Boosting Cross-task Transferability of Adversarial Patches with Visual Relations

Tony Ma[1], Songze Li[2,*] Yisong Xiao[2], and Shunchang Liu[1,2,3,✉]
[1]Shen Yuan Honors College, Beihang University
[2]State Key Lab of Software Development Environmen, Beihang University
[3]Zhongguancun Laboratory
{tonyyyma}@gmail.com, {20373043, xiaoyisong, liusc}@buaa.edu.cn

Abstract

*The transferability of adversarial examples is a crucial aspect of evaluating the robustness of deep learning systems, particularly in black-box scenarios. Although several methods have been proposed to enhance cross-model transferability, little attention has been paid to the transferability of adversarial examples across different tasks. This issue has become increasingly relevant with the emergence of foundational multi-task AI systems such as Visual ChatGPT, rendering the utility of adversarial samples generated by a single task relatively limited. Furthermore, these systems often entail inferential functions beyond mere recognition-like tasks. To address this gap, we propose a novel **V**isual **R**elation-based cross-task **A**dversarial **P**atch generation method called **VRAP**, which aims to evaluate the robustness of various visual tasks, especially those involving visual reasoning, such as Visual Question Answering and Image Captioning. VRAP employs scene graphs to combine object recognition-based deception with predicate-based relations elimination, thereby disrupting the visual reasoning information shared among inferential tasks. Our extensive experiments demonstrate that VRAP significantly surpasses previous methods in terms of black-box transferability across diverse visual reasoning tasks.*

1. Introduction

Adversarial examples are input samples that have been intentionally perturbed to deceive deep learning models [3, 7]. They have become a vital tool in evaluating the robustness of machine learning systems [12, 21]. An essential consideration when assessing the effectiveness of adversarial examples is their transferability, which refers to their ability to generalize across different models. Numerous studies have demonstrated that strong black-box transferability of adversarial samples is critical for assessing the robustness of deep learning models in real-world settings [9, 14, 15].

There has been considerable research aimed at improving the transferability of adversarial examples across models [2, 8, 17], however, very little work has focused on their transferability across tasks. Recently, some foundational multi-task visual systems such as Visual ChatGPT [18] and GPT-4 [1] have emerged. These systems no longer rely on a single machine-learning model to handle specific tasks, but rather, can solve problems in different scenarios, leading to a reduction in the effectiveness of adversarial examples depending on a single task. Moreover, with the continued development of artificial intelligence, reasoning tasks are expected to become the mainstream tasks of these foundational models. Visual reasoning tasks encompass a wide range of task paradigms, requiring neural network models not only to recognize objects in images but also to understand the relationships between objects and use this information to complete downstream tasks. Thus, visual reasoning tasks provide more prior information than simple visual recognition tasks, and the combination of visual representation and reasoning prior is a promising approach for robust multi-task integration [20]. However, existing cross-task attack methods only consider attacks on basic visual recognition tasks [10, 22], and do not take into account the possible visual reasoning information, making them less effective when attacking the aforementioned multi-task models.

To address the problem, it is important to develop a method that is able to generate adversarial examples with strong cross-task transferability, particularly in the domain of cross-reasoning-task transferability. To achieve this, we propose a novel approach called VRAP, which is a visual relation-based cross-task adversarial patch generation method. The central idea behind VRAP is to enable the threatening model to learn the potential relation information present within the image. In doing so, we seek to disrupt the visual reasoning information shared across various inferential tasks, resulting in a positive effect on enhancing the cross-task transferability. VRAP utilizes scene graphs [19],

[*]Intern student. [✉] Corresponding author.

a data structure that represents the objects in a visual scene and their relationships with one another, to combine object recognition-based deception with predicate-based relations elimination. Specifically, the approach leverages object detection to identify objects within the image and then creates a scene graph that represents the relations between these objects. This graph is then used to generate an adversarial patch that is designed to disrupt the relations between the objects. Based on this, we can generate adversarial patches with strong cross-task transferability. Our extensive experiments demonstrate that VRAP outperforms previous methods in terms of black-box transferability across diverse visual reasoning tasks.

2. Preliminary and Method

Recent research suggests that the transferability of adversarial attacks is largely influenced by task-invariant characteristics [13, 22]. In this paper, we aim to identify task-shared characteristics that significantly impact model performance. Specifically, we focus on visual reasoning tasks and seek to enable the attacker to learn the potential relation information present within the image. To achieve this, we propose a novel approach for generating an adversarial perturbation δ constrained to a localized patch, which can fool visual reasoning models into making incorrect predictions.

Specifically, given a clean image x, an additive adversarial patch perturbation $\delta \in \mathbb{R}^z$, and a location mask $M \in \{0,1\}^n$, we can generate an adversarial example x_{adv} as following,

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

where $\odot$ is the element-wise multiplication, f is a visual reasoning model, and y is corresponding ground truth.

To obtain a better understanding of the relationships shared by visual reasoning tasks, we introduce a scene graph generation model f_G. Scene graphs provide a structured representation of the visual content by modeling the relationships between objects and their attributes using predicates, such as a car “on” a bridge or a car “has” wheels. Specifically, the scene graph generation model consists of two parts: an object detection module and a predicate classification module. The object detection module detects and localizes objects within the image, and the predicate classification module assigns predicates to the relationships between the detected objects. By feeding the adversarial example into the model f_G, we can obtain two probabilities, $P_C \in \mathbb{R}^{n \times m_c}$ and $P_R \in \mathbb{R}^{n \times n-1 \times m_r}$, representing the object classification probability and the predicate classification probability, respectively. n is the number of predicted targets, m_c is the number of target labels, and m_r is the number of predicate labels.

To effectively break the relationships between objects in an image, we propose the use of a relationship elimination loss $\mathcal{L}_r$. This loss term enables the adversarial patch to capture the relationship features between objects by reducing the probability of predicted relationships between all pairs of objects. Specifically, the relationship elimination loss can be formulated as:

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

where μ,ν represent object pairs. The loss is computed by adding up the scores of the most likely type of each relation between object pairs. This aggregated score represents the total relation-related score of the entire relation graph. By minimizing this loss, we encourage the patch to disrupt the relationships between objects, thereby increasing the chances of the reasoning information being able to successfully evade recognition by visual reasoning models.

In addition to breaking the relationships between objects in an image, we also explore the use of a target detection deception loss $\mathcal{L}_d$ to mislead object detection and classification models. This loss term is designed to alter the prediction of the model from the perspective of the misleading model detection labels. Specifically, it can be formulated as:

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

where $y_i^{f_G}$ represents the predicted label of the model and y_i^{GT} represents the real label of the target object i. By doing so, we aim to deceive the model into misidentifying the target objects in the image.

The generation of adversarial perturbations is achieved by utilizing a combination of two losses, *i.e.*,

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

where λ controls the contributions of each term. The approach employed utilizes a gradient-based iterative algorithm to optimize the adversarial patches. In each iteration, an initial adversarial patch is generated at a random position. Subsequently, a forward pass is conducted to obtain the bounding box feature and the corresponding target detection deception loss $\mathcal{L}_d$. The visual relation graph is then computed, and the relationship elimination loss $\mathcal{L}_r$ is obtained. Finally, the adversarial patch is updated using the back-propagation algorithm, enabling it to attack the shared relation information of visual reasoning models. Thus, the resulting perturbation exhibits a high degree of transferability across visual reasoning tasks.

3. Experiment

Datasets. We use the Visual Genome dataset [5], a giant visual relation dataset with 108,077 images and 2.3 million relationships, for patch generation. Besides, we also use VQAv2 dataset [4] and COCO [6] dataset for evaluation.

Table 1. Attacking Results on the SGG task. Lower R@K and mR@K means stronger attack.

Subtasks	Method	R@20	R@50	R@100	mR@20	mR@50	mR@100
PredCls	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**
SGCls	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**
SGDet	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**

Table 2. Attacking Results on the IC task. Lower values of the metrics means stronger attack

Method	Bleu_1	Bleu_2	Bleu_3	Bleu_4	METEOR	ROUGH_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**

Target tasks and models. We conduct attacks towards three typical visual reasoning tasks: Scene Graph Generation (SGG), Vision Questions Answering (VQA), and Image Captioning (IC) to verify our method. SGG consists of 3 different subtasks, including Predicate Classification (PredCls), Scene Graph Classification (SGCls) and Scene Graph Detection (SGDet). The inputs to these subtasks are different. PredCls uses ground-truth bounding boxes and object labels as inputs, SGCls uses only ground-truth bounding boxes, and SGDet does not require any ground-truth information. For SGG, we conduct attack on classical causal Neural-MOTIFS model [11]. For VQA and IC, we take the OFA-Base [16] as the target model. It is a unified sequence-to-sequence pretrained model that unifies modalities and tasks, which achieves the state-of-the-art performance at the COCO Leaderboard for image captioning task.

Evaluation metrics. For SGG, we use Recall@K (R@K) and Mean Recall@K (mR@K) as our metrics following [11]. The values of K are taken as 20, 50 and 100, respectively. For IC, we use the metrics following [16].

Baselines. To the best of our knowledge, we are the first work to study the transferability of adversarial examples across visual reasoning tasks. In this paper, we compare only with the state-of-the-art attack approach, Dispersion Reduction (DR) [10], aiming to transfer across recognition tasks. We will consider more baselines in future.

Implementation details. For patch generation, we only use the validation set of Visual Genome which contains 5000 images for the training stage due to the time consumption. We randomly initialize a 80×80 square adversarial patch and conduct training with batch size 1 by $T = 5$ iterations every epoch with an attack step size α of 0.04, and a maximum of 2 epochs. The position and orientation of the patch are randomly chosen, which makes our adversarial patches able to universally attack all scenes. For our method, we set the relation loss weight λ as 0.01. For DR, we choose to attack the dispersion of feature after the third block of backbone. All of our codes are implemented in PyTorch. We conduct all experiments on a NVIDIA GTX2080Ti GPU with 10GB Memory.

3.1. Attacking Results on SGG

First we conduct experiments for scene graph generation task. We generate adversarial patches using the SGCls model and further perform attacks on each subtask accordingly. As shown in Table 1, in contrast to DR, our method achieves lower R@K and mR@K on both white-box and black-box settings, which means stronger attacking ability.

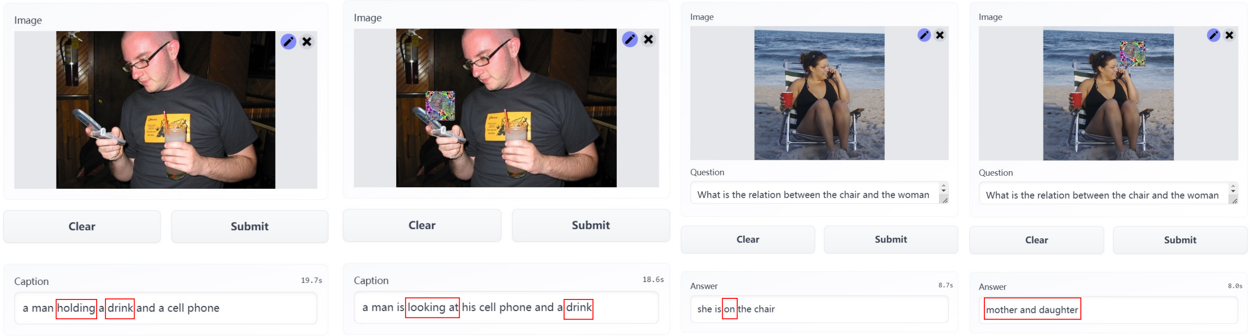
3.2. Attacking Results on VQA and IC

For VQA and IC, we directly pasted the patches generated from the SGG task on the test images for black-box evaluation. Table 2 shows that our method almost achieves lower accuracy for the image caption task, leading to a significantly stronger black-box attacking transferability. We also visualize the results of IC and VQA through the Hugging Face spaces [1]. As shown in Figure 1, it can be seen that our adversarial patch successfully misled the model's judgment towards the relationship.

4. Conclusion

In this paper, we present a novel cross-task adversarial patch generation method, named VRAP, to evaluate the robustness of various visual reasoning tasks. VRAP leverages scene graphs to disrupt the shared information of those tasks by combining object recognition-based deception with predicate-based relations elimination. The experiments conducted show that VRAP outperforms previous methods in terms of black-box transferability. Future research can explore the potential of VRAP in other visual tasks and its application in real-world scenarios.

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



Caption 19.7s

a man holding a drink and a cell phone

Caption 18.6s

a man is looking at his cell phone and a drink

Question

What is the relation between the chair and the woman

Answer 8.7s

she is on the chair

Question

What is the relation between the chair and the woman

Answer 8.0s

mother and daughter

Figure 1. Our patch can successfully mislead the model relation prediction in image caption and visual question answering.

References

[1] OpenAI (2023). Gpt-4 technical report. 2023. 1

[2] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 1

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 2

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 2

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[7] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *AAAI*, 2019. 1

[8] Aishan Liu, Jiakai Wang, Xianglong Liu, Bowen Cao, Chongzhi Zhang, and Hang Yu. Bias-based universal adversarial patch attack for automatic check-out. In *ECCV*, 2020. 1

[9] Shunchang Liu, Jiakai Wang, Aishan Liu, Yingwei Li, Yijie Gao, Xianglong Liu, and Dacheng Tao. Harnessing perceptual adversarial patches for crowd counting. In *ACM CCS*, 2022. 1

[10] Yantao Lu, Yunhan Jia, Jianyu Wang, Bai Li, Weiheng Chai, Lawrence Carin, and Senem Velipasalar. Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In *CVPR*, 2020. 1, 3

[11] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020. 3

[12] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan Yuille, et al. Robustart: Benchmarking robustness on architecture design and training techniques. *arXiv preprint arXiv:2109.05211*, 2021. 1

[13] Jiakai Wang, Aishan Liu, Xiao Bai, and Xianglong Liu. Universal adversarial patch attack for automatic checkout using perceptual and attentional bias. *IEEE TIP*, 2021. 2

[14] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *CVPR*, 2021. 1

[15] Jiakai Wang, Zixin Yin, Pengfei Hu, Aishan Liu, Renshuai Tao, Haotong Qin, Xianglong Liu, and Dacheng Tao. Defensive patches for robust recognition in the physical world. In *CVPR*, 2022. 1

[16] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022. 3

[17] Yuxuan Wang, Jiakai Wang, Zixin Yin, Ruihao Gong, Jingyi Wang, Aishan Liu, and Xianglong Liu. Generating transferable adversarial examples against vision transformers. In *ACM Multimedia*, 2022. 1

[18] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 1

[19] Pengfei Xu, Xiaojun Chang, Ling Guo, Po-Yao Huang, Xiaojiang Chen, and Alexander G Hauptmann. A survey of scene graph: Generation and application. *IEEE Trans. Neural Netw. Learn. Syst*, 1, 2020. 1

[20] Yuan Yang, James C Kerce, and Faramarz Fekri. Logicdef: An interpretable defense framework against adversarial examples via inductive scene graph reasoning. In *AAAI*, 2022. 1

[21] Chongzhi Zhang, Aishan Liu, Xianglong Liu, Yitao Xu, Hang Yu, Yuqing Ma, and Tianlin Li. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. *IEEE TIP*, 2021. 1

[22] Jin Zhang, Wenyu Peng, Ruxin Wang, Yu Lin, Wei Zhou, and Ge Lan. Enhance domain-invariant transferability of adversarial examples via distance metric attack. *Mathematics*, 2022. 1, 2