Masked Images Are Counterfactual Samples for Robust Fine-tuning

Yao Xiao Ziyi Tang Pengxu Wei * Cong Liu Liang Lin
Sun Yat-sen University
{xiaoy99,tangzy27}@mail2.sysu.edu.cn {weipx3, liucong3}@mail.sysu.edu.cn
linliang@ieee.org

分享人：丁伯瑞

2024.05.09

CVPR2023

目录

目录

- 作者介绍

林倞

所属研究所、院系: 大数据与计算智能研究所

职称: 教授

E-mail: linlng@mail.sysu.edu.cn

个人主页: www.linliang.net

研究领域:

多模态感知与理解

— 场景语义解析；跨模态因果推断；跨领域泛化理解

— 自监督学习及预训练大模型；强化学习；认知及常识推理

多模态内容生成及交互

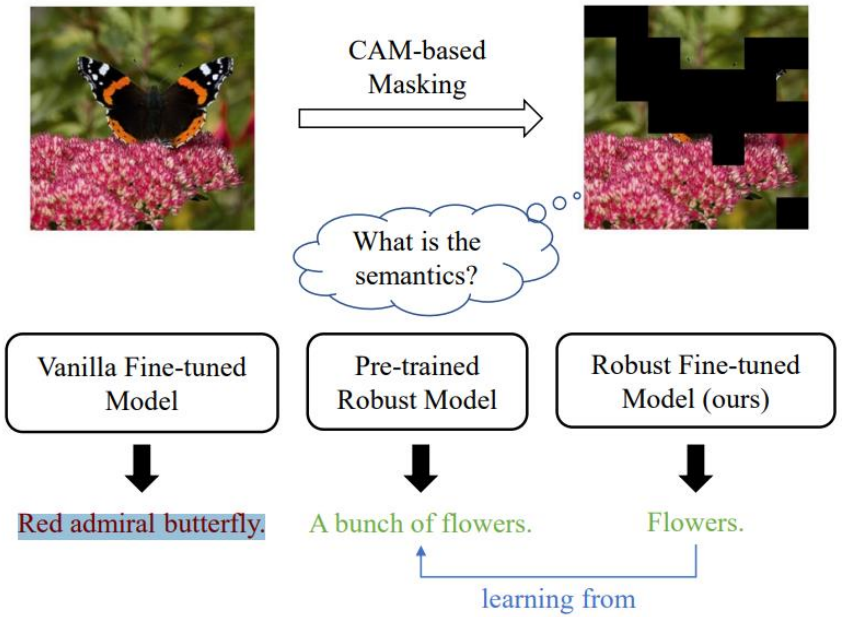
— 精准可控的图像视频生成；3D场景生成及编辑；数字人及元宇宙

— 具身智能与交互学习；自主机器人

目录

研究动机

- 存在问题：
  - 大模型对下游任务微调后，其会适应该数据集内部分布，导致out-of-distribution的效果会变差。



具体来说，下游场景训练图像的非语义表征和语义表征是高度纠缠的。例如，将 CLIP 模型迁移应用到蝴蝶这一下游场景时，许多训练图像中的蝴蝶都在花上。此时，微调训练可能使模型学习到依赖花这一非蝴蝶的语义表征来预测图像的语义。但是，这种相关性并不一定是真实的。

研究动机

- 现有的方法：
  - 限制预训练权重的扭曲：这种方法主要是涉及限制在微调过程中原始预训练模型权重的改变程度。其思想是保留初始化训练中学到的泛化能力。
  - 使用模型集成：另一种策略是使用多个模型或集成方法来保持鲁棒性。

但是并没有直接解决OOD鲁棒性问题。主要通过间接保持预训练模型的能力，而不是直接增强模型处理OOD数据的能力。

目录

方法介绍

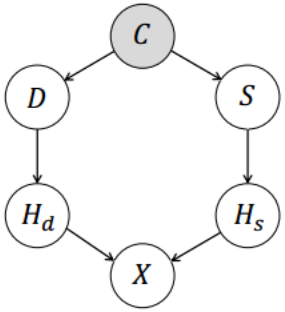
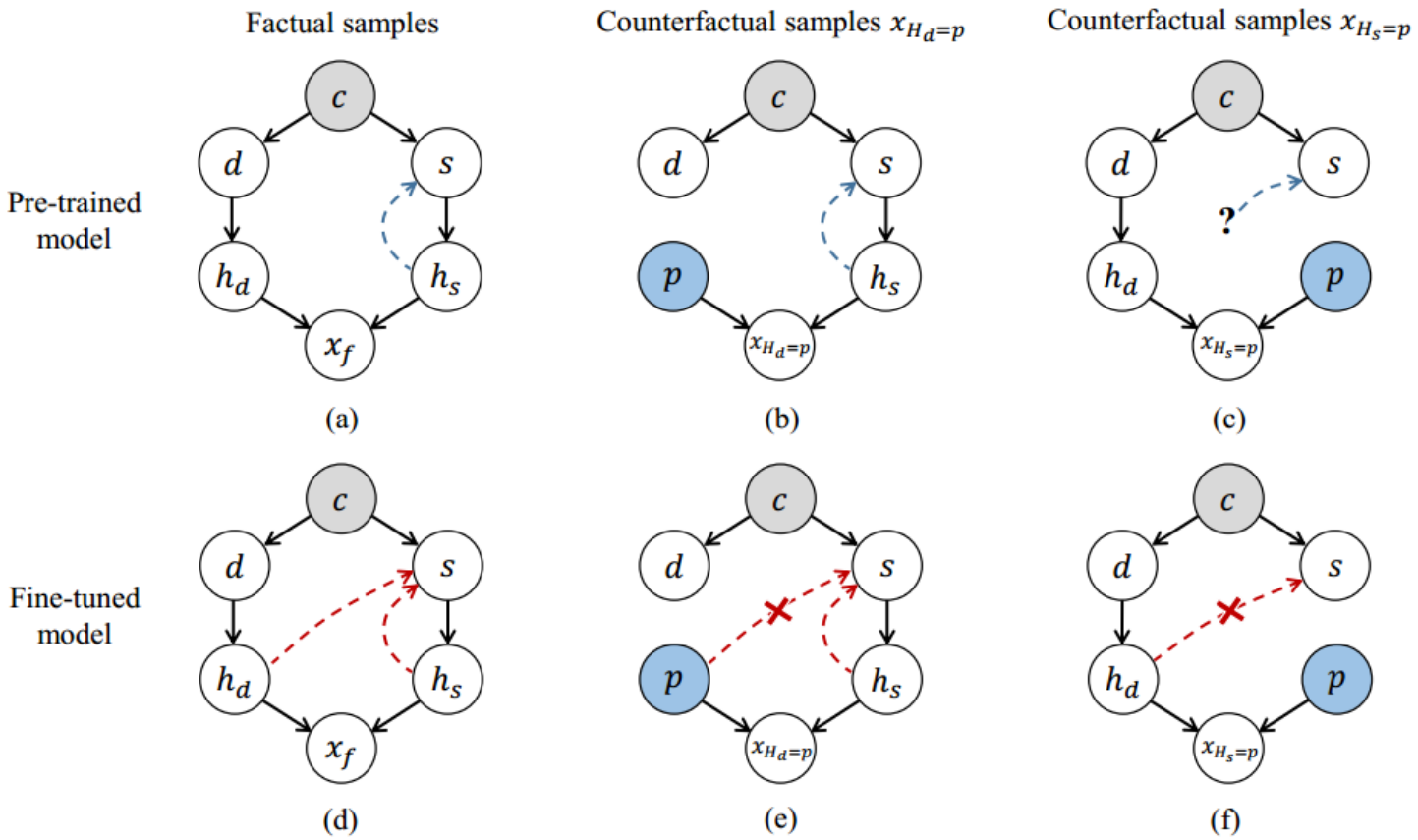


Figure 2. The causal graph of underlying object-centric image generation process across domains. C: confounder; D: domain; S: object semantics; H_d: (non-semantic) domain representation; H_s: semantic representation; X: image.

C：混杂因素
D：非语义因素
S：语义因素
H_d：图像非语义表征
H_s：图像语义表征

如图2所示，在Hd和Hs之间存在一条后门路径，即Hd←D←C→S→Hs，使得两者之间存在伪相关。从单一来源或环境中为下游任务收集数据可能导致语义部分Hs和领域相关部分Hd之间存在强烈的虚假相关性。换句话说，可能存在强烈的选择偏差。

方法介绍


Factual samples
Counterfactual samples $x_{H_d=p}$
Counterfactual samples $x_{H_s=p}$
Pre-trained model
Fine-tuned model
c
d
s
h_d
h_s
x_f
p
$x_{H_d=p}$
$x_{H_s=p}$
?
(a)
(b)
(c)
(d)
(e)
(f)

方法介绍

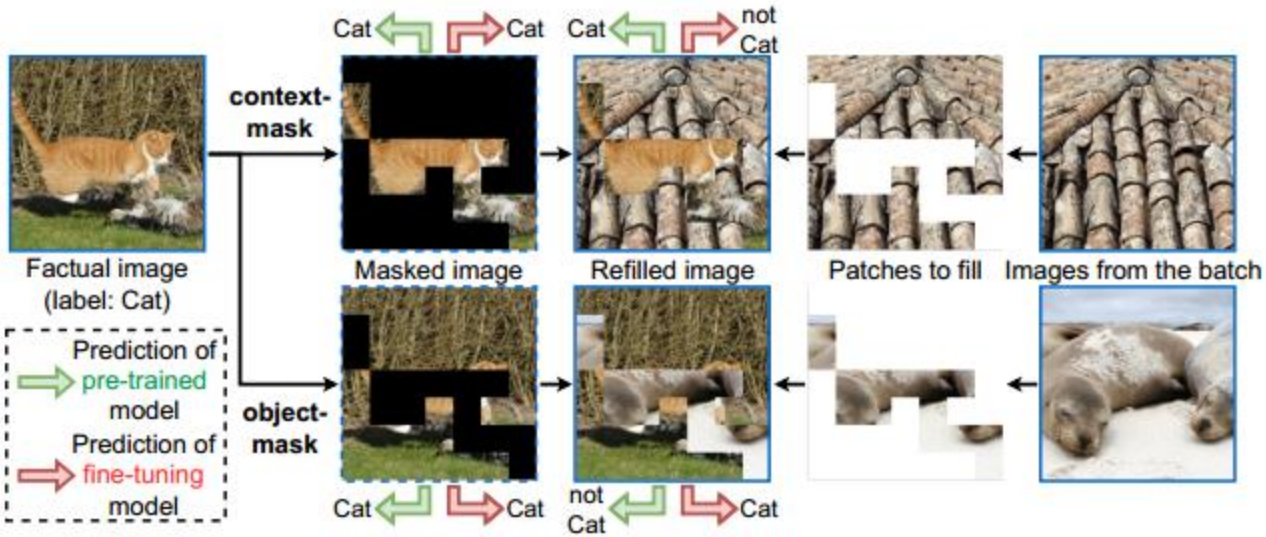


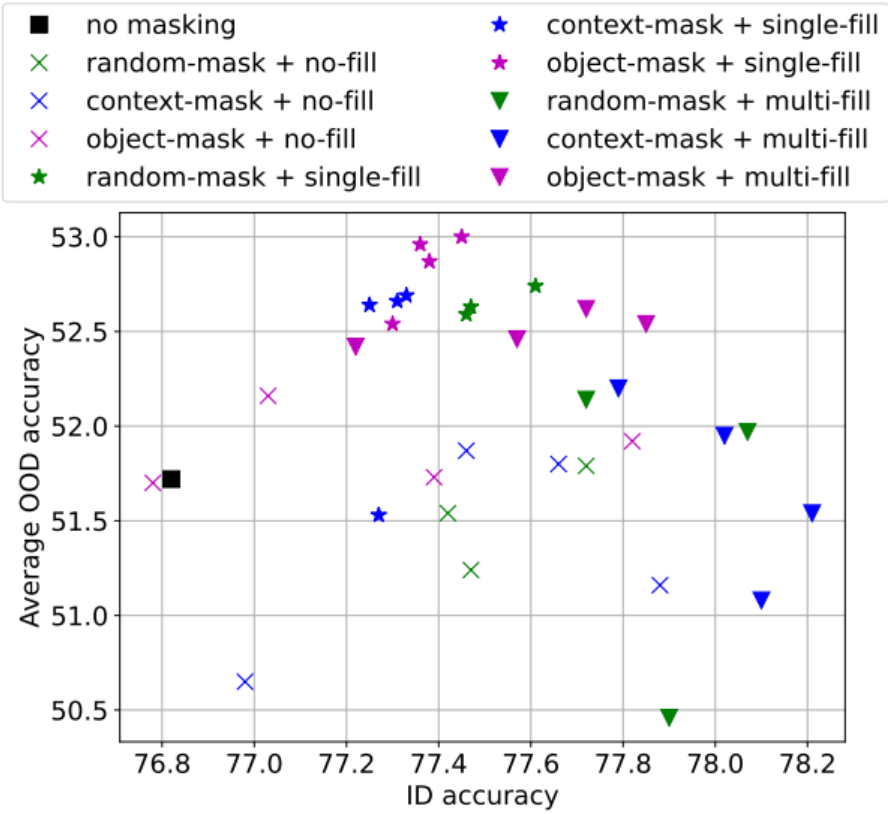
Figure 4. Illustration of the mechanism of masking and refilling.

直接用样本的标签与其反事实样本进行训练是不合适的，因为原始的语义信息可能会被扭曲。由于预训练模型具有捕获语义信息的强大能力，其图像级特征表示通常包含丰富的语义信息。

$$\mathcal{L} = \mathcal{L}_{\text{CE}}\big(g(f(x)), y\big) + \beta \mathcal{L}_{\text{MSE}}\big(\hat{f}(x_{cf}), f(x_{cf})\big),$$

- 作者介绍
- 背景介绍
- 研究动机
- 方法介绍
- 实验结果

实验结果



(1)大多数掩蔽和填充策略的组合比无掩蔽基线获得了更好的ID-OOD权衡，这表明其方法中图像掩蔽的有效性。(2)用其他图像的补丁填充被蒙住的图像(即，单填充或多填充)比单独删除被蒙住的补丁要好。(3)对比两种填充策略，单次填充总体上具有更好的OOD精度，而多次填充具有更好的ID精度。(4)对比掩蔽策略，object-mask在OOD准确率上普遍优于random-mask和context-mask。

实验结果

Method	Masking	Refilling	IN	IN-V2	IN-R	IN-Sketch	ObjectNet	IN-A	OOD avg.
Zero-shot [34]	/	/	63.4	55.9	69.3	42.3	44.5	31.4	48.7
Vanilla fine-tuning	/	/	75.9	64.7	57.0	39.8	39.5	20.0	44.2
Ours	no masking	/	76.9	66.5	69.2	45.6	45.3	29.8	51.3
Ours	random-mask	no-fill	77.5	66.9	66.4	45.7	46.5	30.8	51.2
	context-mask		77.8	67.4	66.7	45.6	45.9	30.0	51.1
	object-mask		77.7	67.1	67.6	46.2	46.8	31.5	51.9
Ours	random-mask	single-fill	77.6	67.1	69.0	46.4	47.8	33.4	52.7
	context-mask		77.2	66.9	68.8	46.5	47.8	32.9	52.6
	object-mask		77.5	67.1	**69.7**	**46.9**	**48.0**	**33.8**	**53.1**
Ours	random-mask	multi-fill	78.0	67.4	67.4	46.1	46.7	31.9	51.9
	context-mask		**78.2**	67.4	66.5	45.5	45.9	30.0	51.1
	object-mask		77.9	**67.7**	68.1	46.6	47.5	33.0	52.6

总结：

这篇文章在一定程度上打开了预训练大模型从深度学习范式中继承的“黑盒子”，从因果角度去分析和解决大模型的可解释性和可控性。

Thank for your attention !