Counterfactual Interpolation Augmentation (CIA): A Unifed Approach to Enhance Fairness and Explainability of DNN

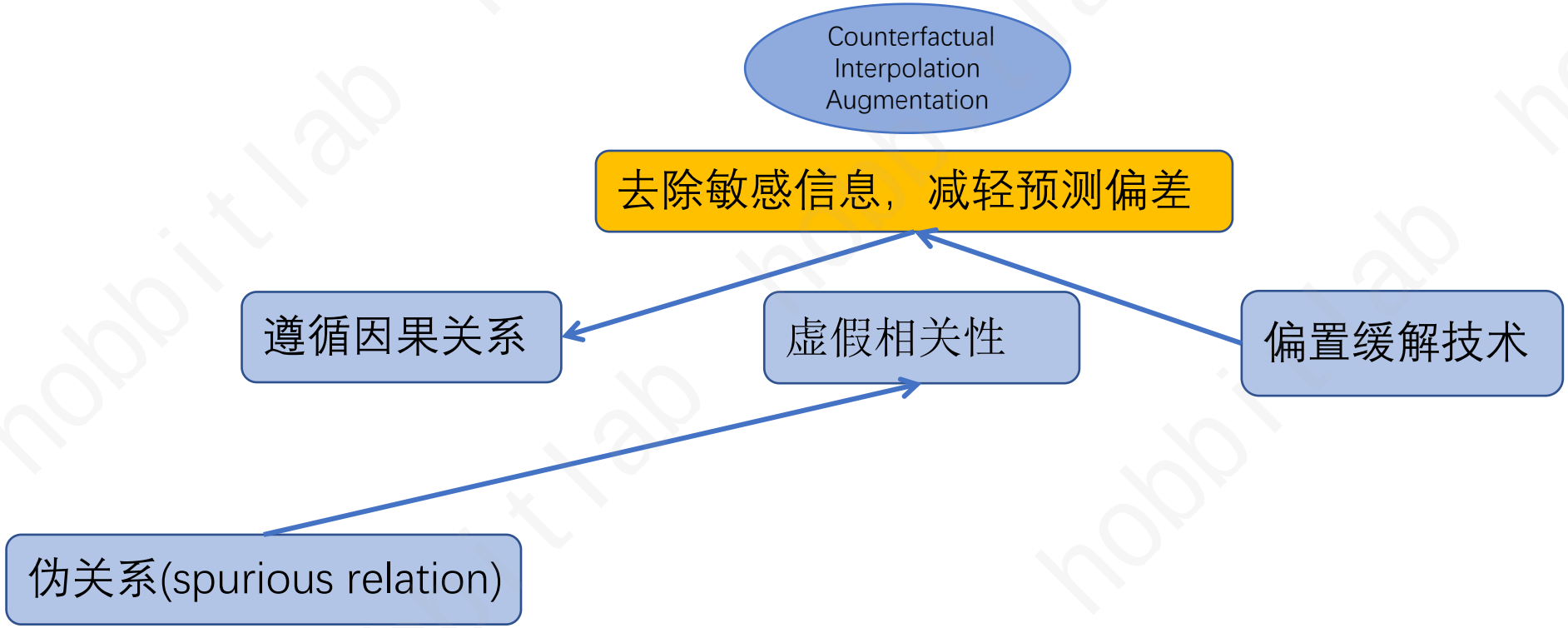
Huo Mingda

Jinan University, Guangzhou

May 11, 2023

研究内容：反事实（counterfactual）

利用因果理论中的反事实（counterfactual）框架来提高算法的稳定性和可解释性。



研究内容：公平性感知预测

解释方法：

$Argmax_E Q(E|Human, Data, Task)$

预处理：

通过数据增强来消除偏差和提高训练集的质量

$Argmax_{E,Model} Q(E|Model, Human, Data, Task)$

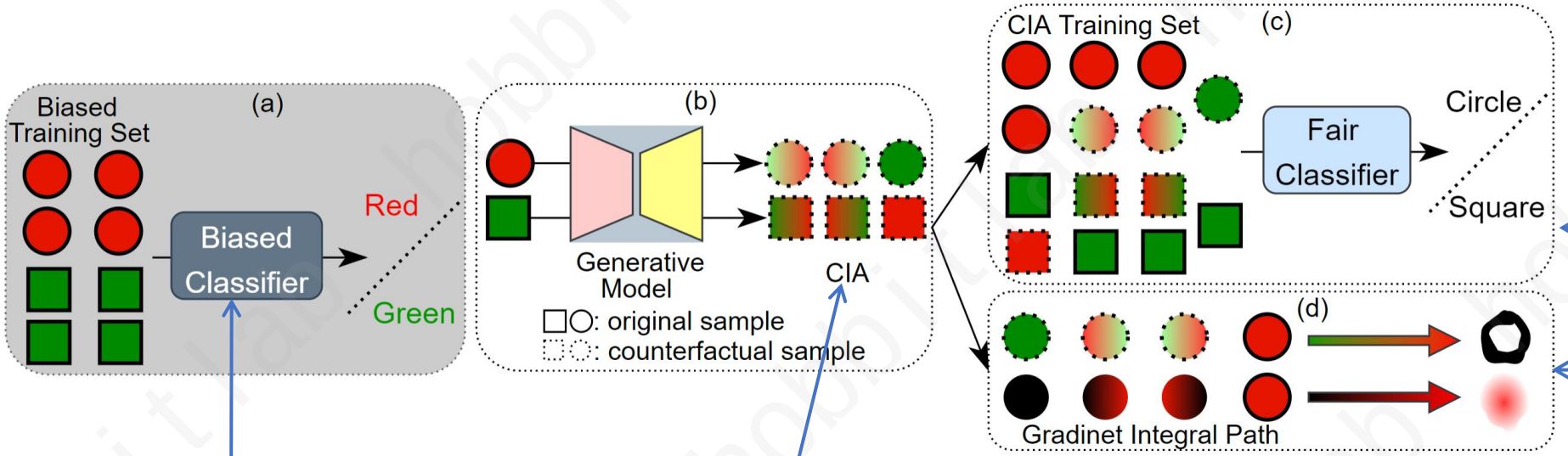
内处理：

在训练过程中从学习到的特征中去除敏感信息

后处理：

根据推理时的敏感属性来校准或修改预测

研究内容：公平性感知预测



有偏分类器：虚假相关预测

反事实数据增强：
减少偏差加强解释

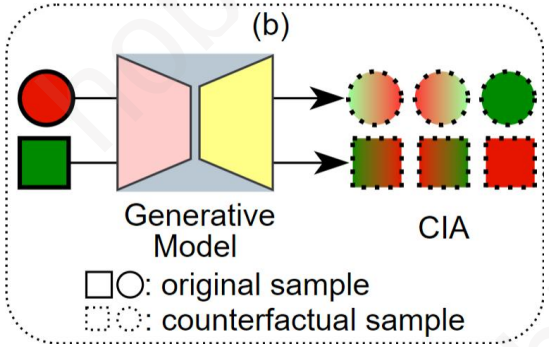
梯度完整路径：伪关系击穿

CIA公平分类器：产生解耦合能力

研究内容：伪关系解耦合

1. 数据增强策略

发现对抗样本，伪关系解耦合



将目标与训练数据中的敏感属性解除关联，来提高DNN的公平性

2. 模型解释方法

梯度积分

侧重于预期属性，不会被敏感属性分心

消除敏感属性的负面影响来产生更有意义的解释

研究内容：基于梯度解释的特征归因

积分梯度

$$\mathrm{IG}_i(x) = (x_i - x_i') \cdot \int_0^1 \frac{\partial F(x_i' + \alpha \cdot (x_i - x_i'))}{\partial x_i} d\alpha$$

黎曼近似求和得到线性路径上的积分梯度

偏导数得到网格灵敏度

基线直接影响归因质量

(1)最大距离基线

(2)模糊基线

(3)高斯基线

(4)统一基线

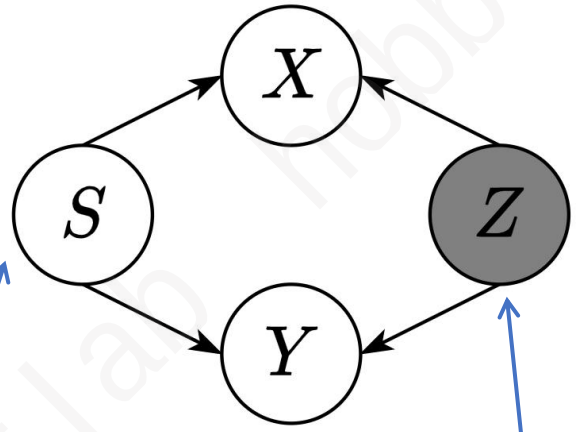
反事实插值路径积分梯度

可解释性

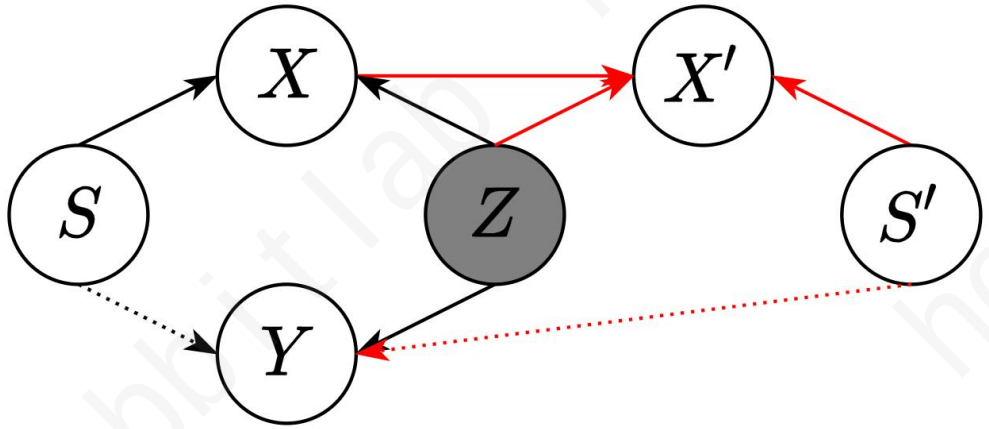
研究内容：反事实（counterfactual）构造

Huo Mingda, Jinan University

制造混淆因子S'产生反事实样例X'



(a) Original causal inference



(b) Counterfactual causal inference

敏感属性

潜在因素

混淆因子S'

研究对比：反事实插值构造

反事实因果推理

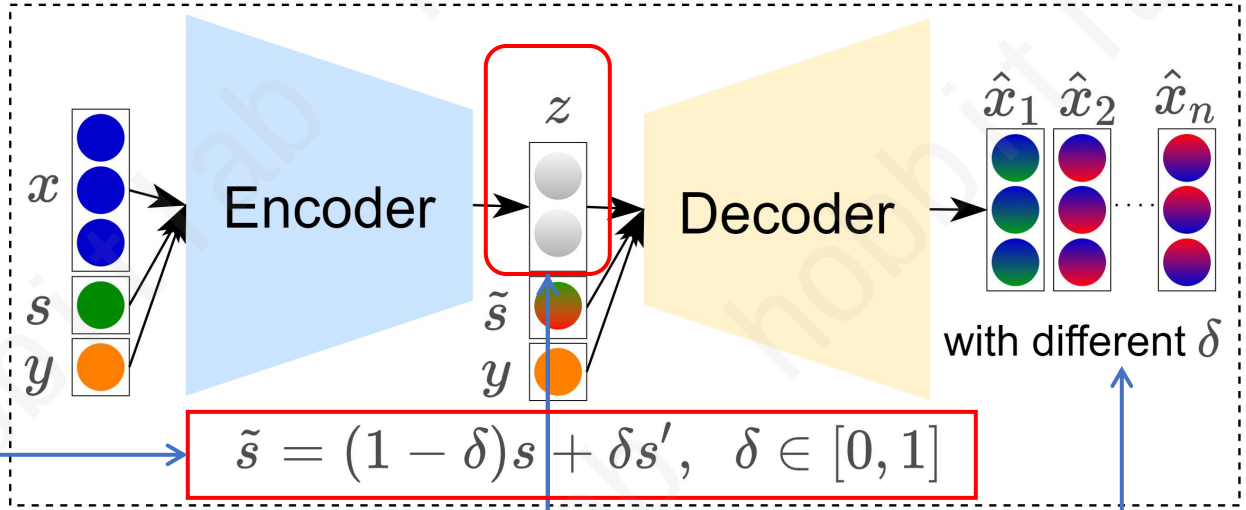
$$p(\hat{Y}_{S\leftarrow s} = y|X = x, Y = y, S = s)$$
$$=p(\hat{Y}_{S\leftarrow s'} = y|X = x, Y = y, S = s').$$

因果推理的视角来解决训练数据偏差问题

以事实目标y为条件，对模型的反事实公平提出标准

因果推理的视角来解决训练数据偏差问题

从事实例子*x*过渡到它的反事实例子*x'*



学习潜在特征

条件解码：
从特征Z和Y
到输入空间

插值路径， δ 决定插值数量

Huo Mingda, Jinan University

$$\mathrm{CGI}_i(x) = (x_i - x_i') \int_{\delta=0}^{1} \frac{\partial F(\gamma(\delta))}{\partial \gamma_i(\delta)} \frac{\partial \gamma_i(\delta)}{\partial \delta} d\delta.$$

$$\gamma(\alpha) = x' + \alpha(x - x')$$

路径积分的方向：从反事实样本x'过渡到输入x

通过CIA生成的反事实插值

实验对比：在MNIST、CELEBA数据集上的基线测试

Huo Mingda, Jinan University

LAFTR.

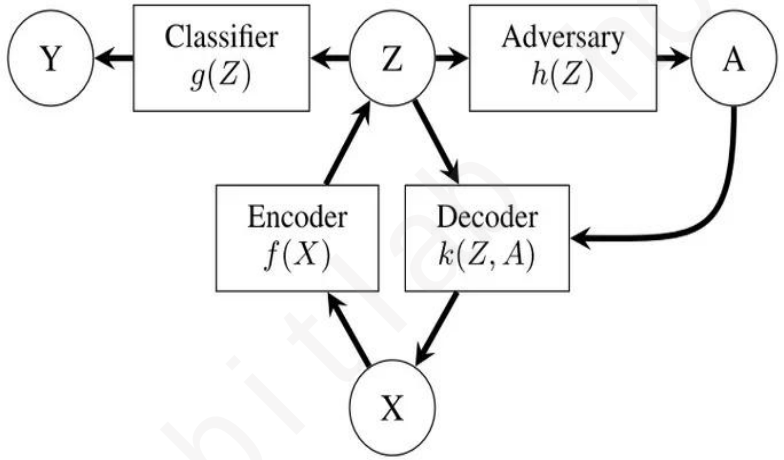


Figure 1. Model for learning adversarially fair representations. The variables are data X, latent representations Z, sensitive attributes A, and labels Y. The encoder f maps X (and possibly A - not shown) to Z, the decoder k reconstructs X from (Z, A), the classifier g predicts Y from Z, and the adversary h predicts A from Z (and possibly Y - not shown).

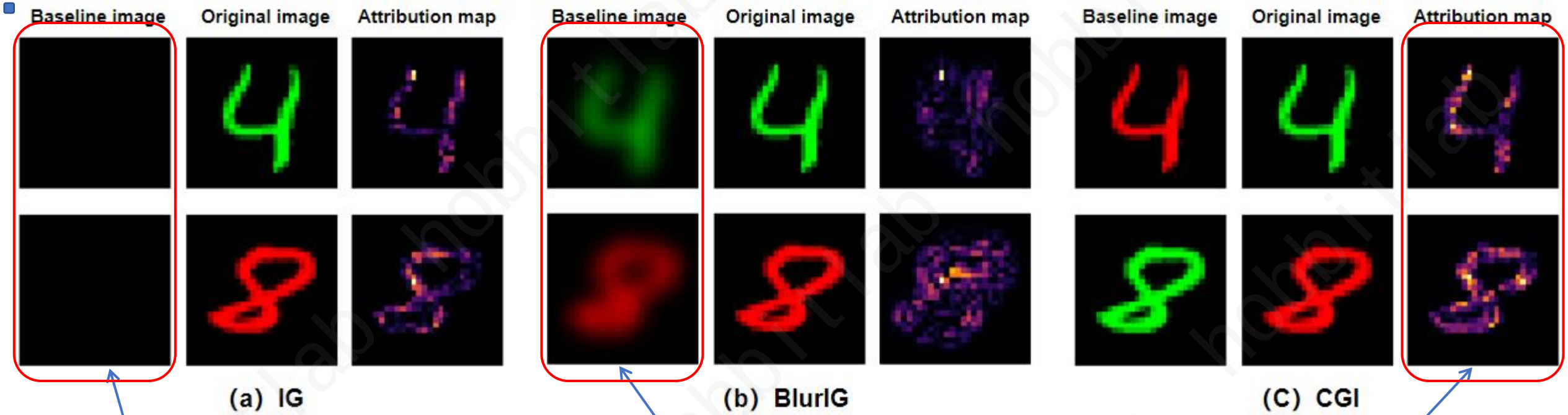
Method	Training Acc	Test Acc
Vanilla	79.48	18.08
LAFTR	74.14	75.22
PriorTraining	74.62	75.46
CIA-10	79.64	78.16
CIA-20	79.95	78.23
CIA-30	**79.97**	**78.69**

错误依赖颜色和数字的虚假相关性

CIA：样本的反事实插值越多准确度越高

Vanilla：训练过程中无法学习数字形状

实验对比

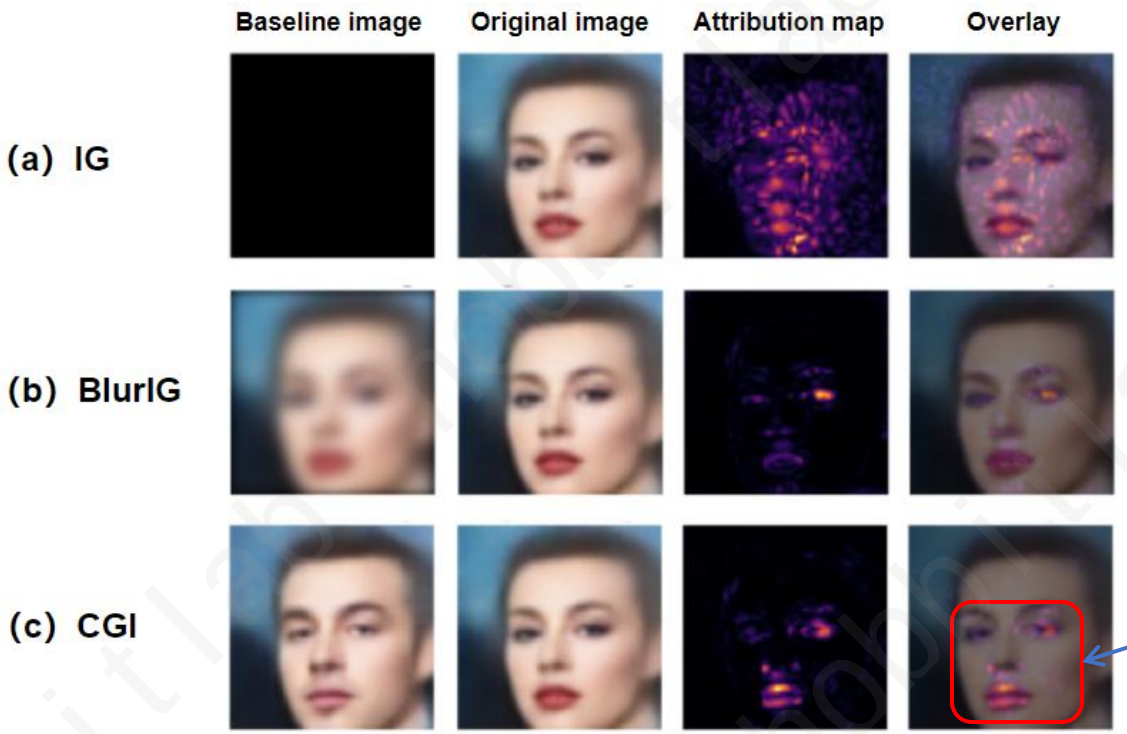


IG积分梯度算法使用黑色图像作为梯度积分的基线

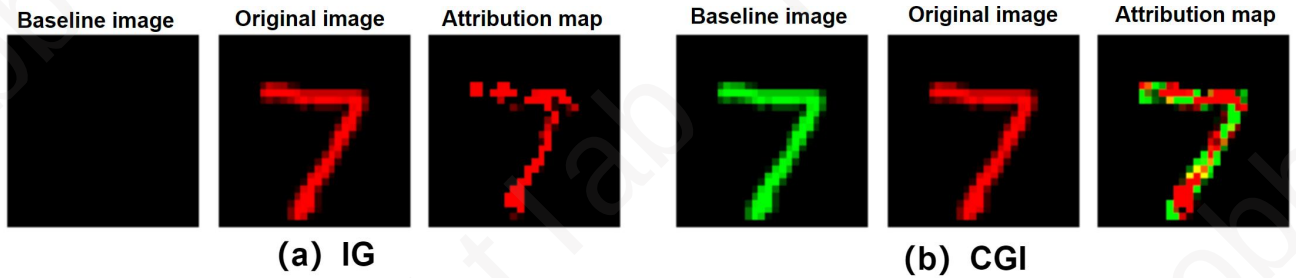
模糊IG:BlurIG通过连续模糊原始输入来消除路径积分

反事实梯度积分算法CGI具备清晰的归因热图

实验对比： 可解释公平性分析



归因属性更准确



可解释公平性

Thanks!