Last-Layer Fairness Fine-tuning is Simple and Effective for Neural Networks
(ICML 2023)

SeHyun Park

February 22, 2024

Seoul National University

Outline

Preliminaries

Introduction

- Recent research has shown that adding fairness constraints to the objective function leads to severe over-fitting to fairness criteria in large models, especially when the training data is imbalanced.
- To handle this, in this paper, the authors propose a simple and effective framework by fine-tuning only the last layer of a pre-trained model.

Notation

- $D = \{(x_i, a_i, y_i)\}_{i=1}^{N}$: **Dataset**
  - $x_i \in \mathcal{X}$: Feature vector
  - $a_i \in \mathcal{A}$: Sensitive attribute (such as race or gender)
  - $y_i \in \mathcal{Y}$: Label
  - For simplicity, in this paper, it is assumed that a and y are binary variables.

- **Cross entropy loss**
  - $\hat{L}(h) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(h(x_i))$
  - where $h \in \mathcal{H} : \mathcal{X} \mapsto [0, 1]$

Fairness

1. **Equalized odds (EO)**
   - Definition : $\mathbb{P}(\hat{y} = 1|a = 0, y = y) = \mathbb{P}(\hat{y} = 1|a = 1, y = y) \ \ \forall y \in \{0, 1\}$
   - Objective : $\min_h \left[\hat{L}(h) + \alpha(f_{pr} + f_{nr})\right]$

     where $f_{pr} = \left|\frac{\sum_i h(x_i)(1-y_i)a_i}{\sum_i a_i} - \frac{\sum_i h(x_i)(1-y_i)(1-a_i)}{\sum_i (1-a_i)}\right|$ &

     $f_{nr} = \left|\frac{\sum_i (1-h(x_i))(1-y_i)a_i}{\sum_i a_i} - \frac{\sum_i (1-h(x_i))(1-y_i)(1-a_i)}{\sum_i (1-a_i)}\right|$
2. **Accuracy equality (AE)**
   - Definition : $\mathbb{P}(\hat{y} \neq y|a = 0) = \mathbb{P}(\hat{y} \neq y|a = 1)$
   - Objective : $\min_h \left[\hat{L}(h) + \alpha \left|\hat{L}^{a+}(h) - \hat{L}^{a-}(h)\right|\right]$

     where $\hat{L}^{a+}(h)$ and $\hat{L}^{a-}(h)$ are the cross entropy loss of samples with respectively $a = 1$ and $a = 0$
3. **Max-Min fairness (MMF)**
   - Definition : $\max \min_{y \in \mathcal{Y}, a \in \mathcal{A}} \mathbb{P}(\hat{y} = y|y, a)$
   - Objective : $\min_h \max\{\hat{L}^{(y+,a+)}(h), \hat{L}^{(y-,a+)}(h), \hat{L}^{(y+,a-)}(h), \hat{L}^{(y-,a-)}(h)\}$

     where $\hat{L}^{(y',a')}(h)$ denotes the cross-entropy loss on the training samples where $y = y'$ and $a = a'$

Problem Background

Problem Background

- **Challenges in training fair neural networks with in-processing techniques**
  - Cherepanova et al. (2021)[1] finds that large models tend to overfit fairness objectives, leading to poor performance on unseen data in terms of fairness, especially when the training data is imbalanced.
  - Two main approaches
    1) train the full neural network with fairness constraints
    2) train a model without fairness constraints and fine-tune the full neural network

Inspiration

- **Standard training can still learn core features on imbalanced datasets**
  - Kirichenko et al. (2022)[2] found that when fixing the representation in a pre-trained model and only fine-tuning the last linear layer, the performance was similar or even better compared to the conventional approach, even when dealing with imbalanced datasets.

Inspiration

- **Access to the preceding property**
  - Model : pre-trained ResNet-18 on the original CelebA dataset using ERM
  - Data : customized hair-only CelebA($= D_H$)
  - experiment :
    1) Divide the D_H into two sets: D_H^{Tr} and D_H^{Te}.
       where D_H^{Tr} is evenly balanced comprise 107 images from each (a, y).
    2) fine-tune the last layer of the model on D_H^{Tr} and evaluate it on D_H^{Te}
  - Result

Train	Test (Worst/Mean)	
	Original	Hair-only
Original	0.268/0.946	0.863/0.878
Balanced	0.789/0.835	0.827/0.843

Fair Deep Feature Reweighting(FDR)

Fair Deep Feature Reweighting(FDR)

- Fair Deep Feature Reweighting(FDR)
  - Step 1: pre-train a representation
    1) train a neural network $\mathcal{N}$ with ERM
    2) Let $\mathcal{N} = w \circ \Phi$, where w is the last layer.
  - Step 2: fine-tune the last layer with reweighting and fairness constraints.
    1) fix Φ
    2) sample a small dataset D_r from the training dataset D such that each (a, y) group in D_r has the same number of samples.
    3) fine-tune w^{new} on D_r with standard ERM and fairness constraints.
    4) final model is $\mathcal{N}^{new} = w^{new} \circ \Phi$

Experiments

Experiments

- Setup
  - Dataset : CelebA
  - Compared methods
    - **FullFT-Reg** : Impose the fairness constraints on the training objective and train the full neural network.
    - **LastFT** : Fine-tune the last layer of a pre-trained model on the imbalanced validation dataset
    - **LastFT-RW** : Fine-tune the last layer of a pre-trained model on the balanced dataset.
    - **LastFT-Reg** : Fine-tune the last layer of a pre-trained model on the imbalanced validation dataset with fairness constraints.

Experiments

- Setup
  - Metrics
    - WACC : weighted accuracy
    - AUC : area under the ROC Curve
    - $\mathrm{EO}_{\mathrm{Diff}} : \max_{y \in \mathcal{Y}} \{|\mathbb{P}(\hat{y} = 1 \mid a = 0, y) - \mathbb{P}(\hat{y} = 1 \mid a = 1, y)|\}$
    - $\mathrm{AE}_{\mathrm{Diff}} : |\mathbb{P}(\hat{y} \neq y \mid a = 0) - \mathbb{P}(\hat{y} \neq y \mid a = 1)|$
    - $\mathrm{WA} : \min_{a \in \mathcal{A}, y \in \mathcal{Y}} \{\mathbb{P}(\hat{y} = y \mid a, y)\}$
    - AF = (WACC - $\mathrm{EO}_{\mathrm{Diff}}$ or WACC - $\mathrm{AE}_{\mathrm{Diff}}$ or WACC + WA)

Experiments

- Result

Fairness Notion 1: EO	WACC		AUC		EO_Diff		AF
	Train	Test	Train	Test	Train	Test	Test
FullFT-Reg	1.000	**0.914**	1.000	0.969	0.000	0.499	0.415
LastFT	0.918	0.913	0.974	**0.971**	0.308	0.327	0.586
LastFT-RW	0.913	0.908	0.970	0.968	0.100	0.207	0.701
LastFT-Reg	0.898	0.901	0.971	0.969	0.177	0.153	0.748
FDR	0.898	0.892	0.962	0.958	0.031	**0.107**	**0.785**
Fairness Notion 2: AE	WACC		AUC		AE_Diff		AF
	Train	Test	Train	Test	Train	Test	Test
FullFT-Reg	1.000	**0.914**	1.000	0.968	0.000	0.049	0.865
LastFT	0.918	0.913	0.974	**0.971**	0.066	0.043	0.869
LastFT-RW	0.913	0.908	0.970	0.968	0.026	0.020	0.888
LastFT-Reg	0.907	0.904	0.969	0.964	0.016	0.009	0.895
FDR	0.898	0.900	0.963	0.967	0.009	**0.003**	**0.897**
Fairness Notion 3: MMF	WACC		AUC		WA		AF
	Train	Test	Train	Test	Train	Test	Test
FullFT-Reg	1.000	0.910	1.000	0.969	1.000	0.393	1.303
LastFT	0.918	**0.913**	0.974	**0.971**	0.633	0.598	1.511
LastFT-RW	0.913	0.908	0.970	0.968	0.872	0.732	1.640
LastFT-Reg	0.896	0.888	0.960	0.955	0.879	0.717	1.605
FDR	0.902	0.898	0.964	0.962	0.868	**0.803**	**1.701**

Figure 1: Overall performance on CelebA dataset with different fairness constraints

References

[1] V. Cherepanova, V. Nanda, M. Goldblum, J. P. Dickerson, and T. Goldstein, **"Technical challenges for training fair neural networks,"** *arXiv preprint arXiv:2102.06764*, 2021.

[2] P. Kirichenko, P. Izmailov, and A. G. Wilson, **"Last layer re-training is sufficient for robustness to spurious correlations,"** *arXiv preprint arXiv:2204.02937*, 2022.

End