zones (in red) are with major influence on predicted labels.



(a) `Civil Comments` (b) `CelebA`

Figure 6. (Transfer Setting) We report negative DP (left) and negative EO (right) scores. The triggers are firstly trained in a BASE model. Then we evaluate the triggers based on another unseen BASE model.

Contact: {guanhua@ucsb.edu, zhan1908@msu.edu}