arXiv:2304.05402v1 [cs.CV] 11 Apr 2023

利用视觉关系提升对抗补丁的跨任务可迁移性

Tony Ma[1], Songze Li[2], *, Yisong Xiao[2],和 Shunchang Liu[2], 3, ✉[1]北京航空航天大学沈元荣誉学院[2]北京航空航天大学软件开发环境国家重点实验室[3]中关村实验室

{tonyyyma}@gmail.com, {20373043, xiaoyisong, liusc}@buaa.edu.cn

摘要

对抗样本的可迁移性是评估深度学习系统鲁棒性的关键方面，尤其是在黑盒场景中。尽管已有多种方法被提出以增强跨模型可迁移性，但对抗样本在不同任务间的可迁移性却鲜少受到关注。随着 *Visual ChatGPT* 等基础多任务人工智能系统的出现，这一问题变得日益突出，使得由单一任务生成的对抗样本的效用相对有限。此外，这些系统通常具备超越单纯识别类任务的推理功能。为弥补这一空白，我们提出一种新颖的、基于视觉关系的跨任务对抗补丁生成方法，称为***VRAP***，旨在评估各种视觉任务的鲁棒性，尤其是涉及视觉推理的任务，例如视觉问答和图像描述。*VRAP* 利用场景图，将基于目标识别的欺骗与基于谓词的关系消除相结合，从而破坏推理任务之间共享的视觉推理信息。大量实验表明，在多种视觉推理任务的黑盒可迁移性方面，*VRAP* 显著优于以往方法。

1. 引言

对抗样本是经过刻意扰动、以欺骗深度学习模型的输入样本[3, 7]。它们已成为评估机器学习系统鲁棒性的重要工具 [12, 21]。评估对抗样本有效性时，一个关键考量因素是其可迁移性，即它们跨不同模型泛化的能力。大量研究表明，对抗样本强大的黑盒可迁移性对于评估深度学习模型在真实场景[9, 14, 15]中的鲁棒性。

一直以来，大量研究致力于提升对抗样本在不同模型之间的可迁移性 [2, 8, 17], 然而，很少有工作关注其在不同任务之间的可迁移性。近年来，一些基础性的多任务视觉系统（例如Visual ChatGPT [18] 和 GPT-4 [1] ）相继出现。这些系统不再依赖单一的机器学习模型来处理特定任务，而是能够解决不同场景中的问题，从而降低了依赖单一任务的对抗样本的有效性。此外，随着人工智能的持续发展，推理任务预计将成为这些基础模型的主流任务。视觉推理任务涵盖广泛的任务范式，要求神经网络模型不仅能够识别图像中的物体，还能够理解物体之间的关系，并利用这些信息完成下游任务。因此，与简单的视觉识别任务相比，视觉推理任务提供了更多的先验信息，而将视觉表征与推理先验相结合，是实现稳健多任务集成的一种有前景的方法 [20]。然而，现有的跨任务攻击方法仅考虑对基础视觉识别任务的攻击 [10, 22], , 并未考虑可能存在的视觉推理信息，因此在攻击上述多任务模型时效果较差。

为解决这一问题，开发一种能够生成具有强跨任务可迁移性的对抗样本的方法至关重要，尤其是在跨推理任务迁移性方面。为此，我们提出了一种名为 VRAP 的新方法，即一种基于视觉关系的跨任务对抗补丁生成方法。VRAP 的核心思想是使威胁模型能够学习图像中蕴含的潜在关系信息。通过这种方式，我们旨在扰乱各种推理任务之间共享的视觉推理信息，从而有效提升跨任务可迁移性。VRAP 利用场景图 [19],

*实习生。 ✉ 通讯作者。

一种表示视觉场景中对象及其相互关系的数据结构，用于将基于目标识别的欺骗与基于谓词的关系消除相结合。具体而言，该方法利用目标检测来识别图像中的对象，然后创建一个表示这些对象之间关系的场景图。随后利用该图生成一个旨在扰乱对象之间关系的对抗补丁。基于此，我们可以生成具有较强跨任务可迁移性的对抗补丁。大量实验表明，在不同视觉推理任务之间的黑盒迁移性方面，VRAP 优于以往方法。

2. 预备知识与方法

近期研究表明，对抗攻击的可迁移性很大程度上受任务不变特征的影响 [13, 22]。本文旨在识别会显著影响模型性能的任务共享特征。具体而言，我们聚焦于视觉推理任务，并希望使攻击者能够学习图像中潜在的关系信息。为此，我们提出了一种生成对抗扰动的新方法 δ，该扰动被限制在局部补丁中，能够诱使视觉推理模型做出错误预测。

具体而言，给定一张干净图像 x、一个加性对抗补丁扰动 $\delta \in \mathbb{R}^z$以及一个位置掩码 $M\in \{0,1\}^n$，我们可以按如下方式生成对抗样本 x_{adv}：

$$x_{adv} = (1 - M) \odot x + M \odot \delta, \ \ s.t. \ \ f(x_{adv}) \neq y, \quad (1)$$

其中，$\odot$表示逐元素乘法，f 表示一个视觉推理模型，y 表示相应的真实标注。

为了更好地理解视觉推理任务共享的关系，我们引入了一个场景图生成模型 f_G。场景图通过谓词对物体及其属性之间的关系进行建模，从而以结构化形式表示视觉内容，例如汽车“在”桥上，或汽车“有”车轮。具体而言，场景图生成模型由两部分组成：目标检测模块和谓词分类模块。目标检测模块检测并定位图像中的物体，谓词分类模块则为检测到的物体之间的关系分配谓词。将对抗样本输入模型 f_G后，我们可以得到两个概率，$P_C \in \mathbb{R}^{n\times m_c}$ 和 $P_R \in \mathbb{R}^{n\times n-1\times m_r}$，分别表示物体分类概率和谓词分类概率。$n$ 为预测目标的数量，m_c 为目标标签的数量，m_r 为谓词标签的数量。

为了有效打破图像中物体之间的关系，我们提出采用关系消除损失 $\mathcal{L}_r$。该损失项通过降低所有物体对之间预测关系的概率，使对抗补丁能够捕获物体之间的关系特征。具体而言，关系消除损失可表示为：

$$\mathcal{L}_r = \sum_{\mu,\nu} \max P_R(x_{adv}|\mu,\nu), \quad (2)$$

其中，μ,ν 表示对象对。损失通过累加对象对之间每种关系最可能类型的得分来计算。该汇总得分代表整个关系图的关系相关总得分。通过最小化此损失，我们促使补丁扰乱对象之间的关系，从而提高推理信息成功逃避视觉推理模型识别的可能性。

除了破坏图像中物体之间的关系，我们还探索了使用目标检测欺骗损失 $\mathcal{L}_d$ 来误导目标检测和分类模型。这一损失项旨在从误导模型检测标签的角度改变模型的预测。具体而言，其形式可表示为：

$$\mathcal{L}_d = \sum_{i} \max P_C(x_{adv}|y_i^{f_G} = y_i^{GT}), \quad (3)$$

其中 $y_i^{f_G}$表示模型的预测标签，y_i^{GT}表示目标对象 i的真实标签。通过这样做，我们旨在欺骗模型，使其误认图像中的目标对象。

对抗扰动的生成通过结合两种损失来实现即，

$$\arg\min_{\delta} \mathcal{L}_d + \lambda\mathcal{L}_r, \quad (4)$$

其中 λ 用于控制各项的贡献。所采用的方法使用基于梯度的迭代算法来优化对抗补丁。在每次迭代中，首先在随机位置生成一个初始对抗补丁。随后进行一次前向传播，以获取边界框特征及相应的目标检测欺骗损失 $\mathcal{L}_d$。接着计算视觉关系图，并得到关系消除损失 $\mathcal{L}_r$。最后，使用反向传播算法更新对抗补丁，使其能够攻击视觉推理模型共享的关系信息。因此，生成的扰动在不同视觉推理任务之间具有很强的可迁移性。

3. 实验

数据集。 我们使用 VisualGenome数据集 [5], 这一拥有 108,077 张图像和 230 万个关系的庞大视觉关系数据集来生成补丁。此外，我们还使用VQAv2 数据集 [4] 和 COCO [6]数据集进行评估。

表1. 场景图生成任务上的攻击结果。R@K 和 mR@K 越低，攻击效果越强。

子任务	方法	R@20	R@50	R@100	mR@20	mR@50	mR@100
谓词分类	Raw	59.64	66.12	67.97	11.44	14.59	15.84
	DR	59.42	65.98	67.86	11.22	14.26	15.47
	VRAP	**59.30**	**65.94**	**67.83**	**11.18**	**14.25**	**15.44**
场景图分类	Raw	36.00	39.24	40.05	6.49	8.02	8.50
	DR	34.08	37.07	37.83	6.15	7.50	7.97
	VRAP	**33.63**	**36.61**	**37.35**	**6.02**	**7.41**	**7.85**
场景图检测	Raw	25.40	32.45	37.24	4.37	5.80	7.06
	DR	23.75	30.02	34.39	4.11	5.39	6.53
	VRAP	**21.96**	**28.61**	**32.98**	**3.69**	**5.02**	**6.09**

表2. 图像描述任务上的攻击结果。指标值越低，表示攻击越强。

方法	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Raw	83.75	69.32	54.97	42.60	31.33	61.53	146.49	25.48
DR	81.53	66.71	52.41	40.22	30.45	60.03	**138.16**	24.30
VRAP	**81.19**	**66.35**	**52.07**	**39.94**	**30.27**	**59.80**	138.62	**24.20**

目标任务和模型。 我们针对三种典型的视觉推理任务开展攻击，以验证我们的方法：场景图生成（SGG）、视觉问答（VQA）和图像描述（IC）。SGG包含3 个不同的子任务，包括谓词分类（PredCls）、场景图分类（SGCls）和场景图检测（SGDet）。这些子任务的输入各不相同。谓词分类使用真实边界框和物体标签作为输入，场景图分类仅使用真实边界框，而场景图检测不需要任何真实信息。对于 SGG，我们针对经典的因果 Neural-MOTIFS 模型 [11]开展攻击。对于 VQA 和图像描述，我们将 OFA-Base [16] 作为目标模型。它是一种统一的序列到序列预训练模型，统一了不同模态和任务，并在 COCO排行榜的图像描述任务中取得了当前最先进的性能。

评估指标。 对于场景图生成，我们按照 [11]使用 Recall@K（R@K）和 Mean Recall@K（mR@K）作为评估指标。K 分别取 20、50 和 100。对于图像描述，我们按照 [16]使用相应的评估指标。

基线方法。 据我们所知，我们是首个研究对抗样本在不同视觉推理任务间可迁移性的工作。本文仅与当前最先进的攻击方法 Dispersion Reduction（DR）[10], 进行比较，该方法旨在实现跨识别任务的迁移。未来我们将考虑更多基线方法。

实现细节。 由于时间消耗问题，在训练阶段，我们仅使用包含 5000 张图像的Visual Genome验证集。我们随机初始化一个 80×80 方形对抗补丁，并以批量大小1进行训练，每个周期进行 $T = 5$ 次迭代，攻击步长为 α 0.04，最多训练 2 个周期。补丁的位置和方向均随机选择，这使我们的对抗补丁能够攻击所有场景。对于我们的方法，我们将关系损失权重 λ 设置为 0.01。对于 DR，我们选择攻击特征的离-在骨干网络的第三个模块之后。我们的所有代码均使用 PyTorch 实现。我们在配备10GB内存的 NVIDIA GTX2080Ti GPU 上进行所有实验。

3.1. 场景图生成上的攻击结果

首先，我们针对场景图生成任务开展实验。我们使用场景图分类模型生成对抗补丁，并相应地对各个子任务进行攻击。如表 1所示，与 DR 相比，我们的方法在白盒和黑盒设置下均取得了更低的 R@K 和 mR@K，这意味着攻击能力更强。

3.2. 视觉问答和图像描述上的攻击结果

对于视觉问答和图像描述，我们直接将从场景图生成任务中生成的补丁粘贴到测试图像上，以进行黑盒评估。表2 显示，我们的方法在图像描述任务上几乎都能达到更低的准确率，从而带来显著更强的黑盒攻击可迁移性。我们还通过Hugging Face空间将图像描述和视觉问答的结果进行了可视化 [1]。如图1所示，可以看出，我们的对抗补丁成功误导了模型对关系的判断。

4. 结论

本文提出了一种新颖的跨任务对抗补丁生成方法，称为VRAP，用于评估各种视觉推理任务的鲁棒性。VRAP通过将基于目标识别的欺骗与基于谓词的关系消除相结合，利用场景图破坏这些任务的共享信息。实验结果表明，在黑盒迁移性方面，VRAP 优于以往的方法。未来的研究可以探索 VRAP 在其他视觉任务中的潜力及其在真实场景中的应用。

[1]https://huggingface.co/OFA-Sys

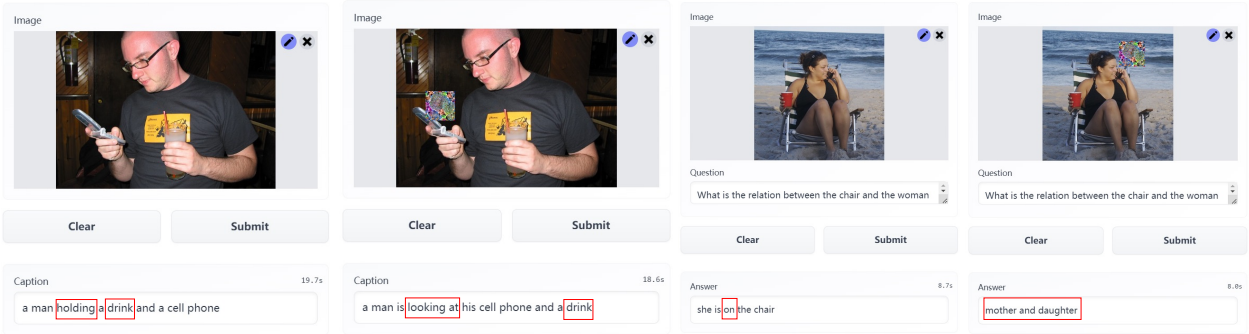

图1. 我们的补丁能够成功误导模型在图像描述和视觉问答中的关系预测。

参考文献

[1] OpenAI（2023）。Gpt-4 技术报告。2023年。1[2] Yinpeng Dong、Fangzhou Liao、Tianyu Pang、Hang Su、Jun Zhu、Xiaolin Hu 和 Jianguo Li。利用动量增强对抗攻击。发表于 CVPR，2018年。1[3] Ian J Goodfellow、Jonathon Shlens 和 Christian Szegedy。解释并利用对抗样本。arXiv 预印本 arXiv:1412.6572，2014年。1[4] Yash Goyal、Tejas Khot、Douglas Summers-Stay、Dhruv Batra 和 Devi Parikh。让视觉问答中的「视觉」发挥作用：提升图像理解在视觉问答中的作用。发表于 CVPR，2017年。2[5] Ranjay Krishna、Yuke Zhu、Oliver Groth、Justin Johnson、Kenji Hata、Joshua Kravitz、Stephanie Chen、Yannis Kalan- tidis、Li-Jia Li、David A Shamma 等。Visual Genome：利用众包密集图像标注连接语言与视觉。IJCV，2017年。2[6] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。发表于 ECCV，2014年。2[7] Aishan Liu、Xianglong Liu、Jiaxin Fan、Yuqing Ma、Anlan Zhang、Huiyuan Xie 和 Dacheng Tao。用于生成对抗补丁的感知敏感型生成对抗网络。发表于 AAAI，2019年。1[8] Aishan Liu、Jiakai Wang、Xianglong Liu、Bowen Cao、Chongzhi Zhang 和 Hang Yu。用于自动结账的基于偏差的通用对抗补丁攻击。发表于 ECCV，2020年。1[9] Shunchang Liu、Jiakai Wang、Aishan Liu、Yingwei Li、Yijie Gao、Xianglong Liu 和 Dacheng Tao。利用感知对抗补丁进行人群计数。发表于 ACM CCS，2022年。1[10] Yantao Lu、Yunhan Jia、Jianyu Wang、Bai Li、Weiheng Chai、Lawrence Carin 和 Senem Velipasalar。通过 Dispersion Reduction 增强对抗样本的跨任务黑盒迁移性。发表于 CVPR，2020年。1, 3[11] Kaihua Tang、Yulei Niu、Jianqiang Huang、Jiaxin Shi 和 Hanwang Zhang。从有偏训练中生成无偏场景图。发表于 CVPR，2020年。3[12] Shiyu Tang、Ruihao Gong、Yan Wang、Aishan Liu、Jiakai Wang、Xinyun Chen、Fengwei Yu、Xianglong Liu、Dawn Song、Alan Yuille 等。RobustArt：对架构设计与训练技术的鲁棒性进行基准测试。arXiv 预印本 arXiv:2109.05211，2021年。1

[13] Jiakai Wang、Aishan Liu、Xiao Bai 和 Xianglong Liu。利用感知偏差和注意偏差针对自动结账的通用对抗补丁攻击。IEEE TIP，2021年。2[14] Jiakai Wang、Aishan Liu、Zixin Yin、Shunchang Liu、Shiyu Tang 和 Xianglong Liu。双重注意力抑制攻击：在物理世界中生成对抗性伪装。发表于 CVPR，2021年。1[15] Jiakai Wang、Zixin Yin、Pengfei Hu、Aishan Liu、Renshuai Tao、Haotong Qin、Xianglong Liu 和 Dacheng Tao。用于物理世界鲁棒识别的防御补丁。发表于 CVPR，2022年。1[16] Peng Wang、An Yang、Rui Men、Junyang Lin、Shuai Bai、Zhikang Li、Jianxin Ma、Chang Zhou、Jingren Zhou 和 Hongxia Yang。OFA：通过简单的序列到序列学习框架统一架构、任务和模态。CoRR，abs/2202.03052，2022年。3[17] Yuxuan Wang、Jiakai Wang、Zixin Yin、Ruihao Gong、Jingyi Wang、Aishan Liu 和 Xianglong Liu。针对视觉变换器生成可迁移的对抗样本。发表于 ACM Multimedia，2022年。1[18] Chenfei Wu、Shengming Yin、Weizhen Qi、Xiaodong Wang、Zecheng Tang 和 Nan Duan。Visual ChatGPT：利用视觉基础模型进行对话、绘图和编辑。arXiv 预印本 arXiv:2303.04671，2023年。1[19] Pengfei Xu、Xiaojun Chang、Ling Guo、Po-Yao Huang、Xiaojiang Chen 和 Alexander G Hauptmann。场景图生成与应用综述。IEEE Trans. Neural Netw. Learn. Syst，1，2020年。1[20] Yuan Yang、James C Kerce 和 Faramarz Fekri。LogicDef：一种基于归纳式场景图推理的可解释对抗样本防御框架。发表于 AAAI，2022年。1[21] Chongzhi Zhang、Aishan Liu、Xianglong Liu、Yitao Xu、Hang Yu、Yuqing Ma 和 Tianlin Li。利用神经元敏感度解释并改进深度神经网络的对抗鲁棒性。IEEE TIP，2021年。1[22] Jin Zhang、Wenyu Peng、Ruxin Wang、Yu Lin、Wei Zhou 和 Ge Lan。通过距离度量攻击增强对抗样本的域不变迁移性。Mathematics，2022年。1，2