北京交通大學

打地鼠困境

A Whac-A-Mole Dilemma:
Shortcuts Come in Multiples Where Mitigating One Amplifies Others

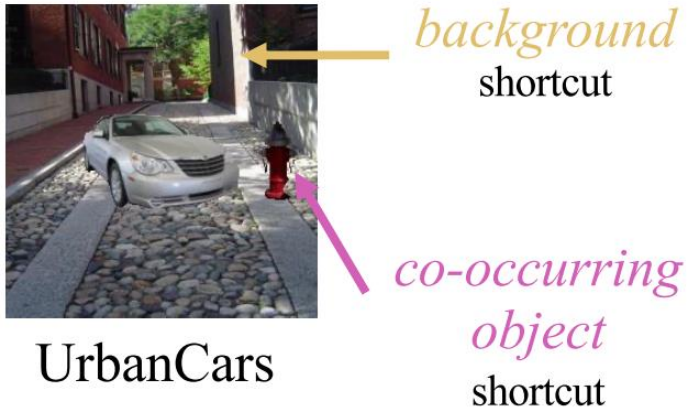
†Zhiheng Li[2] ‡Ivan Evtimov[1] Albert Gordo[1] Caner Hazirbas[1] Tal Hassner[1]
Cristian Canton Ferrer[1] Chenliang Xu[2] ‡Mark Ibrahim[1]
[1]Meta AI [2]University of Rochester
{ivanevtimov,agordo,hazirbas,thassner,ccanton,marksibrahim}@meta.com
{zhiheng.li,chenliang.xu}@rochester.edu

Rui Hu

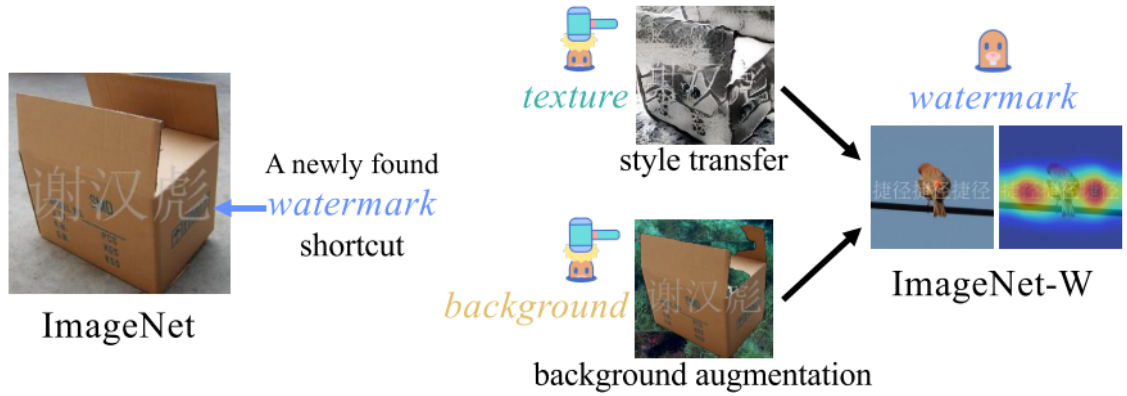
2023.1.11

背景

- 现有的去偏方法都基于一个脆弱的假设, 即数据中只存在single bias; 然而Real-world data会存在multi-bias, 现有方法在multi-bias设置下性能未知;



Target: the car's body type
Shortcut: background & co-occurring object



Target: object
Shortcut: texture & background & watermark

观察性实验

现有方法的性能如何?

- 现有方法面临打地鼠困境: 消除了一个bias会放大另一个bias;

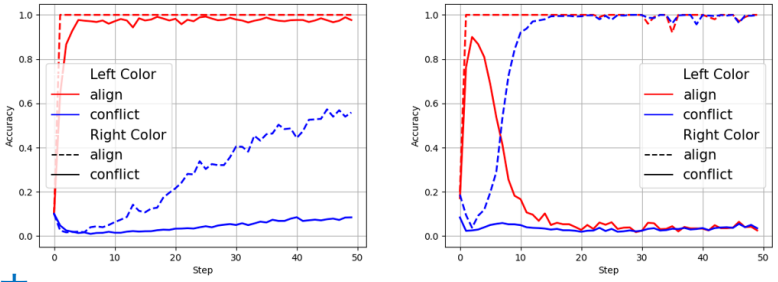
不使用偏见标签

	I.D. Acc	shortcut reliance		
		BG Gap ↑	CoObj Gap ↑	BG+CoObj Gap ↑
ERM	97.6	-15.3	-11.2	-69.2
Mixup	98.3	-12.6	-9.3	-61.8
CutMix	96.6	-45.0 (×2.94)	-4.8	-86.5
Cutout	97.8	-15.8 (×1.03)	-10.4	-71.4
AugMix	98.2	-10.3	-12.1 (×1.08)	-70.2
SD	97.3	-15.0	-3.6	-36.1
CF+F Aug	96.8	-16.0 (×1.04)	**+0.4**	-19.4
LfF	97.2	-11.6	-18.4 (×1.64)	-63.2
JTT (E=1)	95.9	-8.1	-13.3 (×1.18)	-40.1
EIIL (E=1)	95.5	-4.2	-24.7 (×2.21)	-44.9
JTT (E=2)	94.6	-23.3 (×1.52)	-5.3	-52.1
EIIL (E=2)	95.5	-21.5 (×1.40)	-6.8	-49.6
DebiAN	98.0	-14.9	-10.5	-69.0
LLE (ours)	96.7	**-2.1**	-2.7	**-5.9**

使用偏见标签

	shortcut label			shortcut reliance		
	Train	Val	I.D. Acc	BG Gap ↑	CoObj Gap ↑	BG+CoObj Gap ↑
ERM	✗	BG+CoObj	97.6	-15.3	-11.2	-69.2
gDRO	BG+CoObj	BG+CoObj	91.6	-10.9	-3.6	-16.4
DI	BG+CoObj	BG+CoObj	89.0	**-2.2**	-1.0	**+0.4**
SUBG	BG+CoObj	BG+CoObj	71.1	-4.7	**-0.3**	-6.3
DFR	BG+CoObj	BG+CoObj	89.7	-10.7	-6.9	-45.2
ERM	✗	BG	97.8	-14.6	-11.3	-68.5
gDRO	BG	BG	96.0	-4.2	-26.9 (×2.39)	-56.5
DI	BG	BG	94.7	+2.2	-27.0 (×2.40)	-25.2
SUBG	BG	BG	92.6	+1.3	-36.4 (×3.24)	-35.8
DFR	BG	BG	97.4	-9.8	-13.6 (×1.21)	-58.9
ERM	✗	CoObj	97.6	-15.4	-11.0	-68.8
gDRO	CoObj	CoObj	95.7	-31.4 (×2.03)	-0.5	-54.9
DI	CoObj	CoObj	94.2	-36.1 (×2.34)	+2.8	-35.8
SUBG	CoObj	CoObj	93.1	-60.2 (×3.90)	+2.5	-62.4
DFR	CoObj	CoObj	97.4	-19.1 (×1.24)	-8.6	-64.9

无监督去偏方法



(a)Vanilla (b)Iff

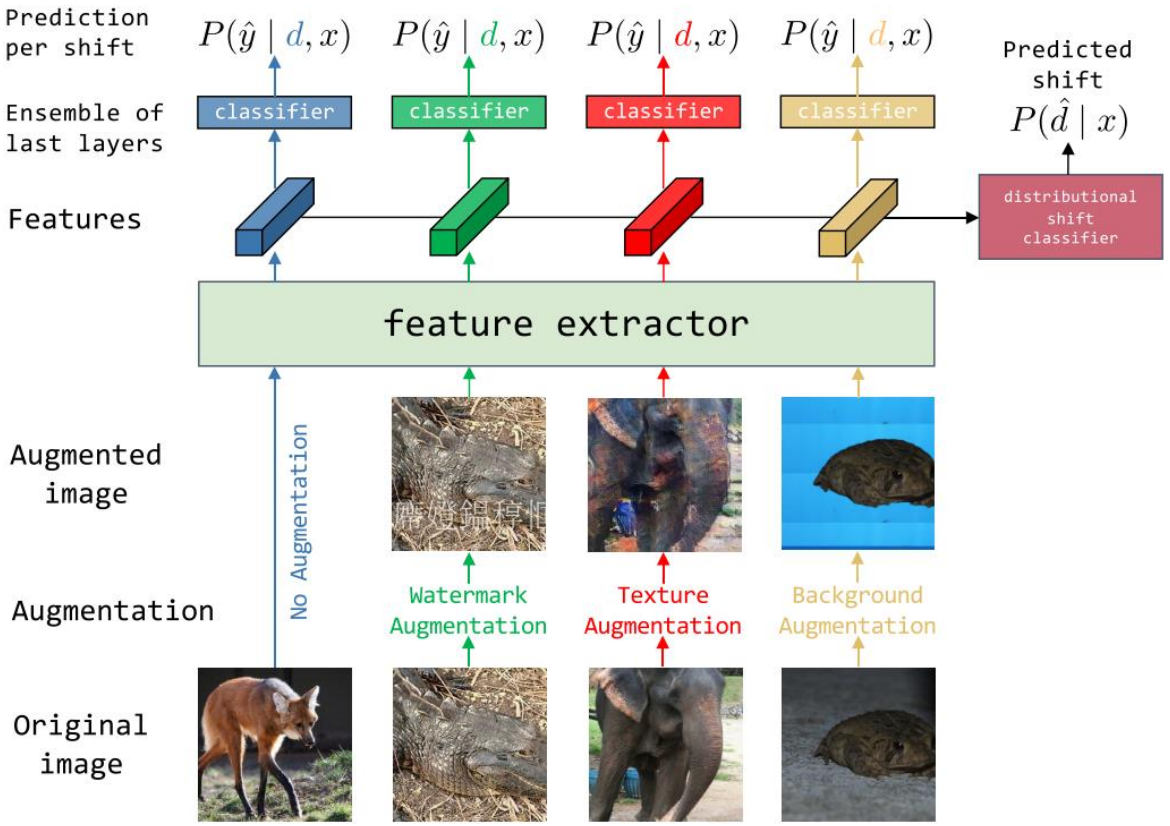
(c)Debian (d)DisEnt

		shortcut reliance				
		Watermark		Texture		Background
	IN-1k	IN-W Gap ↑	Carton Gap ↓	SIN Gap ↑	IN-R Gap ↑	IN-9 Gap ↑
arch: RG-32gf						
ERM	80.88	-14.15	+32	**-69.27**	-52.43	**-6.40**
SEER (FT,IG-1B)	83.35	**-6.50**	**+18**	-73.04 (×1.05)	**-50.42**	-7.14 (×1.11)
arch: ViT-B/32						
ERM	75.92	-8.71	+34	**-57.16**	-49.45	**-6.86**
Uniform Soup (FT,WIT)	79.96	-7.90	+24	-59.67 (×1.04)	**-27.51**	-7.78 (×1.13)
Greedy Soup (FT,WIT)	81.01	**-6.47**	**+16**	-59.61 (×1.04)	-30.01	-7.21 (×1.05)
arch: ViT-B/16						
ERM	81.07	-6.69	+26	-62.60	-50.36	-5.36
SWAG (LP,IG-3.6B)	81.89	-7.76 (×1.16)	+18	-67.33 (×1.08)	**-19.79**	-10.39 (×1.94)
SWAG (FT,IG-3.6B)	85.29	-5.43	+24	-66.99 (×1.07)	-29.55	-4.44
MoCov3 (LP)	76.65	-16.0 (×2.39)	+22	-63.36 (×1.01)	-56.86 (×1.12)	-7.80 (×1.45)
MAE (FT)	83.72	-4.60	+24	-65.20 (×1.04)	-47.10	-4.45
MAE+**LLE (ours)**	83.68	**-2.48**	**+6**	**-58.78**	-44.96	**-3.70**
arch: ViT-L/16 or 14						
ERM	79.65	-6.14	+34	-61.43	-53.17	-6.50
SWAG (LP,IG-3.6B)	85.13	-5.73	**+6**	-60.26	-10.17	-7.26 (×1.12)
SWAG (FT,IG-3.6B)	88.07	-3.16	+20	-63.45 (×1.03)	-12.29	-2.92
CLIP (zero-shot,WIT)	76.57	-4.47	+12	-61.27	**-6.26**	-3.68
CLIP (zero-shot,LAION)	72.77	-4.94	+12	-56.85	-8.43	-4.54
MAE (FT)	85.95	-4.36	+22	-62.48 (×1.02)	-36.46	-3.53
MAE+**LLE (ours)**	85.84	**-1.74**	+12	**-56.32**	-34.64	**-2.77**

预训练大模型

Method

- 面向的设置: 偏见类型已知, 偏见标签未知
- 方案: 针对性数据增强 + 分类器集成



Limitation

- 需要提前知道训练集中的bias类型;
- 并非所有bias都可以被数据增强.

引申想法

通用数据增强方法

- 数据增强方法应用广泛, 但是不同方法在偏见问题上表现不同, 例如CutMix可以缓解Co-obj偏见, 而Cutout可以缓解BG偏见, 原因是什么? 能否博采众长, 提出更好的数据增强方法;

无监督去偏见方法

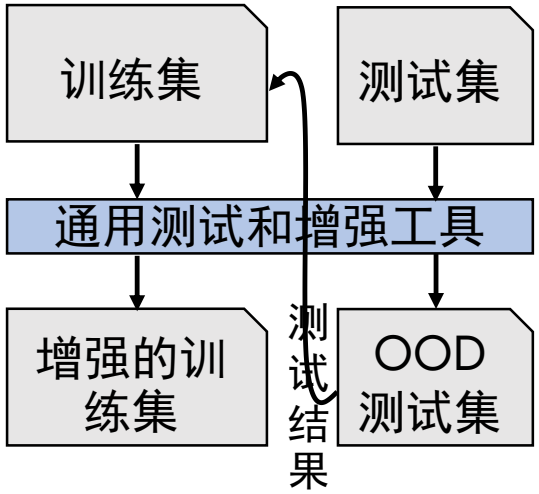
- 对不同无监督去偏见方法在多偏见问题上的差异进行分析比较;

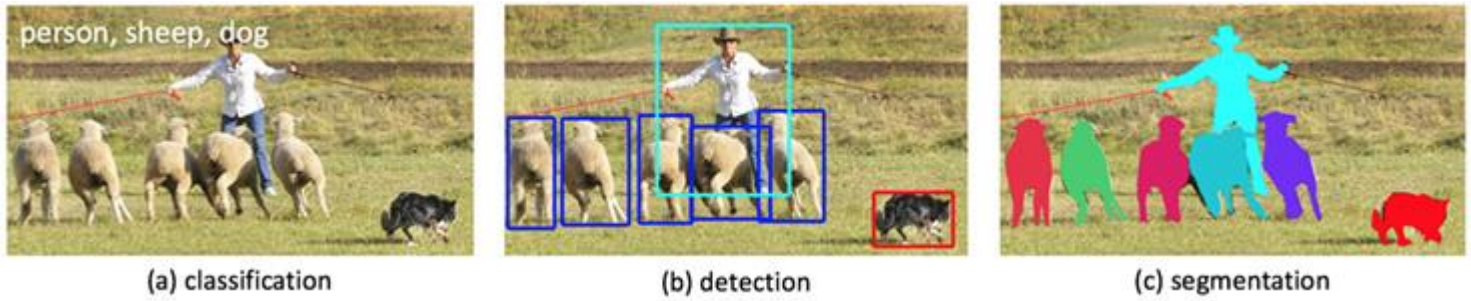
预训练方法避免ImageNet中的偏见

- ImageNet数据集常常被用于视觉预训练任务, 比如Moco和MAE. 同时ImageNet被证明存在watermark/texture/background等偏见, 现有预训练训练模型或多或少地受到这些bias的影响, 能否提出一种面向ImageNet的改进的预训练方法;

面向视觉模型的通用测试和增强工具

- 大多数数据集(ImageNet, COCO等)都存在一些常见的data bias, 比如背景bias, 纹理bias, 水印bias, 一般的测试集无法评估视觉模型在这些OOD情况的性能;
- 我们提出数据增强工具, 集成常见的bias数据增强, 方便模型开发人员生成OOD测试集;
- 同时, 在发现模型bias后, 模型开发人员可以选择增强训练集, 重新训练模型, 改进模型效果.





谢 谢!