Fair Learning aid by Amplifying Bias

Rui Hu

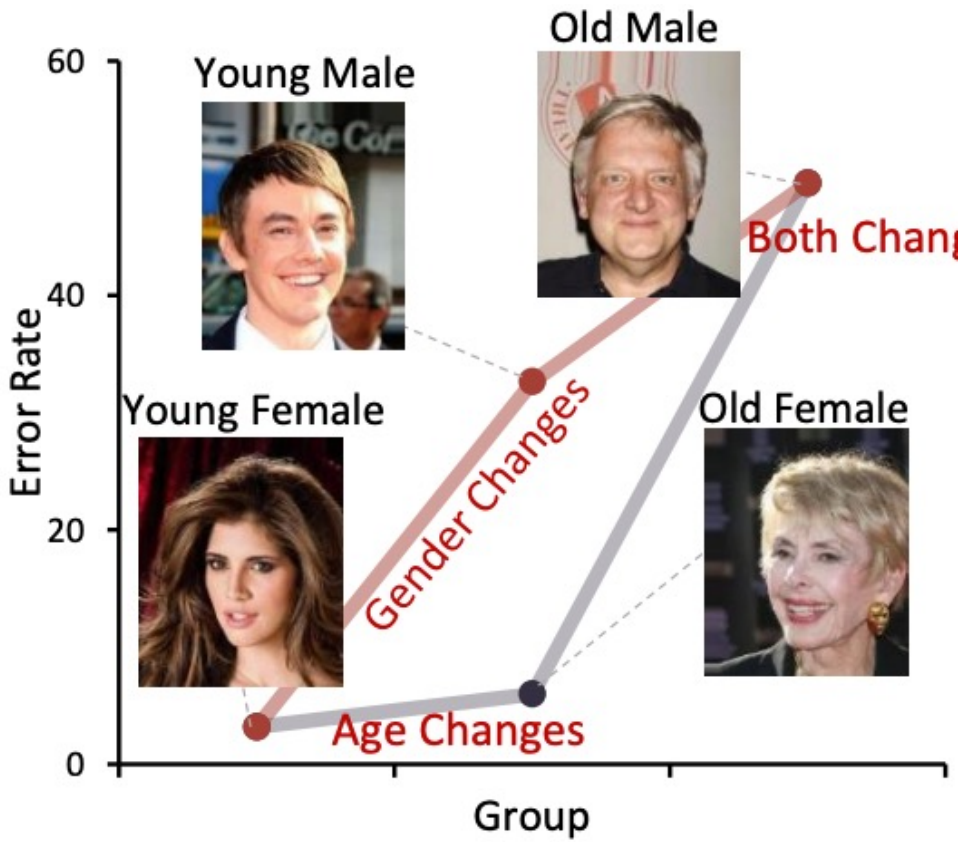
Motivation

多偏见

- 模型会学习多偏见，例如微笑分类任务，受到性别和年龄两种偏见属性影响
- 根据两种偏见属性可以将每个类别分为4个群体，当偏见属性发生改变时，群体准确率下降

无偏见标签

- 偏见标签获取成本大



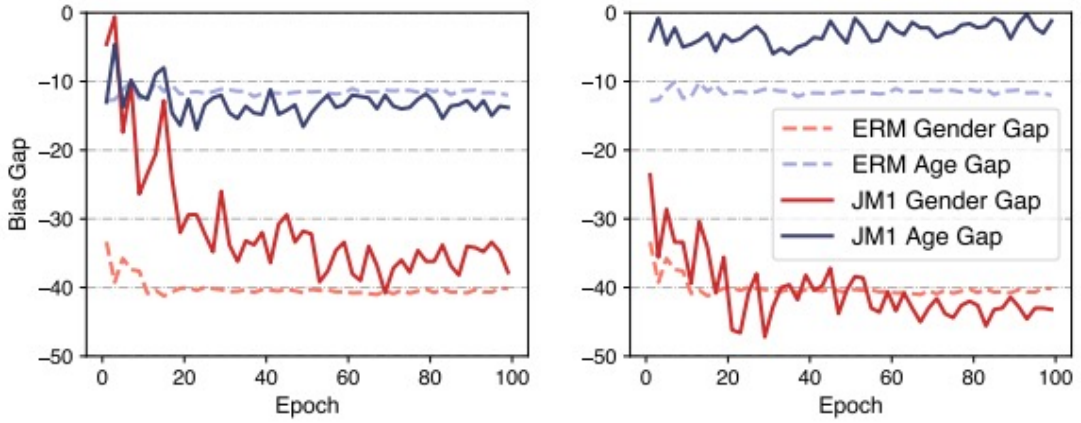
现有方法

多偏见

- 有监督方法可以直接适用多偏见场景，比如GroupDRO，但是需要偏见标签，并且由于核心群体（比如男性老人）的数量过少，去偏效果有限

无偏见标签（无监督）

- 现有的无监督方法大多分为两个步骤
  1，获取伪偏见标签（重要）
  2，利用伪偏见标签去偏



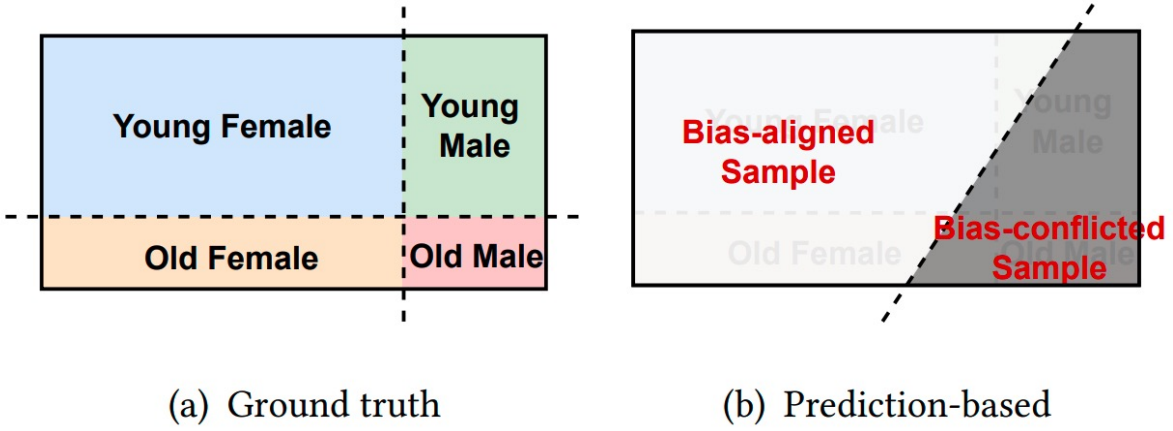
(a) Gender Label　　(b) Age Label

- 获取伪偏见标签主流有两种做法
  1，使用早停的ERM作为偏见模型，基于的是easy-to-learning假设
  2，使用GCE训练偏见模型，有工作提出GCE损失可以让模型放大偏见

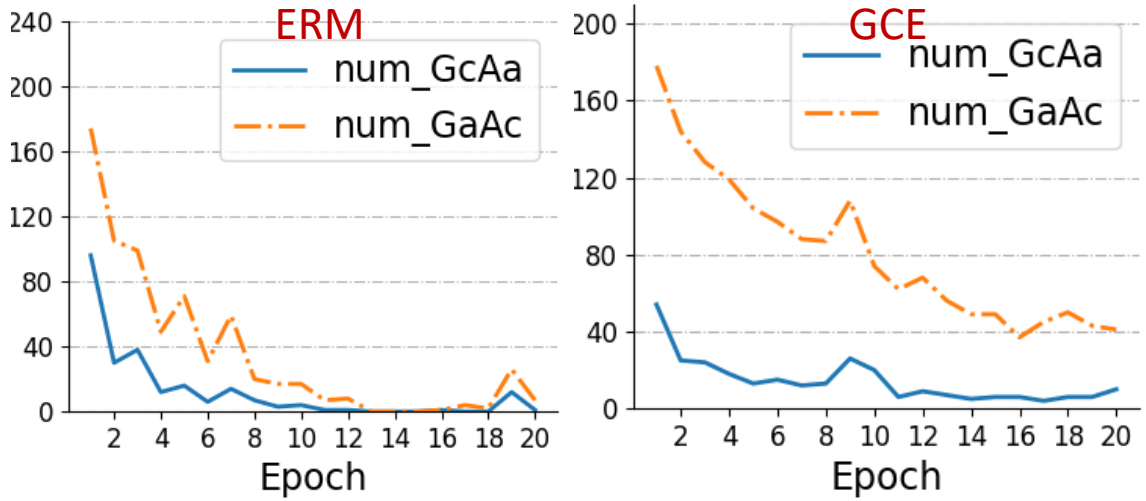
现有伪偏见标签获取方法的缺点

- 获取伪偏见标签主流有两种做法
  1，使用早停的ERM作为偏见模型，基于的是easy-to-learning假设
  2，使用GCE训练偏见模型，有工作提出GCE损失可以让模型放大偏见

- 问题：随着训练的进行，模型对训练集逐渐拟合，
  - 使用早停的话，会引入超参数，且不稳定
  - 偏见模型同步训练的话，找出的偏见冲突样本越来越少

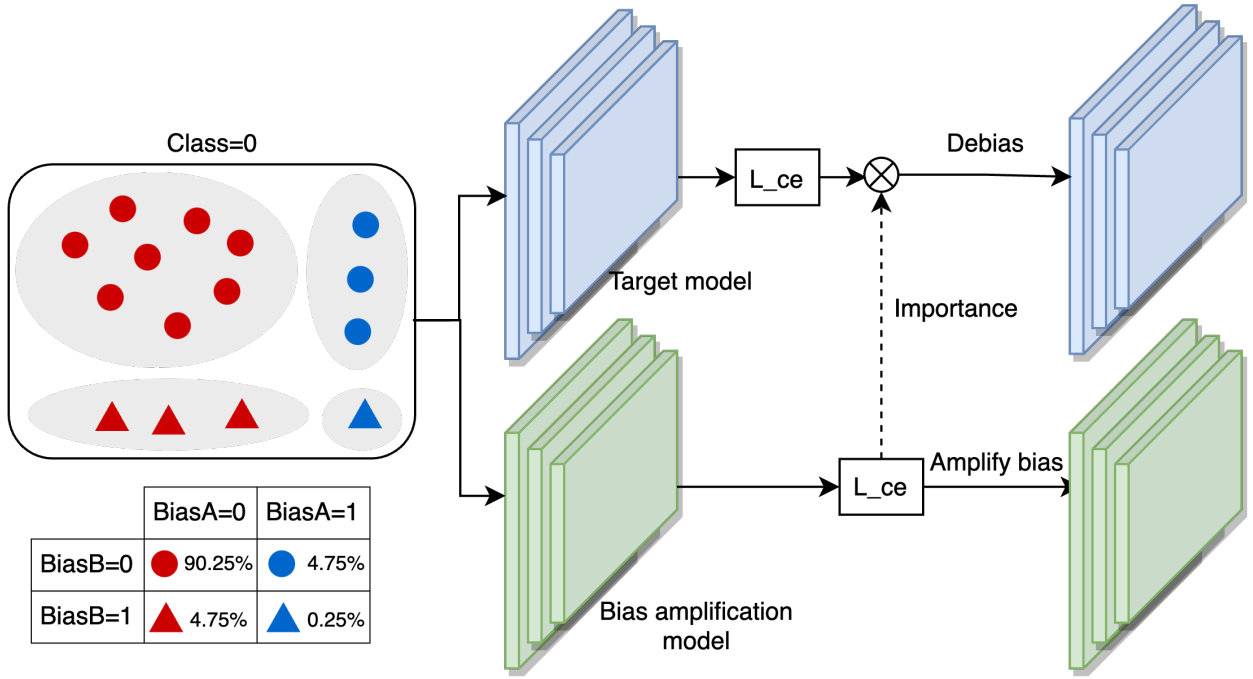


(a) Ground truth　　(b) Prediction-based



方法

- 基于重要程度的样本重加权
  - 从标签角度看：强bias的bias冲突样本是最重要的，弱bias的bias冲突样本是次重要的，偏见对齐样本重要性是最低的
  - 从预测角度看：每次epoch后错误的样本是比较重要的，因为错误样本很可能是偏见冲突的 -> 将错误样本降低权重，避免模型学习它



具体做法：每个样本初始权重为1，每个epoch后，bias模型分类错误样本权重减半

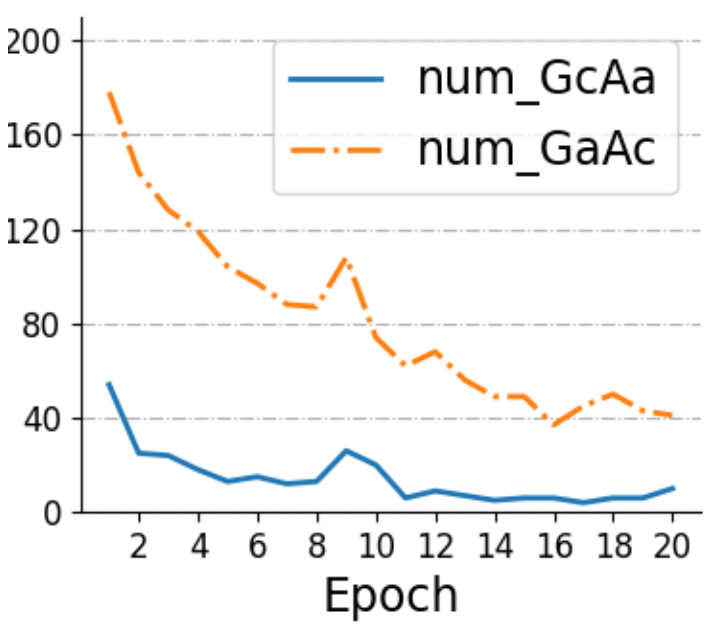
- 减少了偏见冲突样本的权重，使得不被过拟合
- 变相增强了偏见对齐样本里面easy-to-learn样本的权重，使得偏见模型更加集中学习偏见

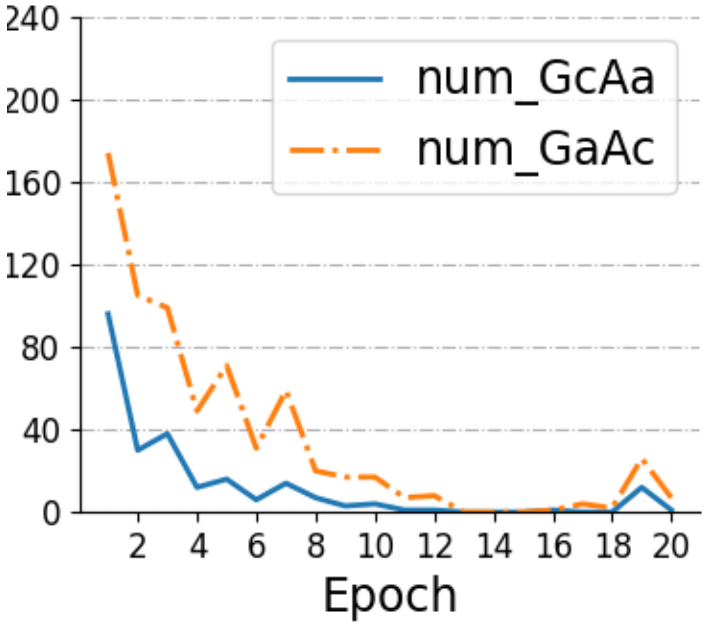
重要性 = 1 / 样本权重

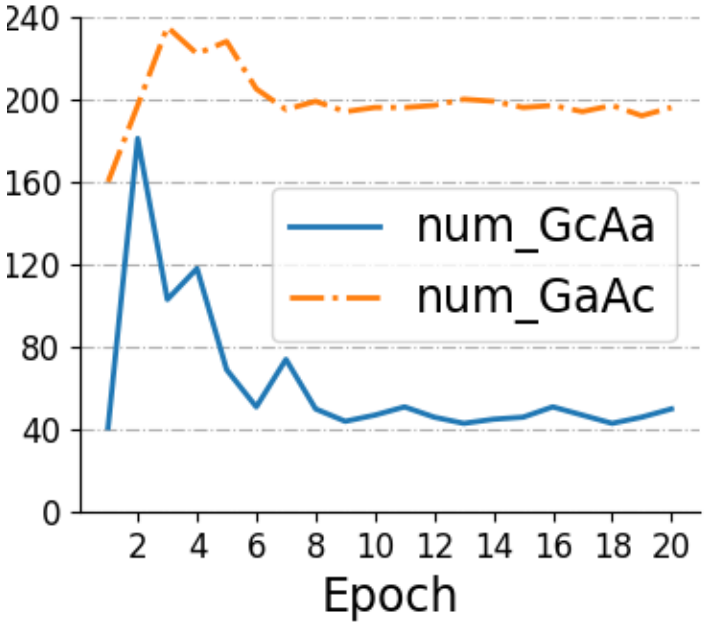
实验结果

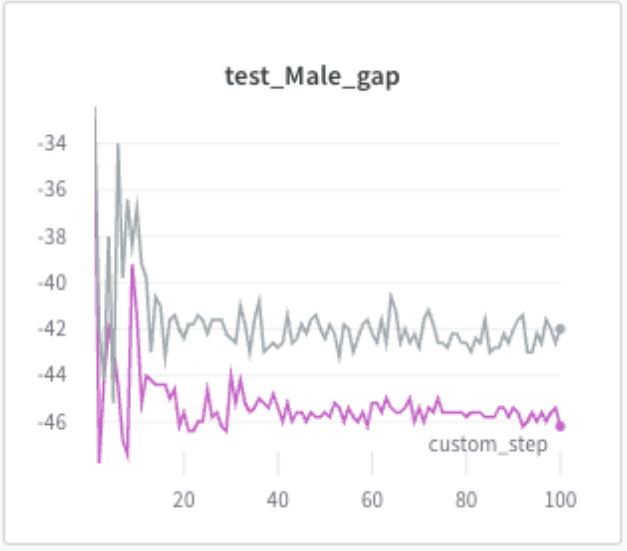
Method	Group avg acc	Worst group acc	Gender gap	Age gap	Avg bias
ERM	75.8	44.0	-38.4	-6.4	-22.4
JTT	76.6	39.2	-38.4	-10.0	-24.2
LfF	77.8	63.2	-21.6	-1.6	-11.6
JM1	66.5	50.4	-29.0	-0.6	-14.8
DebiAN	77.4	43.2	-39.2	-12.8	-26.0
BAR	78.1	52.8	-31.4	-11.0	-21.2
FaLA	77.1	63.2	-9.4	-3.0	-6.4

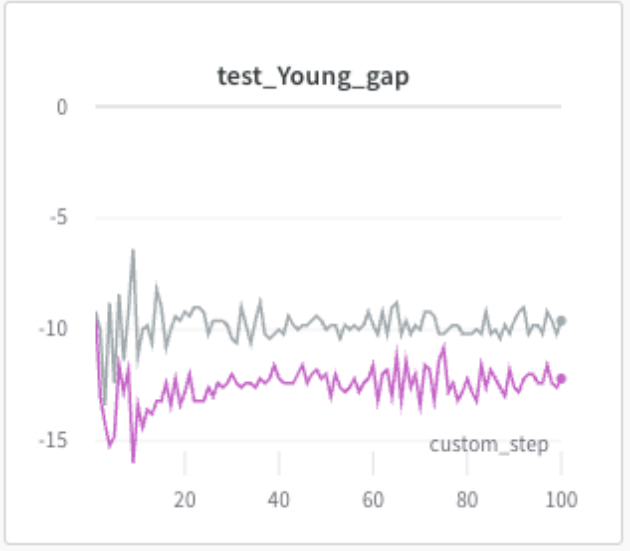
偏见放大模型效果


num_GcAa
num_GaAc
Epoch


num_GcAa
num_GaAc
Epoch


num_GcAa
num_GaAc
Epoch


test_Male_gap
custom_step


test_Young_gap
custom_step