计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(9): 2208−2232, 2025

多模态视觉语言表征学习模型及其对抗样本攻防技术综述

曾 诚[1] 葛云洁[2] 赵令辰[1] 王 骞[1]
[1](武汉大学国家网络安全学院 武汉 430072)
[2](武汉数学与智能研究院(武汉大学) 武汉 430072)
(chengzeng@whu.edu.cn)

Survey of Multimodal Vision-Language Representation Learning Models and Their Adversarial Examples Attack and Defense Techniques

Zeng Cheng[1], Ge Yunjie[2], Zhao Lingchen[1], and Wang Qian[1]
[1](*School of Cyber Science and Engineering*, *Wuhan University*, *Wuhan* 430072)
[2](*Institute for Math & AI* (*Wuhan University*), *Wuhan* 430072)

Abstract With the rapid development of computer vision, natural language processing, and deep learning technologies, multimodal vision-language representation learning models have demonstrated outstanding performance in tasks such as image captioning, text-to-image generation, and visual question answering. Such models have consequently become a focus of research in both academia and industry. However, the multimodal nature and complexity of these models provide attackers with more diverse attack vectors. Attackers can craft adversarial examples to mislead these models into outputting incorrect, harmful, or false content, posing an increasingly serious security threat to such models. This paper systematically reviews the current state of research on multimodal vision-language models. It also categorizes and summarizes the adversarial examples-based attack methods that have emerged in recent years against these models, as well as their defense strategies. Furthermore, it provides a detailed overview of the fundamental principles, implementation approaches, and conclusions of relevant studies. On this basis, this paper discusses the current state and future directions of security research in multimodal vision-language representation learning. Finally, the paper envisions the future prospects of applying vision-language representation learning technology in combination with interpretability techniques.

Key words representation learning; vision-language learning; adversarial attack; adversarial defense; cyber security

摘 要 随着计算机视觉、自然语言处理与深度学习技术的快速发展,多模态视觉语言表征学习模型在图像描述、文本生成图像、视觉问答等任务中展现出了卓越的性能,已成为当前学术界与工业界共同关注的研究热点.然而,这类模型的多模态特性和复杂性为攻击者提供了更加多样的攻击途径,攻击者可以通过对抗样本引导模型输出错误的、有害的或虚假的内容,使该类模型面临的安全威胁日益严峻.系统地梳理了多模态视觉语言模型的研究现状,同时,对近年来出现的针对该类模型的对抗样本攻击方法及其防御策略进行了分类总结,详细归纳了相关研究的基本原理、实施方法与研究结论.在此基础上,对多模态视

收稿日期: 2025-06-06; **修回日期**: 2025-07-09
基金项目: "叶企孙"科学基金项目(U2441240); 国家自然科学基金项目(62441238, U21B2018, 62441237); 国家资助博士后研究人员计划(B)档(GZB20250384); 湖北省博士后尖端人才引进项目(2004HBBHJD066)
This work was supported by the "Ye Qisun" Science Foundation (U2441240), the National Natural Science Foundation of China (62441238, U21B2018, 62441237), the Postdoctoral Fellowship Program (Grade B) of China Postdoctoral Science Foundation (GZB20250384), Hubei Provincial Postdoctoral Advanced Talent Introduction Project (2004HBBHJD066).
通信作者: 王骞(qianwang@whu.edu.cn)

觉语言表征学习的安全研究现状与未来方向进行了探讨，并展望了视觉语言表征学习技术在未来结合可解释性技术的应用前景.

关键词 表征学习；视觉语言学习；对抗性攻击；对抗性防御；网络空间安全

中图法分类号 TP18

DOI: 10.7544/issn1000-1239.202550410 **CSTR:** 32373.14.issn1000-1239.202550410

随着社会信息化程度的不断提高，多模态数据(multimodal data)呈现出了爆发式增长的趋势，在各行各业中逐步占据了主导地位. 多模态数据指的是包含多种信息来源或形式的数据，常见的如语言文本、视觉图像、视频和音频等. 近年来，随着社会需求的扩大和信息数据收集手段的增加，人们对人工智能的数据处理能力要求已经从少量、单模态的数据逐步转移到了大量、多模态的数据，这也使得多模态相关研究逐渐受到了研究人员的关注.

表征学习[1]是机器学习中学习特征技术的方法集合，其目的是将原始数据转换成为能够被机器学习有效开发的数据，后者常有更高、更多的抽象表示层次和更高效的有用信息. 针对多模态的表征学习技术[2]不仅要从视觉、语言、声音等多个模态的数据中提取出各自的关键语义信息，还需要对这些数据进行分析和融合，是计算机视觉(computer vision, CV)、自然语言处理(natural language processing, NLP)和语音识别(automatic speech recognizaion, ASR)多个领域技术的交叉研究方向. 多模态表征学习通过将文本、图像、音频、视频等多源异构数据投射到同一潜在空间，实现跨模态一致性和语义对齐，从而让模型能够在不同信息形式之间互相补充并融合特征. 因此，从多模态表征学习中获得的表征不仅能显著提升人工智能系统跨模态检索(如根据文字查找最相关图像或根据图像检索对应描述)的准确性[3]，还在视觉问答[4]、图像自动描述[5]、音视频情感识别[6]等任务中展现出强大的能力；在医疗影像与病历联合分析[7]、智能交互与辅助驾驶[8]等场景中，亦能有效增强系统的鲁棒性和泛化性能. 通过模拟人类多感官融合的机制，多模态表征学习为人工智能系统提供了更全面、更自然、更可靠的信息表达方式，支撑了从娱乐推荐到安全监控等诸多前沿应用的发展.

在视觉语言多模态模型部署前，通常采用“预训练-微调”的范式对模型进行训练. 模型的预训练(pre-training)通过大量数据和随机初始化来训练模型，这种模型被称为视觉语言预训练(vision-language pre-training, VLP)模型；微调(fine-tuning)技术则是在预训练模型基础上，使用特定的数据再次训练模型，使模型适应新任务的过程. 通过“预训练-微调”范式训练得到的模型能从多个模态中提取融合后的表征，使用者则可以根据自身需求对这些表征进行进一步的利用，完成不同类型的下游任务.

对抗样本攻击(adversarial attack)是使用对抗样本诱导人工智能模型产生错误输出的攻击，其概念由 Szegedy 等人[9]在单模态图像分类模型上提出. 在人工智能模型的输入样本上故意添加一些人眼无法察觉的细微的干扰，即可生成对抗样本. 与之相反，对抗样本防御技术通过采用不同的检测、筛除和训练手段，增强人工智能模型的鲁棒性，或使得模型免于对抗样本攻击. 对抗样本攻击技术和防御技术经过近十年的博弈发展，在图像、文本和音频等单模态表征学习领域都趋向成熟. 近年来，研究人员的重心也逐渐向多模态表征学习领域转移，尤其是 VLP 模型，在传统单模态对抗样本攻防技术的基础上，针对 VLP 模型的对抗样本攻防技术也考虑到了多模态信息的融合攻击[10]，产生了新的安全隐患.

现有的针对多模态视觉语言表征学习的系统梳理工作[11]将表征学习划分为基于相似性和基于预训练的 2 种架构，并对这 2 种架构的模型设计、预训练任务和下游性能评估进行了系统阐述；文献 [12] 则对多模态视觉语言预训练模型的常用技术、预训练数据和下游任务进行了详细总结. 但一方面，近年来多模态表征学习发展迅猛，产生了更多具有卓越性能的 VLP 模型[13-18]，这部分模型的细粒度梳理工作亟待进行；另一方面，多模态信息之间的交互特性已经被攻击者利用来设计更强的融合攻击[19-20]，当前综述针对多模态表征学习的对抗性攻击与防御研究关注不足. 随着多模态表征学习的下游模型集成应用的兴起，攻击者可以在表征学习模型部署的任何阶段发起对抗性攻击，造成资金损失、信息泄露等实质性危害. 而面对不断演变的攻击手段，当前的防御技术仍难以覆盖其全部，因此，评估和提升多模态表征学习模型的抗攻击能力已经成为该领域的关键问题.

本文针对上述问题，从原理层面对近年来的视

觉语言表征学习技术、针对视觉语言表征学习模型对抗性攻击和防御技术进行汇总分析,尽可能全面地展现这一研究领域的最新研究方向和关键问题.本文的主要架构如图 1 所示. 总体来说,本文的主要贡献可以总结为 4 个方面:

1)系统回顾了近年来多模态视觉语言表征学习的技术演进,从细粒度对齐、轻量高效化与多任务通用 3 个研究角度,清晰呈现了近年研究中的结构特征与代表方法.

2)从对抗安全角度,全面回顾了视觉语言表征学习面临的对抗样本攻击技术,按照攻击动机与策略将攻击方法划分为模态联合攻击、通用扰动与补丁攻击以及面向具体任务的攻击三大类. 系统总结比较了其技术机制与迁移能力.

3)总结并分类了当前多模态表征学习中的对抗防御策略,包括对抗训练、结构增强、防御检测 3 类方法,结合视觉语言模型的特点分析其优势与局限.

4)讨论并展望了多模态视觉语言模型在安全鲁棒性方面的研究趋势与挑战,指出当前防御方法仍难全面抵御高强度多模态攻击,同时指出对模型可解释性的研究同样亟待推动.

1 研究背景

1.1 视觉语言表征学习

视觉语言表征学习的核心是最早在自然语言处理领域提出的 Transformer[21] 框架,一种基于自注意力机制的深度神经网络架构. Transformer 摒弃了传统循环或卷积的序列处理方式,通过引入多头注意力(multi-head attention)机制,并行学习不同子空间的表示,计算输入序列中任意位置间的注意力权重,有效建模长距离依赖,提升了模型对多样化特征的捕捉能力.

ViT(vision Transformer)[22] 将 Transformer 架构应用于端到端的图像表征学习. ViT 将输入图像划分为若干个固定大小的补丁,补丁可以看作是文本 Transformer 中的词元嵌入,在每个补丁中加入可学习的位置编码,用以编码二维平面上的补丁坐标位置. 将包含额外分类标记(classification token, [CLS])的补丁序列输入到多个 Transformer 编码器层后,每层依次应用多头自注意力和前馈网络,并通过残差连接与层归一化提高训练稳定性和收敛速度. 在预训练或微调阶段, ViT 会将编码器输出的 [CLS] 向量用作图像分类或其他视觉任务的特征表示,实现端到端的表征学习.

Transformer 为视觉语言表征学习提供了高效编码的方式. 基于使用 Transformer 的方式,在模型架构上,视觉语言表征学习模型可以分为双流模型和单流模型两大类. 双流模型分别为图像和文本设置了独立的编码器,分别对视觉和语言进行独立编码,并通过跨模态注意力(co-attention)层或者对比学习交互 2 个模态的信息,因此支持更灵活的检索任务;单流模型则将图像特征与文本标记拼接输入同一 Transformer,通过融合注意力(merged attention)完成早期融

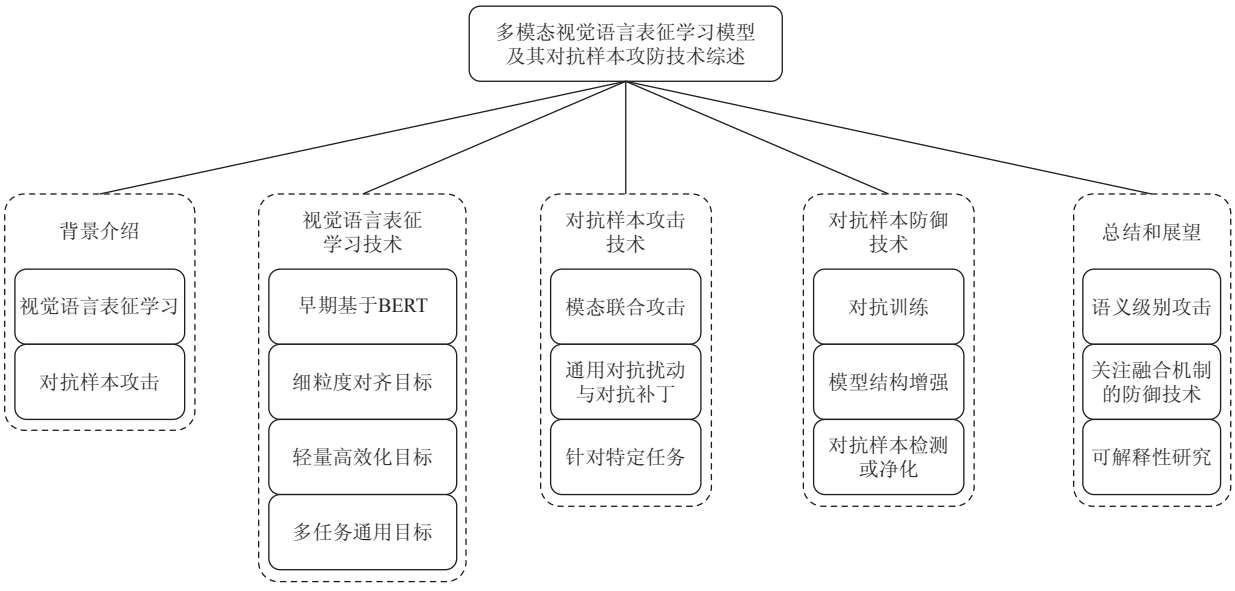


Fig. 1 Framework diagram of our paper

图 1 本文框架图

合，拥有较少的参数和统一的编码，计算高效. 对于双流模型而言，又可以从模态的交互机制将其分为融合编码器（fusion encoder）、双编码器（dual encoder）和二者结合的混合方案 3 类. 融合编码器直接在编码器内部通过自注意力或交叉注意力实现跨模态交互；双编码器仅在嵌入空间进行对比或匹配；混合方案兼顾两者优势，既有对比对齐又有深度融合，对比学习往往是深度融合的前置任务，用于加快模态的融合速度.

视觉语言表征学习模型常用的预训练任务包括图文匹配任务（image-text match, ITM）、图文对比学习任务（image-text contrastive learning, ITC）、图像掩码建模（masked image modeling, MIM）、文本掩码建模（masked language modeling, MLM）等[12]. 与下游任务不同，这些任务通常不直接对应实际应用需求，只通过特定的损失函数来训练模型提取表征. 在预训练数据集方面，视觉语言表征学习使用的大部分的数据集都由图像+文本描述构成，如 COCO[23]、Conceptual Captions（CC）[24]、SBU Captions（SBU）[25]、LAION-400M[26]、ALIGN[3] 等. 也有数据集用于特定的任务，如用于视觉定位的数据集 Visual Genome（VG）[27]，其含有区域级别的描述；用于问答的 VQA 数据集[28]，其含有 1 100 多万的图像问答实例. 包含丰富视频资源的 YFCC100M[29]、仅有类别标签但拥有近 30 亿图像的 JFT-3B[30] 也常常用于视觉语言表征学习模型的预训练.

采用“预训练-微调”范式训练得到的视觉语言预训练模型已广泛应用于图文检索（image-text retrieval, ITR）、视觉问答（visual question answering, VQA）、图像自动描述（visual captioning, VC）、视觉推理（visual reasoning, VR）、视觉定位（visual grounding, VG）及视觉蕴含（visual entailment, VE）等多种下游任务，取得了突破性的成果[12]，加速了视觉语言表征学习产业化的步伐，使得 VLP 模型成为研究和开发中不可或缺的关键技术. 通过评估表征学习模型微调后在下游任务上的准确率，我们可以确认表征学习是否从数据中提取到了有效的表征信息；在数据投毒和对抗样本的场景下，也能够确认表征学习模型的鲁棒性，为进一步的安全相关工作提供指导.

1.2 对抗样本攻击

对抗样本攻击是指利用对抗样本诱导深度神经网络产生错误输出的攻击，由 Szegedy 等人[9] 发现，随后在各类神经网络模型中被广泛验证. 当前学术界尚未对对抗样本的成因产生统一共识，Szegedy 等人认为[31]，对抗样本是深度神经网络的极端非线性特性造成的，即对抗样本位于数据流形的低概率区域. 而另外一些观点认为[32]，对抗样本产生的原因是高维空间中的线性行为，即模型的局部线性特征，因此在模型使用了线性激活函数，如 ReLu 和 Maxout 的时候，其低鲁棒性也十分明显.

对抗样本具有可迁移的性质[33]，即针对一个深度学习模型生成的对抗样本，在攻击其他模型时也能产生一定的误导效果. 根据攻击者可获取的模型信息，对抗样本攻击可分为白盒攻击与黑盒攻击 2 种. 白盒攻击情景下，整个模型对于攻击者完全透明，攻击者了解模型的结构、训练数据集以及参数等具体细节，通过模型梯度来设计对抗样本；而黑盒攻击情景下，攻击者并不知道模型的内部结构. 在现实应用场景中，视觉语言表征学习模型往往具有不同的预训练策略和下游任务数据集，即模型对于攻击者而言往往是黑盒模型，因此评估针对 VLP 的对抗样本攻击时，往往首先评估攻击的可迁移性能，即在一个模型上生成的对抗样本，攻击其他结构不同、数据集不同甚至下游任务不同的模型时的成功率.

目前，针对单模态表征学习的对抗样本攻击技术已经产生了良好的攻击效果和可迁移性，如针对视觉图像的 FGSM[32] 攻击、PGD[34] 攻击和 C&W[35] 攻击，以及针对语言文本的 BERT-ATTACK[36] 等. 但文献 [10] 发现，直接将上述针对单模态表征学习的攻击技术迁移到多模态表征学习模型时，2 种模态的攻击可能会相互冲突，往往无法取得较好的攻击效果. 因此，不同模态的对抗扰动应该协同考虑，而不是独立考虑. 近年来，围绕多模态视觉语言表征学习模型的攻防机制与安全性问题的研究逐渐增多，并取得了显著进展，揭示了视觉语言预训练模型在面对多模态对抗样本时的脆弱性.

2 视觉语言表征学习模型介绍

已有文献主要针对 2022 年及以前的视觉语言表征学习模型进行总结与分析[11-12]，但在 2022 年以后，随着视觉语言表征学习的广泛应用与不同需求，其发展路径逐渐呈现出多元化趋势，视觉语言表征学习模型也进入了一个新的迅速发展阶段. 本节首先在 2.1 节简单回顾了视觉语言表征学习模型的早期发展脉络，随后将 2022 年以后的新颖技术分为 3 种主要类型：以细粒度对齐为目标的表征学习、以高效轻量化为目标的表征学习、以多任务通用为目标的

表征学习, 在 2.2~2.4 节分别介绍, 最后在 2.5 节对比小结.

2.1 基于 BERT 拓展的视觉语言表征学习

从 2019 年开始, 基于 BERT[37] 构建的视觉语言表征学习技术开始了在多模态视觉语言领域的初步尝试. BERT 是于 2018 年提出的一种基于 Transformer 的双向编码器预训练模型, 通过掩码语言建模和下一个句子预测任务, 在大规模无标签文本上学习上下文关联表示, 极大提升了模型在自然语言处理方面的多项基线性能. 早期的视觉语言表征学习工作拓展了 BERT 的结构, 将视觉信息以不同的方式编码到模型中, 如: ViLBERT[38] 将 BERT 拓展为双流结构, 分别编码视觉与文本, 并通过共注意力模块进行跨模态融合, 以图文匹配作为预训练目标, 实现了良好的图文检索效果; VisualBERT[39] 保持了单流结构, 直接将文本词元与图像区域特征拼接后输入标准 BERT 模型, 加上了模态分割编码, 通过 Transformer 进行对齐和融合; VL-BERT[40] 在保持单流结构的同时对输入层次和视觉特征做了更细致的 4 层设计, 保留了 BERT 的掩码语言建模和掩码区域分类作为预训练目标. 这些基于 BERT 架构的研究探索了模态间的融合方式与预训练任务的差异, 验证了跨模态信息交互的可行性与有效性.

自 2021 年起, 以 CLIP[41] 为代表的大规模对比学习成为了视觉语言表征学习的主流. 对比学习是一种常见的自监督学习方法, 通过分割正样本对与负样本对, 优化模型使得正样本对的表征在共享空间中更接近, 负样本对的表征更远离, 实现模型对表征的良好语义提取. 在模型结构方面, CLIP 使用了双流结构分别提取视觉特征和语言特征, 图像编码器使用 ViT 或者传统的卷积神经网络如 ResNet[42], 文本编码器则使用 BERT 来完成, 将图像和文本的特征映射到同一个表征嵌入空间进行相似度计算. 由于 CLIP 的自监督学习依赖于图像和其对应的文本描述之间的自然配对关系, 无需人工标注类别标签, 从而能够利用规模庞大的自然图文对数据集, 在零样本学习、多模态检索和图像生成辅助[43] 中均有明显的优势, 对之后的工作产生了很大的影响.

CLIP 的对比学习任务可以与其他任务结合, 实现表征在其他下游任务上的性能平衡. Mu 等人[44] 在 CLIP 的对比目标之外加入了图像掩码建模任务, 实现了在图像线性分类任务和零样本学习任务上的性能提升. Singh 等人[45] 则更进一步, 同时加入了文本掩码建模任务, 并在单模态数据上联合训练, 实现了在 35 项视觉、语言与多模态任务上的平衡性能. Yu 等人[46] 设计了一种编码器-解码器架构, 将对比学习与自回归生成无缝融合, 通过解码器自回归地生成文本, 实现了对齐与描述能力的统一. Li 等人[47] 从博弈论视角重新审视视觉与语言的对比交互提出了 LOUPE, 将图像补丁与文本词元视为合作博弈的玩家, 使用沙普利(Shapley)交互值衡量每对输入的语义贡献, 建模视觉区域与文本短语的细粒度交互. 相比 CLIP 的直接计算余弦相似度, LOUPE 能高效地在语义级别强化表征之间的区域-短语对齐, 在视觉检索、视觉定位与多模态推理任务上实现最优性能. Lu 等人[48] 则致力于构建通用视觉基础模型, 将对比学习扩展到了(图像, 类别标签, 文本描述)的三元组合, 在同一个对比学习框架下进行联合训练, 在正样本三元组中两两拉近元素在共享表征空间中的距离, 实现了模型在少样本情况下的多任务迁移能力.

2.2 以细粒度对齐为目标的视觉语言表征学习

CLIP 的双流设计仅在损失层面对图像和文本整体进行全局匹配, 未在模型内部引入跨模态注意力, 因而无法捕获局部的对应关系, 导致仅靠对比学习训练的模型提取出的多模态表征在需要深度交互的下游任务(如视觉问答、视觉推理和视觉定位)和生成任务上表现欠佳. 因此, 近年来许多研究探索了在对比学习基础上强化多模态视觉语言表征细粒度交互对齐的方法, 以追求表征在精细下游任务上的良好表现.

为了让多模态表征能够更细粒度地体现图像区域和文本词元的关系, Li 等人[13] 在 ALBEF 中提出了“对齐+融合”(align before fuse)的多模态表征学习框架. ALBEF 的网络结构如图 2 所示, 其在图像编码器与文本编码器输出的表征上引入中间图文对比损失(image-text contrastive, ITC), 完成模态表征的全局对齐, 然后将对齐后的表征输入跨模态的 Transformer 获得多模态表征, 捕获视觉区域与文本词元的细粒度对应关系. 针对自监督学习中的图文数据存在文本噪声的问题, ALBEF 还引入了一套动量模型, 由动量模型生成的软标签作为伪目标, 平滑对比损失的学习信号, 使对齐和融合过程更稳健. Yang 等人[14] 在 ALBEF 的基础上进一步考虑了模态内的对比自监督和局部-全局互信息最大化, 提出了三重对比学习预训练模型 TCL. TCL 在同一模态内的图像区域或文本短语之间执行对比学习, 以确保相似同模态样本在嵌入空间中保持紧密邻近; 同时通过最大化局部区域与全局表示之间的互信息量, 使模型能够捕捉

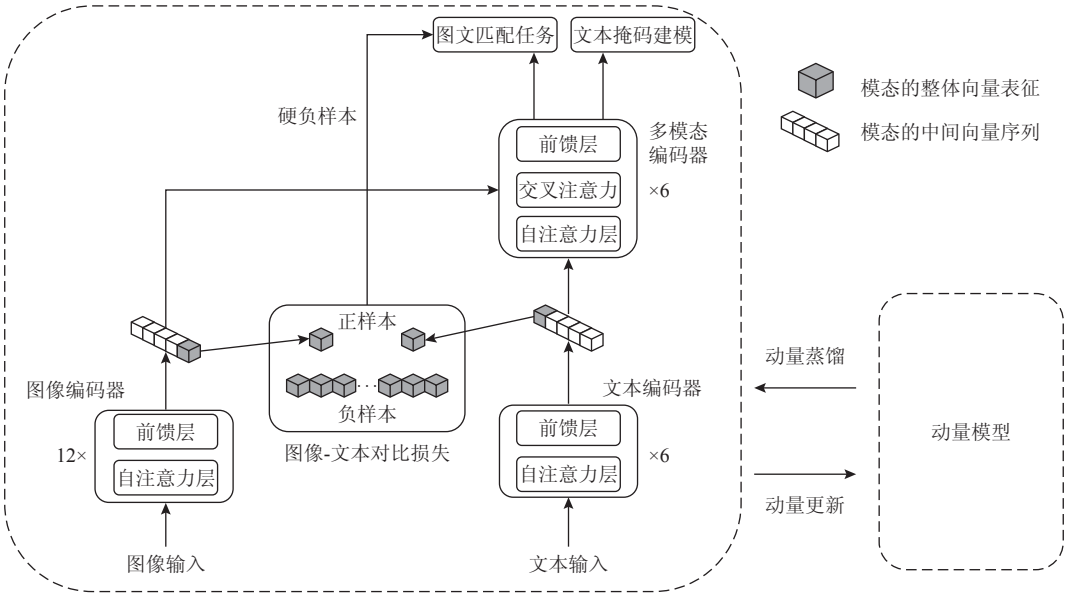


Fig. 2 Architecture of ALBEF

图 2 ALBEF 的架构

图像结构和文本语义的细粒度对应关系，有效弥补了单一全局对比对齐的不足，在多项基准下游任务性能上超过了 ALBEF.

部分研究认为，增强模型的视觉理解能力，让模型在如视觉定位和物体检测等下游任务上提升性能，可以获取更细粒度的对齐表征. Zeng 等人[18] 在 2022 年提出了 X-VLM，在无需大量人工标注的情况下，实现图像和文本从全局到局部的多粒度语义对齐. X-VLM 在预训练中构造多粒度对齐任务，在全局对比学习的基础上引入了视觉概念定位与匹配机制，利用了少量有边框标注的数据，从文本中提取短语/实体(如名词短语)，让模型提取表征去预测图像中对应的区域位置，并通过区域-短语对的对齐损失进行训练. 同时，X-VLM 也设计了无标注数据上的检测式预训练任务，通过将预训练的探测器提取的区域提议与文本短语进行匹配，从弱标注数据中学习表征的细粒度对应，为模型提取复杂场景中多个物体及其关系的表征奠定了基础. Yao 等人[49] 提出了 FILIP，通过引入跨模态的晚期交互机制，利用视觉和文本标记之间的最大相似度，引导模型学习更精细的跨模态表示. Zhang 等人[50] 提出的 GLIP 统一了物体检测和短语定位(phrase grounding)两种任务，给模型输入图像和一系列包含所有类别名称的文本短语或图像描述，模型需要在图像中框出对应短语提及的物体，从而利用大规模目标检测数据以及短语定位数据共同训练，并设计了自监督的框预测机制来利用无标注的海量图文对. GLIPv2[51] 在 GLIP 的基础上将预训练任务扩展为 3 项，加入了区域-词对比任务和文本掩码建模任务，实现了定位预训练和跨模态语义预训练的有机结合，其提取的跨模态表征为未来同时涉及视觉定位和语言理解的复杂任务打下了基础.

训练细粒度的表征学习往往需要付出额外的训练数据标注成本. Zhong 等人[52] 借鉴了模型蒸馏的思想，提出了 RegionCLIP，首先利用一个预训练的 CLIP 模型，把图像划分的若干区域提案与相应的文本描述进行初步匹配，自动生成区域级别的伪标签；然后用这些(区域，文本)对作为训练数据，对一个新模型进行对比学习预训练，以对齐区域视觉特征和文本概念，利用大规模图文对弱标签实现了无监督下的细粒度区域对齐表征学习. Liu 等人[53] 引入了教师-学生蒸馏策略提出了 UTA，首先利用预训练 CLIP 模型作为教师，训练一个 ViT 学生模型，使其每个未掩盖的图像补丁特征对齐到 CLIP 视觉编码器对应的图像补丁特征. 通过这种逐补丁对齐，学生 ViT 被自动对齐到 CLIP 的多模态空间，从而即使不直接用图文对数据训练，预训练后的 ViT 也具备良好的跨模态表示能力和细粒度表征对齐能力. 相比纯视觉自监督的方法，UTA 避免了预训练与微调目标不一致的问题，而且由于不需要添加掩码标记等额外设计，训练更为高效.

多尺度和局部-全局融合机制也是视觉语言表征学习细粒度对齐的重要方向. 例如 Ouyang 等人[54] 提出的 SLViT 在 Transformer 中引入了语言引导的多尺度融合注意力机制，以同时捕获图像的全局语义和

局部细节, 在细粒度理解任务上取得了显著提升. Gao 等人[55] 提出的 PyramidCLIP 在模型中引入图像特征的金字塔表示, 让不同层次的视觉特征与文本逐级交互对齐, 使模型既关注整体图景, 又不遗漏小物体或细节属性.

部分工作将结构化知识融入细粒度对齐过程, 也取得了不错的效果. 例如在视觉问答等需要复杂推理的任务中, 引入多模态图结构可以帮助模型细粒度理解和对齐跨模态表征. Wang 等人[56] 提出了 SGEITL 框架, 通过引入视觉场景图(scene graph)来增强视觉常识推理任务中的图文对齐. 该方法在 Transformer 中设计了多跳的图结构注意力机制, 将图像中对象及其关系的先验结构注入自注意力计算, 从而引导跨模态对齐更加关注语义相关的实体对. 这种场景图增强的模型有效提升了模型对实体关系的细粒度表征提取能力, 在视觉常识问答等复杂推理任务上取得了显著的性能提升. 类似地, Ye 等人[57] 将外部知识图谱引入视觉语言表征学习, 提出了 DANCE, DANCE 将知识图谱中的子图线性化为文本描述, 作为对图文对的数据增强, 在模型训练中动态注入这些常识, 通过这种方式提取的细粒度融合表征在常识推理等需要外部知识的任务上表现突出.

2.3 以轻量高效化为目标的视觉语言表征学习

随着视觉语言表征学习在各类任务上取得了卓越的效果, 其庞大的参数规模也带来了模型部署和训练上的困难. 因此, 近年来涌现出许多轻量化或高效架构的新方法, 旨在在降低模型规模和计算量的同时保持模型提取良好视觉语言表征的性能.

对视觉语言表征学习轻量化的首次尝试来自 2021 年 Kim 等人[16] 提出的 ViLT, 其结构如图 3 所示, ViLT 将双流结构中臃肿的视觉编码器完全移除, 采用纯粹的 Transformer 单流架构, 将图像直接划分为固定大小的图像补丁, 线性嵌入后与文本序列一起输入一个单一 Transformer, 实现视觉和文本特征的融合交互. 为了保证良好的表征提取效果, ViLT 在预训练中结合了文本掩码建模任务和单词-补丁细粒度对齐任务等, 为后续轻量高效的视觉语言模型设计提供了新思路. Khan 等人[58] 在 ViLT 的基础上提出了 SIMLA, 解决了 ViLT 未显示关注局部细节的问题. SIMLA 采用了和 ViLT 相似的单流架构, 但设计了 2 个新的预训练任务来实现多级对齐: 1)对称跨模态重建. 随机遮挡图像补丁或文本词元, 让模型从另一模态的信息来重建被遮内容, 促进模型学会细粒度地对齐图像和语言片段以完成重建. 2)伪标签语义概念预测. 利用 Transformer 的注意力找出文本中的关键词, 再通过动量编码器推荐出图像中尚未在文本中提及的重要概念词作为伪标签, 然后训练视觉编码器去预测这些图像中存在但文本标签缺失的概念, 从而使模型学习更高层的语义概念, 将视觉特征和更抽象的语言语义对齐. SIMLA 在轻量化的单一 Transformer 架构下同时实现了全局和细粒度多层次对齐.

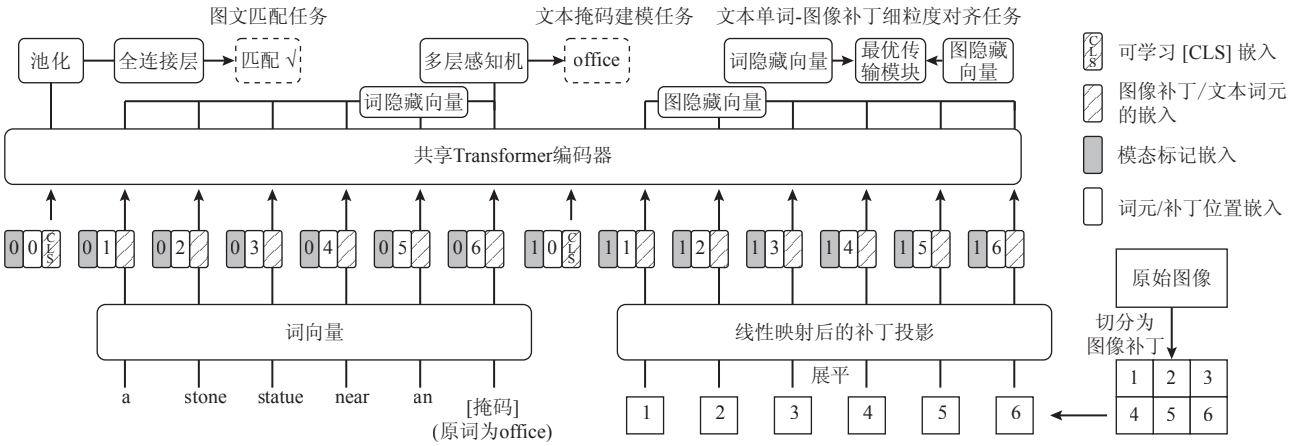


Fig. 3 Architecture of ViLT

图 3 ViLT 的架构

虽然 ViLT 等端到端轻量化模型实现了高效率的视觉语言表征学习, 但这类模型在小规模预训练数据下的性能可能不及使用了视觉编码器方法的性能. Dou 等人[17] 系统性研究了端到端 Transformer 模型的设计取舍, 探究如何在保持高效的同时提升下游任务性能并提出了 METER. METER 框架对多种模型设计维度进行了全面探索, 包括视觉编码器(如 CLIP 的 ViT、Swin Transformer[59] 等)、文本编码器(如

RoBERTa[60]、DeBERTa[61]等)、跨模态融合方式(merged-attention和co-attention)、整体架构形式(编码器解码器或单编码器),以及预训练目标(如加入图像掩码建模任务等),找到了一系列有效配置并构建最佳模型,推动了此后大量端到端模型的发展.

双流模型的多模块架构(双编码器、表征融合模块,其他外部模块如视觉检测器和光学字符识别模块等)和多重预训练目标显著增加了模型的复杂性.针对这一问题,微软提出了GIT[62],GIT仅用1个图像编码器和1个文本解码器构成序列到序列模型,以单一生成式目标统一各种任务.由于生成式Transformer在预训练和微调阶段架构一致,因此可同时处理描述生成和问答等任务,只需通过扩大全数据规模来提升性能.GIT采用ViT提取图像表征,将其输出作为文本解码器的条件输入,通过前缀语言模型目标来训练.虽然GIT只使用了语言模型损失,但大规模的多样化数据让模型学会了图像到文本的映射,从而增强了学习到的表征的泛化能力.

Xu等人[63]针对双流模型融合不充分、单流模型无法获取强大单模态表征的问题,提出了融合两者优势的高效架构BridgeTower.BridgeTower依旧采用了双编码器,但在视觉编码器和文本编码器的顶层之间插入了多重桥接层(bridge layers),在2种架构之间搭建“桥梁”,将视觉和文本的高层表示在每一层跨模态Transformer中都进行交互,而不仅仅在最后融合.这些桥接层相当于在每层融合模块中,将提取的视觉和文本表征逐层引入,从底层到高层逐步对齐不同语义层次的图像和文本表征.这种设计保留了各自模态编码器的强大表示能力,同时通过多层次融合实现细粒度对齐.BridgeTower在几乎不增加参数量的情况下,仅用约400万图文对就取得了优异表现.

传统的知识蒸馏和模型剪枝等方式也能用于视觉语言表征学习模型的轻量化,如EfficientVLM[64]通过直接的“蒸馏+剪枝”两阶段框架,将大型预训练视觉语言表征学习模型压缩为仅约9 300万参数(教师模型的44%),并保留了98.4%的模型性能.另一类方法专注于解决双流模型中视觉编码器过于臃肿的问题,如BUS[65]提出了“由下至上”的图像补丁摘要机制,在ViT视觉主干中嵌入一个文本语义感知的补丁选择器,以粗粒度选出与文本相关的图像补丁,然后接入一个Transformer补丁抽象解码器对选定的补丁序列进一步抽象.这种自下而上的补丁摘要机制使得模型训练效率提高了50%,且模型学习到的表征性能不降反升.Jian等人[66]提出的SimVLG则采用了逐渐合并相似图像补丁的策略,在训练过程中动态压缩图像序列信息.这种单阶段、单损失的框架有效保留了语义,将视觉语言表征学习模型的预训练收敛速度提高约5倍,且只需原来约1/10的数据量即可达到与现有模型相当的性能.

也有不少工作尝试通过替代Transformer结构来提升视觉语言表征学习的效率与可扩展性,如Gu和Dao提出的Mamba系列模型[67].Mamba系列模型用结构化状态空间模型替代了传统Transformer的自注意力机制,通过连续卷积和递归计算高效建模长距离依赖,这种架构在处理长序列时仅具有线性复杂度.最新的VL-Mamba模型[68]基于该架构,集成了一个Mamba LLM、预训练视觉编码器,以及专门的多模态连接器(multi-modal connector, MMC)来桥接二维图像与一维序列信号.MMC采用双向扫描和交叉扫描等机制高效提取空间上下文,将视觉补丁特征转换为序列模型可处理的格式.由于去除了复杂的自注意力计算,此类无Transformer架构在保持表征提取性能的同时显著降低了计算开销,加速了推理,适合实时应用.

2.4 以多任务通用为目标的视觉语言表征学习

随着视觉语言表征学习的应用场景增多,表征学习模型在实际应用之前往往需要用全新的数据集和下游任务进行微调,这个过程不仅复杂,而且受数据集规模的限制.因此,将多种视觉语言任务统一到同一表征学习模型中训练,让模型提取更加通用的表征,也是近年来的重要研究方向之一.

一些代表性工作通过统合模型架构和多个训练目标,成功地在单个表征学习模型中实现了对理解类和生成类任务的兼顾.Li等人在2022年提出的BLIP[15]就是其中的典型代表,BLIP提出了一个多模态的编码器-解码器混合架构,其结构如图4所示.一方面,依旧使用图文对比损失训练图像和文本的双编码器,以对齐视觉和语言表征;另一方面,引入了图像条件文本编码/解码器,通过图文匹配损失和语言建模损失来训练模型进行图文对判断和图像描述生成任务.其文本编码器在接收图像特征时加入了交叉注意力层用于融合多模态信息,文本解码器则与编码器共享视觉交叉注意力模块,用因果自注意力生成描述.BLIP在图片描述、视觉问答、跨模态检索等多项任务上取得了优秀表现,充分验证了统一理解与生成表征学习的可行性.

另一些工作尝极简化预训练目标并利用海量弱

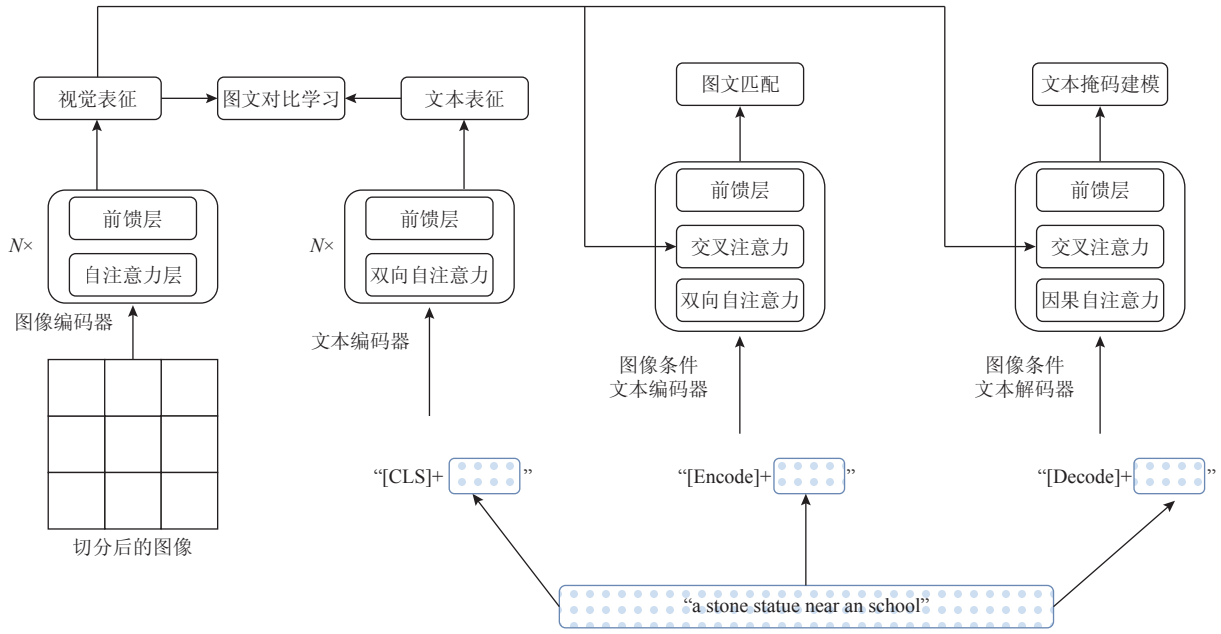


Fig. 4 Architecture of BLIP

图 4 BLIP 的架构

标注数据, 以降低训练复杂度、提升表征的通用泛化能力. 如 Wang 等人[69] 提出的 SimVLM 将图像补丁序列当作前缀与文本序列拼接输入 Transformer 编码器, 然后使用解码器根据该前缀生成后续文本. 为了弥补弱目标的训练信号问题, SimVLM 利用了 10 亿级别的图文对作为弱对齐数据进行训练. 事实上, 其训练方式与纯大语言模型类似, 仅通过 PrefixLM 生成目标, 在序列开头增加了图像信息, 无需额外的下游任务, 却能在弱对齐图文数据上同时学习通用性良好的视觉语言对齐表示. 微软公司提出的 OFA[70] 则基于编码器-解码器的序列到序列框架, 将多种跨模态和单模态任务(如图像字幕生成、视觉问答、图像分类、视觉定位和纯语言建模等)都转换成了“给定输入序列, 生成目标序列”的统一形式, 在预训练和微调时都采用指令驱动学习, 即将不同的下游任务转换成不同类型的问答任务指令, 让模型仅在这些问答任务上回答并删除, 让不同任务共享完全相同的模型参数, 在只使用了 2 000 万公开图文对的情况下实现了跨任务的知识融合和通用表征提取. Zhu 等人提出的 Uni-Perceiver[71] 系列模型则采用了单一 Transformer 编码器配合轻量模态特定编码器的架构, 不同模态的输入(图像补丁、文本词元等)首先通过各自的特定编码器编码为统一格式的序列表示, 然后送入共享的 Transformer 编码器进行处理, 并将各种任务都统一抽象为最大似然估计问题. 对于输入, 对应的输出目标(类别标签、文本序列、检测框等)都表示成向量, 模型学习让正确的目标与输入表示的相似度最大. Uni-Perceiver 提取的表征展现了强大的通用泛化能力, 在一些新任务上零微调即取得了有竞争力的表现, 其最新版本[72] 通过引入任务无关解码器和区域提案网络(region proposal network, RPN)作为视觉输入达到了新的通用性能高度.

Bao 等人提出的 VLMo[73] 创新性地提出了多模态专家混合(mixture of modality experts, MoME) Transformer 架构, 同时支持双流模型的高效表征和单流模型的融合表征. VLMo 的每个 Transformer 层内, 不再是一个固定的前馈子网络(feed-forward neural network, FFN), 而是包含多组专家模块如图像专家 FFN、文本专家 FFN 等, 以及一个共享的自注意力子层. 输入数据中的图像补丁将被路由到视觉专家, 文本词元则被路由到文本专家, 表征融合时则使用共享注意力层打通. 这种模块化设计使模型参数在一定程度上按模态分块, 既能作为统一 Transformer 融合使用, 又可在需要时部分充当双编码器, 展示了统一视觉语言表征学习模型多用途的前景. 其多路 Transformer 架构也被后来的模型所借鉴, 如 Wang 等人提出的 BEiT-3[74] 就延续了 VLMo 灵活的模块化架构, 但在训练任务上, BEiT-3 将文本掩码建模任务扩展到了多模态. BEiT-3 将图像用离散的“视觉词汇”表示, 利用 VQ-VAE[75] 将图像补丁编码为视觉单词, 称为“图像语”(imglish), 然后在 Transformer 上对图像序列、文本序列以及图像-文本混合序列都执行文本掩码建模任务

作为训练，模型在训练中需要同时预测遮挡的图像语、文本和图文对混合词元，也要预测遮挡的文本词元，还有遮挡的图文对混合词元. 通过这种并列的方式，BEiT-3 能够同时学习 2 种模态的通用基础表征，在大规模数据下取得了当时最好的下游任务表现.

2.5 小 结

在本节中，我们系统梳理了视觉语言表征学习模型近年来的发展脉络. 早期的视觉语言表征学习以 BERT 等语言模型为基础，将视觉信息编码到和语言模型相同的表征空间中进行融合，验证了不同模态表征融合的可行性. 随着以 CLIP 为代表的大规模无监督对比学习技术和“预训练+微调”范式的推广，研究者们开始从不同的实际需求出发，展示了视觉语言表征学习技术的多维发展潜力：通过引入更细粒度的对齐融合框架、将视觉模态进行细粒度分割以及采用从全局到局部的多尺度融合机制等方案，可以获得在精细下游任务上表现良好的细粒度表征；通过优化单流架构、模型蒸馏和剪枝等方案，可以在显著降低模型规模和计算量的情况下高效获取表征；通过统合多个预训练目标或极简预训练目标，可以获取更加通用、微调难度更低的多模态表征. 我们在表 1 中给出了近年来部分视觉语言表征学习模型的总结，并在表 2 中给出了第 2 节提及的所有模型按照结构类型和训练目标进行分类的汇总表.

Table 1 Summary of Vision-Language Representation Learning Models

表 1 视觉语言表征学习模型总结

研究目标	实现方法及原理	代表工作
细粒度对齐	通过“对齐+融合”机制，捕获视觉区域与文本词元的局部对应关系	ALBEF[13]、TCL[14]
	用视觉定位等特定下游任务增强模型的视觉理解能力，从弱标注数据中学习细粒度对应	X-VLM[18]、GLIP[50]
	采用多尺度的图像融合机制，自下而上逐步对齐表征，实现整体与细节的平衡	SLViT[54]、PyramidCLIP[55]
	借助知识图谱等结构化知识，帮助模型细粒度理解跨模态表征	SGEITL[56]、DANCE[57]
轻量高效化	采用单流架构端对端模型，结合多模态掩码建模任务等技巧实现模型轻量化	ViLT[16]、METER[17]
	以单一生成式目标统一各种任务，避免使用多重预训练目标依赖的多模块架构	GIT[62]
	采用桥接高效架构融合单流与双流模型的优势，几乎不增加参数量	BridgeTower[63]
	知识蒸馏、模型剪枝等传统模型压缩方案	EfficientVLM[64]、BUS[65]
多任务通用	统合模型架构和多个训练目标，在单个模型中实现对理解类和生成类任务的兼顾	BLIP[15]
	极简化预训练目标并利用海量弱标注数据提升表征的通用泛化能力	SimVLM[69]、OFA[70]

3 针对视觉语言表征学习的对抗样本攻击

自 2021 年 CLIP 出现以来，大量工作开始探讨针对视觉语言表征学习的对抗样本攻击方法，由于视觉语言表征学习模型在模型结构、预训练任务和数据集上往往有较大的区别，模型对攻击者而言往往是黑盒模型，针对视觉语言表征学习的对抗样本攻击研究均会考虑黑盒场景下的性能. 因此与传统的黑盒攻击/白盒攻击分类方法不同，本节根据对抗样本攻击的研究动机和内容将其分为 3 类：模态联合攻击、通用对抗扰动与对抗补丁攻击，以及针对特定下游任务的攻击，并在 3.1~3.3 节分别介绍这 3 类攻击，最后分析小结.

3.1 模态联合攻击

在尝试模态联合攻击以前，就有部分工作尝试结合多模态损失仅扰动视觉模态来影响文本相关的预测，如 Vishwamitra 等人[76] 提出的 MUROAN 框架就设计了一种解偶联攻击（decoupling attack），没有采用梯度优化方式，而是启发式地通过选择性移除或扰乱输入中的少量关键视觉特征，来“解耦”图像和文本之间的正确对齐，从而误导模型决策 MUROAN；只需改变不到 1.2% 的输入空间，即可 100% 破坏模型判断. 该工作揭示了多模态融合机制的脆弱性，为之后基于融合特征的攻击奠定了基础.

2022 年，Zhang 等人[10] 系统研究了针对视觉语言表征学习的对抗样本攻击策略，研究了 4 种类型的攻击：只生成图像对抗样本的攻击 PGD[34]、只生成文本对抗样本的攻击 BERT-Attack[36]、同时分别生成 2 种对抗样本的攻击 Sep-Attack（separate multimodal adversarial attack）、联合扰动攻击 Co-Attack（collaborative multimodal adversarial attack）. 他们分析了 4 种类型的攻击对攻击效果的影响，观察到单独攻击图像或文本（前 3 种方案）均不如联合扰动攻击有效，他们对此的解释如图 5 所示. 图 5 展示了分别扰动攻击和联合扰动攻击在样本空间中的区别，其中 x_i，x_t 代表正

Table 2 Summary of Categories of Vision-Language Representation Learning Models

表 2 视觉语言表征学习模型分类汇总

结构类型	目标	模型	预训练任务	预训练数据	下游任务/评测任务
单流	基于 BERT	VisualBERT[39]	ITM+MLM	COCO	VG+VR+VQA
		VL-BERT[40]	ITM+MLM	CC	VG+VR+VQA
	细粒度	SLViT[54]	ITC	COCO	VG
		SGEITL[56]	MNM（三元组掩码预测）	VG	VC+VR+VQA
	轻量化	ViLT[16]	ITM+MLM	COCO+VG+SBU+CC	ITR+VQA+VR
		SIMLA[58]	ITC+MLM+跨模态重构	COCO+VG+SBU+CC	ITR+VG
		GIT[62]	LM（图像到文本生成）	规模化图文字幕数据（>2 亿）	VC+VQA
		BUS[65]	ITC+ITM+MLM	COCO+VG	ITR+VQA+VC
		SimVLG[66]	LM	COCO+SBU+LAION400M+VG+CC	VC+VQA
		VL-Mamba[68]	ITC+多模态结构微调	图文混合数据（LLM 语料+图像）	VQA
	通用性	SimVLM[69]	PrefixLM（前缀语言建模）	ALIGN+C4	VQA+VR+VE+VC
		OFA[70]	MLM+MIM+ITM+VQA 等	COCO+VQA 等	VC+VQA+VG
		Uni-Perceiver[71]	MLM+MIM+ITM	ImageNet+COCO+CC+SBU+VG 等	ITR+VQA+VC
		BEiT-3[74]	MLM+MIM+跨模态 MOM	CC+COCO+VG+SBU 等	VQA+VR+VC+ITR
双流	基于 BERT	ViLBERT[38]	ITM+MLM+MIM	COCO+VG	ITR+VR+VE+VQA
		CLIP[41]	ITC	4 亿网络图文对	OCR（光学字符识别）
		SLIP[44]	ITC+MIM	CC+YFCC100M 等	OCR
		FLAVA[45]	ITC+ITM+MIM+MLM	COCO+SBU+VG+CC 等	ITR+VG+VQA
		CoCa[46]	ITC+MLM	JFT-3B+ALIGN	ITR+VG+VC
		LOUPE[47]	ITC	2 亿 4 000 万网页图文对	ITR+VG
		Florence[48]	MNM	9 亿图文对	ITR+VG+VQA
	细粒度	ALBEF[13]	ITC+ITM+MLM	COCO+VG+SBU+CC	ITR+VR+VQA
		TCL[14]	ITM+ITM+MLM	COCO+VG+SBU+CC	ITR+VQA+VR+VE
		X-VLM[18]	ITC+ITM+MLM+VG	COCO+VG+SBU+CC	ITR+VQA+VR+VG+VC
		FILIP[49]	ITC	CC+YFCC100M+3 亿图文对	ITR
		GLIP[50]	ITC+目标检测	2 700 万网页图文对	VG+短语检测
		RegionCLIP[52]	ITC	COCO+CC	VG
		UTA[53]	无掩码符号对齐	ImageNet	ITR+VR+VQA
		PyramidCLIP[55]	ITC	LAION400M	ITR+VG
		DANCE[57]	ITC+知识蒸馏	ConceptNet 知识增强的图文数据	ITR+VQA
	轻量化	METER[17]	MLM+MIM+ITM	COCO+VG+SBU+CC	ITR+VQA+VR+VE
		BridgeTower[63]	ITM+MLM	CC+SBU+COCO+VG	ITR+VE+VQA
		EfficientVLM[64]	ITM+MLM+知识蒸馏+剪枝	CC+COCO+SBU+VG	ITR+VQA+VR+VC
	通用性	BLIP[15]	ITC+ITM+LM	COCO+VG+CC+LAION400M	ITR+VC+VR+VQA
单双流切换	通用性	VLMo[73]	ITC+ITM+MLM	COCO+SBU+VG+CC	ITR+VQA+VR

常图片与正常文本，x_i'，x_t'分别代表对抗图片与对抗文本，β_s与β_c分别代表在分别扰动攻击与联合扰动攻击下图像扰动方向与文本扰动方向间的夹角. 在分别攻击中，2 个模态的对抗样本均基于正常图片-文本对分别生成；而在 Co-Attack 中，文本对抗样本x_t'是基于图像对抗样本x_i'的生成结果生成，不论是在单流融合模型还是双流对齐模型中，这种联合优化策略保证了文本和图像的对抗扰动是沿着同一最优方向，即互相最远离的方向进行，而分别扰动则会让 2 个扰动的方向产生一定冲突，攻击效果较差，为后来的视觉语言表征学习攻击提供了通用的参考.

Yin 等人[77] 在 2023 年提出了 VLAttack 框架，他

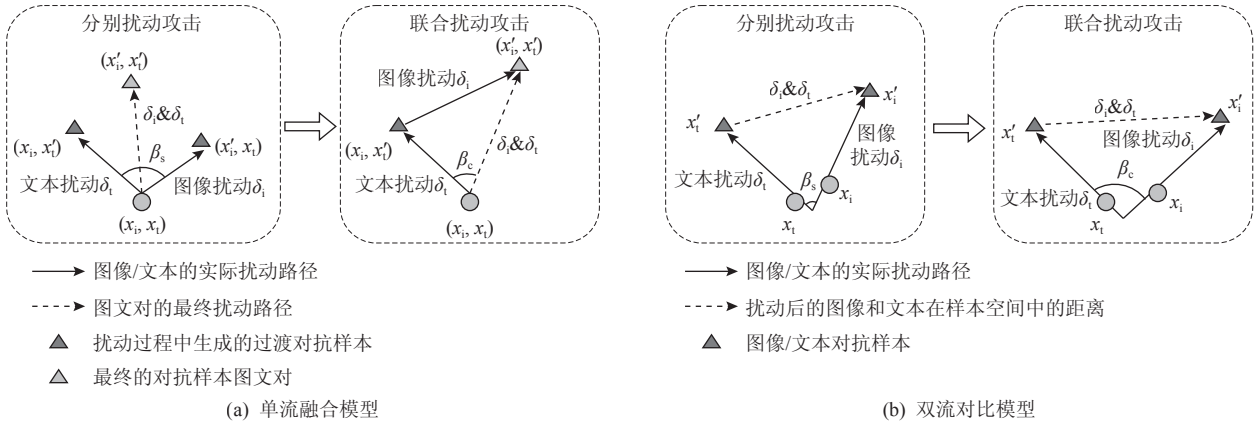


Fig. 5 Illustration of Sep-Attack and Co-Attack in embedding space

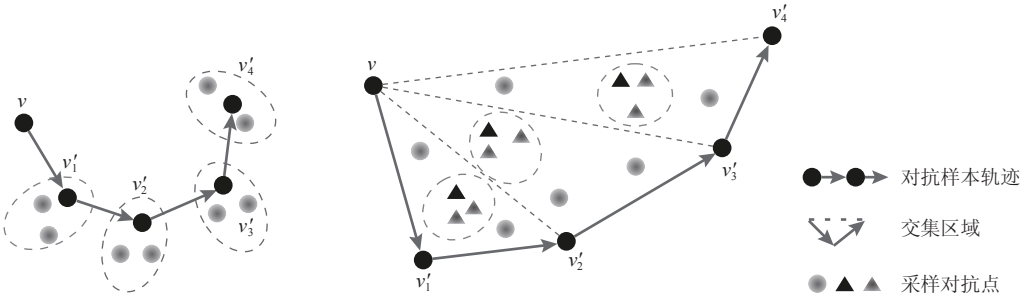
图 5 分别扰动攻击与联合扰动攻击在样本空间中的解释

们强调了“针对视觉语言表征学习的对抗样本攻击应当更关注其迁移性”. VLAttack 在更现实的黑盒场景中攻击多模态模型：攻击者只能访问预训练模型作为代理模型，目标模型是下游任务上微调过的黑盒模型. VLAttack 将生成对抗图像-文本对的过程分为 2 个阶段：首先在单模态层面对图像和文本分别施加扰动，使用块匹配扰动 BSA（block-wise similarity attack）来扰动图像使其通用特征失配，并采用现有文本攻击技术生成文本扰动，这个过程类似 Sep-Attack；然后在多模态层面迭代优化二者的匹配，Yin 等人[77]提出了迭代交叉搜索攻击（iterative cross-search attack, ICSA）算法，通过多次查询黑盒下游模型并微调对抗样本图文对，使其对抗组合在黑盒下游模型上的成功率更高. VLAttack 在 3 种预训练模型和 6 项下游任务上的实验进一步证明，同时攻击图像和文本并结合多模态优化可取得最高的攻击成功率.

Lu 等人[19]在针对视觉语言表征学习的对抗迁移攻击研究上更进一步，他们指出以往的攻击多聚焦于白盒攻击，而黑盒转移攻击，即攻击另一个完全不同的模型的成功率往往显著降低. 与 VLAttack 的攻击场景不同，这里的黑盒场景指代理模型与目标模型在预训练阶段也完全不同，如 Co-Attack 以 ALBEF[13]作为代理模型生成的对抗样本，在攻击 CLIP-vit[41]时往往成功率会降低 70% 甚至更多. 基于此，Lu 等人提出了集合级引导攻击（set-level guidance attack, SGA），SGA 在优化对抗样本时引入了一组语义多样的文本描述，即文本集合来共同指导，在对图像的优化过程中确保对抗图像在模型中的对齐能够远离整个文本集合. 此外，SGA 还结合了保持对齐的增广策略，在对抗样本迭代过程中对图像进行尺度、视角等变化，模拟多样化场景，同时用原始文本指导，以防止对抗噪声破坏正确的跨模态对应关系. SGA 在多个下游任务上显著提升了不同模型间的迁移攻击成功率，专注提升视觉语言表征学习迁移攻击的工作.

He 等人[78]在 2023 年提出了 SA-Attack，他们分析了已有研究，指出模态间交互和数据多样性是影响迁移攻击效果的两大因素. SA-Attack 在生成对抗样本时对输入进行自增强，对图像施加缩放、裁剪等多种数据增广，对文本也做如同义词替换等增广，从而生成一批多样化的对抗样本来共同更新扰动. 数据增广下生成的扰动不会过度拟合单一输入形态，因而在攻击其他模型时更具泛化性. 此外，Zheng 等人[79]提出了特征引导攻击（feature guidance attack, FGA），FGA 使用文本模态的表征向量来直接指导图像扰动，仅修改视觉输入就能生成对抗图像，在白盒与黑盒场景下对 VLP 模型均能造成显著误导. 该方法的一大优势是与许多现有单模态对抗策略兼容，可直接借鉴单模态图像对抗攻击领域的丰富成果. 进而，他们将文本攻击融入该框架，构建了 FGA-T（feature guidance with text attack），同步对图像和文本进行扰动，通过 2 种模态的交互作用进一步提升了黑盒迁移性能.

Gao 等人[20]受 SGA 和 SA-Attack 的启发，创新性地关注了对抗样本轨迹的交叉区域，提出了 DRA. SGA 与 DRA 的演化过程如图 6 所示，SGA 在每一步只从当前对抗点v'_i取样获得局部点，随后随机或按策略选择扰动方向，生成下一步对抗点v'_{i+1}. DRA 则在前一步对抗样本、当前对抗样本和原始样本构成的“交叉三角形”中采样，从而获取更多样的扰动方向，并让新的文本对抗样本远离整个“交叉三角形”，有效消除了扰动对单一模型的局部过拟合，从而提高了黑盒迁移攻击的成功率. Jia 等人[80]于 2024 年在 DRA 基础上提出了更完整的语义对齐对抗演化三角



(a) SGA (b) DRA

Fig. 6 The evolutionary processes of adversarial examples in SGA and DRA

图 6 SGA 和 DRA 的对抗样本演化过程

(semantic-aligned adversarial evolution triangle, SA-AET)攻击, 他们发现原始图像特征中存在的许多冗余非活跃维度主导了模型的相似度计算, 导致扰动容易对特定模型过拟合. 为此, 他们在 DRA 的基础上, 构建了一个基于大规模语义语料的子空间, 将图像和文本嵌入投影到该子空间中进行对抗优化, 使对抗扰动主要影响对语义判别有贡献的特征部分, 而非利用模型特有的冗余特征来欺骗模型, 进一步提升了对抗样本的迁移成功率. Zhang 等人[81] 在扩展数据多样性后更进一步提出了精细化对抗攻击(meticulous adversarial attack, MAA), MAA 提出了重新调整裁剪模块 RScrop(resizing and sliding crop)和多粒度相似性破坏(multi-granularity similarity disruption, MGSD). RScrop 通过对原始图像施加不同尺度和位置的裁剪变换, 生成一系列丰富多样的视角, 并在这些视角上优化初始对抗扰动, 与 SGA 的增广策略不同, RScrop 生成的所有增广图像都会直接用于优化扰动; MGSD 则在模型的多层次表示上施加约束, 在 VLP 模型的不同层级和组件中, 对抗样本的表征表示都尽可能远离原始样本, 以逐层放大扰动对模型的影响, 最大化对抗图像与原图在各个语义层面的嵌入差异. 这种“宽度+深度”的优化思路更进一步地提高了对抗样本的迁移性.

Wang 等人[82] 在 2024 年提出了 TMM(transferable multimodal attack)攻击框架, 系统性地利用了模型内部的跨模态注意力信息来同时破坏“对齐特征”(alignment features)与“差异特征”(divergent features), 从而获得更高的迁移成功率. TMM 首先利用模型的跨模态交叉注意力权重来定位图像与文本的关键对齐区域, 并在这些区域对图像像素进行定向扰动, 以有效破坏模型的“对齐特征”, 提升对抗样本的迁移能力. TMM 同时沿着与跨模态对齐向量正交的方向注入额外扰动, 使得对抗样本图像的嵌入同时具备更多模型间不共享的“噪声”成分, 破坏“差异特征”, 从而在多个模型上都能混淆对齐判断. TMM 在多项视觉语言表征学习下游任务上取得了 20% 以上的攻击成功率提升.

2024 年也产生了许多具有创新性的对抗攻击方案, 如 Fu 等人[83] 提出的语义不变的协同多模态交互攻击 CMI-Attack, 此前工作很少关注视觉语言表征学习模型中模态间的交互机制, 并且难以避免对抗文本对语义的直接破坏. CMI-Attack 提出了嵌入引导和交互增强 2 种策略. 前者在保证语义不变的前提下, 对文本输入进行嵌入层面的对抗扰动; 后者则利用模型中图文交互的梯度信息指导图像扰动的生成, 加强对图像和文本扰动的约束, 实现了对抗样本隐蔽性和泛化性的平衡. Ye 等人[84] 提出了一种图像攻击与文本防御相结合的联合优化框架 Mutual-Modality Attack, 利用预训练的 CLIP 模型作为跨模态语义对齐度量, 交替优化图像和文本. 首先, 对原始图像进行对抗扰动, 使其在 CLIP 的联合嵌入空间中偏离与原文本提示的语义对齐; 随后, 在文本模态上执行对应的“防御”, 通过更新文本提示(如调整提示词或其嵌入), 试图在受到扰动的图像嵌入空间中重新匹配回正确语义. 通过攻防迭代的过程, 最终得到的对抗图像包含了深度的跨模态语义偏移, 因而在不同模型上都能保持攻击有效性.

3.2 通用对抗扰动与对抗补丁攻击

在对抗样本攻击中, 通用对抗扰动(universal adversarial perturbation, UAP)和对抗补丁是一类重要的攻击方法, 它们旨在生成与输入无关的固定对抗扰动, 能叠加在任意图像上对表征学习模型产生误导. 与 3.1 节中强调模型间迁移效果的攻击方法不同, UAP 强调的是在数据集和输入上的通用性, 因而在真实世界场景下具有更大的威胁.

Zhou 等人[85] 在 2023 年提出了面向 CLIP 编码器

的下游无关对抗攻击框架 AdvCLIP, 在不依赖具体下游任务细节的情况下, 生成一个可叠加到任何自然图像上的补丁, 让使用 CLIP 编码器的所有下游任务都受到干扰. AdvCLIP 的创新点是构建了图文特征的拓扑关系图, 以捕捉目标样本与其邻居样本之间的相关位置, 然后采用拓扑偏差约束的生成式对抗网络[86](generative adversarial network, GAN)来产生能够降低图像嵌入与正确文本的相似度、扰乱特征空间分布的补丁, 从而实现无目标的通用对抗攻击. Kim 等人[87] 设计了双重通用对抗扰动(Doubly-UAP), 在不需要任何标注的情况下冻结模型参数, 仅针对视觉编码器内部注意力机制的脆弱组件(中后层的注意力值向量)进行优化, 在一次优化、不访问语言编码细节的情况下同时破坏图像表征提取和表征对齐, 针对以 CLIP 图像编码器为基础的模型时效果良好. Zhang 等人[88] 提出的 ETU(effective and transferable universal adversarial attack)攻击则从黑盒角度出发, 假设攻击者无法获取目标模型梯度, 通过一个代理模型集合来训练获取 UAP. ETU 在优化函数中综合考虑了扰动在全局和局部的作用, 在优化整幅扰动的整体结构的同时, 减少了不同局部扰动单元之间的相互干扰. ETU 使用了一种包括自混合和跨混合的数据增强手段 ScMix, 在不改变语义前提下增加多模态数据的多样性, 在多个主流视觉语言表征学习模型上都取得了良好的攻击效果. Huang 等人[89] 成功构建了有目标的通用对抗扰动攻击(targeted universal adversarial perturbation, TUAP), 同样采用了代理模型集合优化策略, 使任意带有补丁的输入图像在 CLIP 编码空间中都被误导为攻击者预先设定的某个目标文本的表示, 在攻击融合了 CLIP 图像编码器的大型视觉语言模型(如 OpenFlamingo[90]、LLaVA[91]、BLIP-2 [4], 等)时也能产生效果, 并且给补丁提供了全局噪声和小尺寸贴纸 2 种形态.

生成式对抗网络在通用对抗扰动方面也有非常广阔的应用. Fang 等人[92] 在 2024 年提出了对比训练的跨模态扰动生成器(contrastive perturbation generator with cross-modal conditions, C-PGC), C-PGC 在训练时采用“恶意”的对比学习策略: 精心构造“正样本”(不匹配的图文对)和“负样本”(匹配的图文对)来训练生成器, 使其输出的扰动能拉远原本匹配的图文表征距离并拉近不匹配对的表征距离, 即反转了 CLIP 的对比学习目标来破坏表征对齐. C-PGC 一经训练完毕, 即可生成一个对所有输入通用的扰动. Yang 等人[93] 对 C-PGC 的扰动求解进行了优化, 加速了通用扰动的生成过程, 提出了 DO-UAP(direct optimization-based UAP), 实验证明 DO-UAP 能够保持与 C-PGC 相近的攻击效果. Zhang 等人[94] 最新提出的 AnyAttack 利用了海量的未标注数据(LAION-400M[26])进行自监督预训练, 得到了一个无需人工指定攻击目标或标签, 就能将任何输入图像自动转换为攻击不同视觉语言表征学习模型所需的对抗样本的对抗生成模型, 实现了“任意到任意”的攻击框架.

对抗补丁相比 UAP 更加注重物理世界的可实现性. 如 Zhang 等人[95] 提出的 CAPatch 就是一种物理对抗补丁, 将打印的小贴纸贴在拍摄场景中, 仅覆盖图像约 5% 的区域, 能持续误导图像描述模型(如 BLIP 的下游模型)产生错误字幕, 在真实摄像头下, 视频实测攻击成功率达 73.1%. 更多的工作使用图像扩散模型[96] 来生成对抗补丁, 如 Kong 等人[97] 使用扩散模型生成逼真的对抗补丁, 借助模型的跨模态注意力引导将补丁放在影响最大的图像位置, 在白盒设置下的图文检索任务达到了 100% 的攻击成功率, Wang 等人[98] 的工作也类似, 但提升了补丁的可定制性, 适用于更细粒度的视觉推理和对话任务.

3.3 针对特定下游任务的攻击

在 2022 年以前, 视觉语言下游任务模型结构较简单时就有许多针对图文检索[99-100]、图像字幕[101-103]、视觉问答[104-105] 等任务模型的攻击. 但随着“预训练-微调”的视觉语言表征学习范式推广开来, 下游微调模型可能出现不同于预训练模型和以往模型的脆弱点. 在现实场景中, 攻击者也常常只需以这些部署模型为目标, 有较大的安全隐患. 因此, 近年来也有一小部分新的工作聚焦于攻击特定的下游任务模型, 评估实际应用中的风险.

针对 VQA 模型, Zhang 等人[106] 将迁移攻击与查询优化攻击相结合, 面向黑盒 VQA 模型提出了 VADS(Visuo-Adaptive DualStrike)攻击. VADS 首先在一个 VQA 代理模型上生成对抗扰动, 并且根据视觉内容自适应地调整攻击强度(即“Visuo-Adaptive”), 然后在目标黑盒模型上通过有限查询迭代细调扰动. VADS 的双管齐下策略在黑盒场景下实现了显著的攻击效果, 体现了在面对特定下游任务时结合多种攻击手段的重要性. Yin 等人[107] 在迭代优化过程中引入了大语言模型(large language model, LLM), 利用 LLM 生成一个将问题和正确答案融合的文本提示句, 并将其中的正确答案词随机掩盖. 然后将掩盖答案的文本与图像一起送入代理模型的掩码语言模型头, 计算模型在这些位置上预测原正确答案词的概率, 这

些被掩盖位置上模型预测出正确答案的概率分数为“抗恢复损失”,最小化该损失即可使得代理模型无法从扰动后的输入中“恢复”出原本的正确答案信息.他们同时还引入了延迟触发的文本攻击策略,在图像扰动到一定程度后再扰动文本,保持文本的含义.

对于图文检索排序模型,Li 等人[108]认为以往针对检索的攻击场景过于理想化,往往只要求让目标查询的结果不相关,他们提出了绝对-相对排序攻击(absolute relative ranking attack, ARRA).该攻击包含 2 种新的攻击目标:绝对排序攻击,即将指定候选文本推至检索列表的特定位置;相对排序攻击,即颠倒或改变指定候选之间的相对顺序.他们分别设计了绝对排序损失和相对排序损失,并以它们的加权和作为优化目标.ARRA 在黑盒有限查询场景下,使用基于评分反馈的优化方法寻找对抗扰动,成功地在真实在线服务 Huawei ImageSearch 上干预了返回结果顺序,充分说明了图文检索下游模型的实际应用风险.

3.4 小 结

在本节中,我们梳理了针对视觉语言表征学习的对抗样本攻击方法.模态联合攻击从模态间交互的角度,通过互相引导远离实现更好的攻击效果,并且采用丰富的增广策略避免对抗扰动陷入局部最优,实现更好的可迁移性.通用对抗扰动和对抗补丁倾向于获取语义级别的扰动,可以叠加到任何图像上完成对抗攻击,现有的方案通过联合多个代理模型优化或使用大量数据训练生成式对抗网络来实现这一点,但当前的工作仍然局限于图像本身.针对特定下游任务的攻击在模态联合攻击的基础上,结合针对特定下游任务的攻击策略定向优化扰动,或提出新的攻击目标来展现视觉语言表征学习模型的实际应用风险,我们在表 3 中给出了部分针对视觉语言表征学习的对抗样本攻击技术的总结,但这部分研究的关注度较低.

4 针对视觉语言表征学习的对抗样本防御

与针对视觉语言表征学习的对抗样本攻击技术的发展相比,对抗样本防御技术的发展则显得稍慢一些.虽然已有对抗攻击工作考虑了其突破防御方案的性能[82],但涉及的防御方案仍旧是针对单模态图像表征学习模型的净化方案如 BitRed[109]、JPEG 压缩[110]、NRP[111] 等,攻防研究之间缺少一定的交互性.因此,梳理当前已有的针对视觉语言表征学习的对抗防御技术对表征学习的安全性研究同样重要,本节将现有的防御技术分为三大类型:基于对抗训练的防御、基于模型结构增强的防御、基于对抗样本检测或净化的防御,分别在 4.1~4.3 节介绍,最后分析小结.

4.1 基于对抗训练的防御

对抗训练是在表征学习中最常用到的防御技术之一,在模型的训练样本里混入一些对抗样本,即可让表征学习模型具有一定的对抗鲁棒性,但同样可能造成一些模型精度上的损失.2022 年就有工作分析了 CLIP 模型具有一定鲁棒性的原因[112],证实了单纯增加训练数据数量(保持分布不变)并不能提高模型的有效鲁棒性(跨域泛化能力),而采用更多样化的训练数据分布则显著增强了 CLIP 在分布外数据上的表现.Nguyen 等人[113] 随后详细探讨了预训练数据集设计与 CLIP 模型鲁棒性之间的关系,Tu 等人[114]也扩展了扰动的形式、CLIP 的配置等变量,对 CLIP

Table 3 Summary of Adversarial Example Attack for Vision-Language Representation Learning Models

表 3 针对视觉语言表征学习模型的对抗样本攻击总结

攻击类型	主要研究内容和创新点	部分相关工作
模态联合攻击	系统研究了针对视觉语言表征学习的对抗攻击策略,发现联合优化策略的最优性	Co-Attack[10]
	提出集合级引导攻击 / 特征引导攻击和数据增广策略,显著提升了攻击的可迁移性	SGA[19]、FGA[79]
	优化引导攻击中的采样空间和优化空间,有效消除了扰动对单一模型的局部过拟合	DRA[20]、SA-AET[80]
	利用模型内部的跨模态注意力信息同时破坏“对齐特征”与“差异特征”,提升迁移性	TMM[82]
UAP 及对抗补丁	构建图文特征拓扑关系图,采用拓扑偏差约束的 GAN 来产生扰乱特征空间分布的补丁	AdvCLIP[85]
	类似迁移对抗攻击,通过代理模型来训练获取通用扰动,考虑扰动在全局和局部的作用	ETU[88]、TUAP[89]
	用大量数据“反向”训练一个对比表征学习模型,从而生成一个对所有输入通用的扰动	C-PGC[92]
	借助扩散模型生成会被模型提取为其他表征的对抗补丁,放在影响最大的位置误导模型	DiffPatch[98]
针对特定下游任务	针对 VQA 模型,通过迁移攻击结合有限查询优化扰动,兼具攻击效果和隐蔽性	VADS[106]
	针对 VQA 模型,结合 LLM 构建“抗恢复损失”阻碍模型输出正确答案	VQAttack[107]
	针对 ITR 模型,提出绝对排序攻击目标,设计 2 种损失加权优化,实现可控制的排序攻击	ARRA[108]

的鲁棒性和安全性进行了全面的考察，为后续改进视觉语言表征学习模型的鲁棒性指明了方向.

对视觉语言表征学习模型的对抗训练起初将重心放在了视觉表征鲁棒性上. 2023 年，Mao 等人[115]提出了文本引导的对比对抗训练 TeCoA，对视觉语言表征学习模型进行有监督对抗微调，证明了对抗训练可以提升 CLIP 的鲁棒性. TeCoA 利用 ImageNet[116]有标注数据，在固定的 ImageNet 类别文本嵌入下，对 CLIP 模型的视觉编码器进行 PGD 对抗训练，以增强 CLIP 在零样本图像分类下的鲁棒性. TeCoA 显著提高了 CLIP 抵御图像对抗扰动的能力，但也存在明显的局限，由于仅针对 ImageNet 类别进行训练，模型对非 ImageNet 类别的表征发生了扭曲，破坏了原有的零样本泛化能力，在其他下游任务上精度大幅下降. 此外，由于 TeCoA 的对抗训练损失基于余弦相似度，忽视了嵌入向量的幅值信息，也会影响后续多模态任务的性能. Schlarmann 等人[117]针对 TeCoA 的问题提出了无监督的对抗微调方案 FARE(fine-tuning for adversarially robust embeddings)，FARE 收集了没有任何类别标签的大规模图文对数据，并将这些数据中的图像扰动为对抗图像. 在预训练过程中，FARE 使用对抗图像和原始图像共同训练，并让对抗图像的表征仍接近原始图像的表征. FARE 在几乎不损失零样本精度情况下提升鲁棒性，保持了 CLIP 原有表征的语义丰富性. 但 FARE 仍局限于视觉模态的鲁棒性，且需要大量的额外对比数据，开销较大. Hossain 等人[118]提出的 Sim-CLIP 采用了双塔孪生架构——2 路相同的 CLIP 视觉编码器并行处理“干净图像”与“对抗图像”两种输入，且这 2 路网络共享全部权重. Sim-CLIP 在微调阶段的每次迭代中，仅对视觉编码器进行优化，利用 PGD 在当前视觉编码器上生成对抗图像，然后用余弦相似度损失函数来拉近 2 路编码器输出的特征向量，约束对抗扰动下输出保持与原图语义一致. Sim-CLIP 微调后的 CLIP 在增强鲁棒性的同时保持了语义细节，降低了 TeCoA 和 FARE 中对抗训练导致的语义偏移，同时训练效率更高.

另一方面，也有工作认为从语言表征角度引导视觉语言表征学习模型的对抗训练同样重要，Li 等人[119]提出了一种利用语言先验的锚点对抗训练框架 LAAT，针对零样本分类任务进行. LAAT 固定了 CLIP 的文本编码器，将类模版(如“A photo of {}”)输入文本编码器，将输出的表征作为类别锚点来指导视觉编码器训练. LAAT 采用了角度扩展算法优化锚点，避免类别锚点之间余弦相似度过高的问题，同时引入了一种对齐交叉熵损失，有效加大了对抗样本误导模型的难度，实现了零样本条件下的对抗训练，但需要预先定义类别锚点，限制了该框架的使用范围. 更直接的做法是将对抗扰动也引入文本模态，如 Zhou 等人[120]提出的多模态对比对抗训练(multimodal contrastive adversarial training, MMCoA)框架，MMCoA 在对比学习目标中加入了 2 种模态的对抗损失(没有对 2 个模态同时扰动，而是只扰动其中一个进行对抗训练，再扰动另一模态进行对抗训练)，通过同时对图像和文本进行对抗训练，模型不仅对视觉攻击更鲁棒，对文本侧攻击也更具抵抗力；证实了同时对 2 种模态做对抗训练的效果优于只对单个模态做对抗训练的效果. Waseda 等人[121]提出了“一对多”的对抗训练方案，他们发现仅用原始配对的“一对一”图文数据进行多模态对比对抗训练会导致过拟合，于是进一步利用图文数据的“一对多”对应关系，为每张图像生成多个等价描述或为每个文本检索多个相关图像，增广对抗训练集的多样性和对齐性，显著提高了表征在图文检索、视觉定位等任务下的对抗鲁棒性. Yu 等人[122]观察到对抗样本会导致 CLIP 的跨模态注意力偏移，因此提出通过注意力对齐来提升鲁棒性，他们设计了注意力精炼模块，将对抗图片经模型得到的文本引导注意力对齐到干净图片的注意力上，从而约束模型内部表征不因扰动而严重破坏；同时另设计了注意力约束模块确保模型在干净样本上保持原有性能，有效缓解了对抗训练精度下降的问题，但同样有额外的训练开销.

4.2 基于模型结构增强的防御

对抗训练方案需要引入大量对抗样本作为训练集，均面临着训练成本显著增加的问题，因此，少部分工作尝试从模型结构和学习方案上做出优化，以更少的成本增强表征的鲁棒性.

Cui 等人[123]在 2024 年对视觉语言表征学习模型的一项脆弱性研究中就探讨了简单的提示词调整对增强鲁棒性的作用. 他们发现为模型提供图像上下文信息可以有效缓解对抗攻击的影响，在询问图像内容时附加一些有关场景的描述或分解为多条子查询，可显著降低模型被误导的概率，这种无需改动模型参数的防御思路代价很低，具有很大的实用价值. Li 等人[124]于 2024 年提出了一种对抗提示微调(adversarial prompt tuning, APT)方案，他们发现对抗攻击和防御的效果对输入的文本提示词非常敏感，即对抗样本的攻击强度依赖于用于生成这些样本的文本提示. 因此，他们提出冻结预训练模型权重，并且在训练

过程中额外引入对抗样本来学习一个用于加强鲁棒性的可学习文本前缀. 虽然 APT 同样需要在训练过程中生成对抗样本, 但 APT 仅增加了少量参数, 且在使用少量数据时也能有效提高模型对抗攻击的抵抗力, 生成的提示词也具有一定的泛化性能. Zhang 等人[125] 随后提出了 AdvPT, 采用了不同的训练范式进一步降低 APT 的训练成本: 首先用类似 TRADES[126] 方法中的 Kullback-Leibler 散度(KL 散度)作为损失度量, 一次性生成所有训练样本的对抗样本, 而不是简单地以“分类错误”作为对抗样本的目标, 以这种方式生成的对抗样本的表征能最大程度地远离原图像的表征; 随后的对抗提示微调过程就不再实时生成, 而是从对抗表征库中提取, 显著改善了训练效率. 也有工作[127] 将 APT 扩展到了 2 个模态, 共同提升鲁棒性.

Bhagwatkar 等人[128] 从模型结构角度系统性地研究了如何提高视觉语言模型抵御图像对抗攻击的鲁棒性. 他们发现采用在多元数据集上训练的视觉编码器(如 CLIP 或 SigLIP[129])有助于提升模型对简单攻击(如 FGSM)的鲁棒性, 相比之下, 仅在单一数据集上训练的编码器(如纯 ImageNet 预训练的 ViT)更容易被攻击, 但这种优势在更复杂的迭代攻击下几乎消失; 他们同时发现, 集成多个视觉编码器对鲁棒性的提升非常有限. 此外, 针对 APT 方案, 他们发现如果让攻击算法在生成扰动时也接收提示词, 模型的鲁棒准确率会再次下降, 说明熟悉防御策略的自适应攻击仍能攻破这些基于提示的防御, 强调了提示工程的局限性.

2025 年也有一些新的工作, Han 等人[130] 提出了 AlignCLIP, 通过改进注意力分布和标签表示来增强 CLIP 模型对分布变化的适应性. AlignCLIP 在模型结构上引入了注意力对齐损失, 鼓励多头注意力在各层次聚焦于图像中的关键对象而非背景, 从而减少注意力偏差; 同时采用了语义标签平滑策略, 利用类别文本的相似性信息, 约束模型对相近类别做出更平滑的判别, 提升了模型表示在开放环境下的稳定性. AlignCLIP 主要针对分布偏移的问题设计和开放集问题设计, 能够在一定程度上解决噪声问题, 为后续工作提供了参考.

4.3 基于对抗样本检测或净化的防御

针对视觉语言表征学习模型的对抗样本检测或净化方案较少, 由于检测和净化技术只是基于数据本身进行, 因此针对单模态的检测和净化策略能够直接用于多模态, 现有的对抗攻击工作在考虑防御净化方案时也基本是将 2 个模态的净化方案同时使用[82]. 事实上, 已经有相当一部分的工作将视觉语言表征学习模型获取的细粒度表征用于单模态的对抗样本和净化, 取得了不错的效果[131], 这些工作也能应用于视觉语言表征学习模型本身. 下面介绍 3 个专为视觉语言表征学习模型设计的对抗样本检测或净化工作.

Zhang 等人[132] 针对视觉语言问答下游模型提出了一种对抗样本探针检测方法 PIP, 在输入一张待测图像及其问题的基础上, 额外再询问模型一个无关的随机探针问题(例如“这张图片中有钟吗?”). 对于干净图像, 模型在回答无关探针问句时, 其注意力分布会呈现出一种稳定的模式, 通常不会关注图像内容; 但若图像是对抗样本, 则模型可能出现异常的注意力权重偏移. PIP 只需一次额外前向推理即可完成检测, 效率较高, 且能达到 90% 以上的准确率, 为理解视觉语言表征学习模型内部机制提供了新思路.

Fares 等人[133] 提出了一种面向通用视觉语言模型的高效检测方法 MirrorCheck(镜子检查), 他们将图像输入待测的视觉语言模型生成文本描述, 然后另起一个“镜子”生成模型反向生成图像, 比较原始输入图像与生成图像在 CLIP 特征空间的相似度, 若相似度较低则说明原始输入可能被对抗扰动篡改. 该方法与模型架构无关, 通用性强, 但推理成本较高.

Huang 等人[134] 构建了一个大型多样对抗图像数据 RADAR, 并提出了一种利用模型隐藏层单一向量实现高效检测的方案 NEARSIDE(N-time embedding based adversarial image detection). 他们发现, 对于一张被成功攻击的图像, 多模态模型内部某些方向的激活变化尤为显著. NEARSIDE 通过在模型隐藏状态中提取一个“攻击方向”向量来捕捉这种变化, 因此只需在推理时计算待测输入在该方向上的投影即可判断其是否为对抗样本. 实验证明 NEARSIDE 能在 LLaVA 等视觉语言模型上有效区分干净图像和对抗图像, 且这个“攻击方向”在不同模型间具有一定的可迁移性. 但该工作主要针对视觉模态的攻击, 对文本侧攻击检测尚未涉及.

在对抗样本净化方面, Xing 等人[135] 提出了测试时反击(test time counter attack)策略, 在推理阶段对每张输入图像执行若干步的小幅“对抗扰动”, 这个“对抗扰动”是反向的, 即使对抗样本朝着正确分类方向移动, 让模型主动攻击输入来抵消潜在对抗噪声. 这是针对 CLIP 的纯推理防御方法, 实验表明, 该方法对干净精度的影响并不显著, 且可与任何预训练或微调鲁棒模型正交结合, 能进一步提高鲁棒性. 其存

在的局限是需手动设置反击步数和强度，对不同攻击可能需要调参，同时过强的反扰动也可能引入新误差.

4.4 小 结

在本节中，我们梳理了针对视觉语言表征学习的对抗样本防御方法. 早期的对抗训练将重心放在视觉模态上，优化视觉编码器的鲁棒性，却造成了一定的语义偏移和表征泛化能力的降低. 后来的部分工作尝试了用语言表征引导对抗训练，取得了不错的效果，但依旧无法避免训练成本过高的问题. 基于模型结构增强的防御通过提示词微调、改进模型结构和注意力分布等方式，在成本较低的情况下增强模型的抗噪声能力，但这类防御方案在白盒场景下其实表现较差，且需要复杂的调参过程. 对抗样本检测和净化技术在视觉语言表征学习方面发展较慢，但也有工作尝试借助模态注意力融合机制的独有特性，在测试时完成对抗样本检测. 总体来说，针对视觉语言表征学习的对抗样本防御方法依旧处于发展初期，现有研究数量尚未全面覆盖所有的攻击类型与防御策略，该领域具有较大的安全隐患. 我们在表 4 中给出了部分针对视觉语言表征学习的对抗样本防御技术的总结.

Table 4 Summary of Adversarial Example Defence for Vision-Language Representation Learning Models

表 4 针对视觉语言表征学习模型的对抗样本防御总结

防御类型	主要研究内容和创新点	部分相关工作
对抗训练	在 CLIP 的视觉编码器上进行对抗训练，约束对抗样本和干净样本表征一致	TeCoA[115]、Sim-CLIP[118]
	将文本表征采用角度扩展算法优化后作为锚点，加大对抗扰动误导模型的难度	LAAT[119]
	将文本对抗样本也纳入考虑，加入 2 种模态的对抗损失进行联合优化	MMCoA[120]、MAT[121]
模型结构增强	利用对抗样本对文本提示词的高敏感性，训练一个加强鲁棒性的提示词前缀	APT[124]、AdvPT[125]
	改进注意力对齐损失，鼓励模型聚焦图像关键部分，并用语义标签平滑策略约束模型输出	AlignCLIP[130]
对抗样本检测或净化	针对 VQA 模型，通过无关探针问题分析其注意力分布情况，从而识别对抗样本	PIP[132]
	将图像生成模型作为“镜子”反向生成图像样本，与原始图像输入在表征空间中比较判断	MirrorCheck[133]
	通过分析模型隐藏层单一向量的激活变化投影来区分干净图像和对抗图像	NEARSIDE[134]
	提出了测试时反击策略，在推理阶段对每张图像执行反向“对抗扰动”抵消潜在对抗噪声	CounterAttack[135]

5 总结与展望

5.1 本文总结

本文首先介绍了视觉语言表征学习和对抗样本攻击的基本概念，包括表征学习模型的基本架构和对抗样本的相关概念. 随后，我们从细粒度、轻量化、通用性 3 个发展方向出发，回顾了近年来视觉语言表征学习模型，详细分析了它们的技术特点.

在对抗攻防技术方面，本文对现有的对抗攻击和防御方案按照时间顺序进行了较全面的分类回顾，在上述工作的基础上，本文对比了部分攻击方案的异同点和优劣，并特别对视觉语言表征学习安全性研究的现存问题与前沿方向进行了探讨和研究展望. 与此同时，可以识别这些方法的优劣势，如表 5 所示.

5.2 未来展望

当前，视觉语言表征学习技术已经拥有较强的细粒度跨模态对齐能力，也因此被广泛应用于移动

Table 5 Comparison of Adversarial Examples Attack and Defense Technologies on Vision-Language Representation Learning Models

表 5 在视觉语言表征学习模型中的对抗样本攻防技术对比

攻击方法	优势	劣势
模态联合攻击	充分利用模态交互，攻击成功率高，有一定的模型和任务迁移能力	实施流程复杂，查询和计算代价较高
通用对抗扰动与对抗补丁	可以攻击任意图像，可用于物理场景，一次生成即可长期使用	模态交互不足，在模型和任务间的迁移效果有限
针对特定下游任务	定制攻击策略，隐蔽性和攻击成功率高	通用性差，只作用于特定模型的特定任务
防御方法	**优势**	**劣势**
对抗训练	实施简单，直接有效，适用范围广	训练成本高昂，可能降低模型性能
模型结构增强	训练代价小，通常能保持模型性能	实施和设置过程复杂，在白盒场景下能被针对性绕过
对抗样本检测或净化	兼容性最强，能与其它防御方案结合	当前仅集中于图像检测，且存在误判和漏报风险

互联网、法律、工业甚至医学等领域. 但一方面, 由于视觉语言表征学习模型庞大的参数量和较高的训练成本, 其依旧难以被广泛部署甚至产生明显的性能损失, 因此, 轻量化的模型结构和高效的训练方式将成为未来研究的主要趋势之一; 另一方面, 各大领域对视觉语言表征学习技术的需求和下游数据集的缺失之间产生了显著矛盾, 医学、法律等领域缺乏系统可用的下游数据集, 因此, 构建多任务的统一建模, 减少针对单一任务定制模型的需求, 提升表征的通用性也是未来研究的一个重要方向.

在对抗样本攻击技术方面, 现有的模态协同攻击和 UAP 充分关注了模态表征的注意力融合机制, 实现了较高的白盒攻击成功率, 但在攻击的迁移性方面仍有提升空间. 现有的迁移攻击技术仍然会降低 20% 左右的攻击成功率, 且在下游任务间迁移, 尤其攻击大型视觉语言问答模型时表现不够理想[82], 因此, 针对视觉语言表征学习模型的对抗攻击, 未来可以向着语义级别的攻击发展, 以进一步挖掘视觉模态和语言模态的交互逻辑, 开发更具有威胁性的攻击技术, 并同时推动对表征学习模型的注意力机制的研究.

在安全防御技术方面, 现有的防御技术发展较慢, 也因此面临着诸多问题: 1)大部分对抗防御技术仅在 CLIP 甚至是 CLIP 的视觉编码器上展开, 一方面对语言模态的关注度较低, 另一方面对已经广泛推广的 ALBEF 等新模型的防御研究和部署不足. 2)部分防御技术的通用性较差, 如 APT 方案主要应用于问答模型, 在其他如检索模型、定位模型和描述模型上具有一定的应用难度. 3)面临着巨大的额外训练开销与使用开销, 对模型结构和训练方式的改变均会加重模型的训练负担, 现有的对抗样本检测与净化技术则需要在使用时付出额外的推理时间, 且面临着误判的风险. 因此, 防御技术未来的发展方向应同样关注视觉语言表征的注意力融合机制, 当攻击技术普遍攻击机制误导模型判断时, 防御技术也应在表征学习模型的注意力层关注多模态之间的融合交互过程, 并从模型训练角度或模型结构优化角度提升这个过程的鲁棒性, 实现攻击技术与防御技术的对抗提升.

此外, 针对视觉语言表征学习模型的可解释性研究也应成为未来发展的重要组成部分. 随着多模态模型走向并部署在医学、法律等对伦理道德要求较高的领域, 其决策依据的可解释能力将成为实际应用中的重要需求之一. 虽然在单模态视觉表征学习领域应用广泛的类激活映射(class activation mapping, CAM)[136] 和梯度加权类激活映射 Grad-CAM[137] 可以迁移到多模态表征学习模型的视觉编码器中, 但仍欠缺语言模态和专门针对多模态表征学习的直观解释方法. 良好的可解释性机制不仅可以使模型的视觉和语言推理过程透明化, 帮助开发者及时检测并修正模型异常行为, 指导攻击技术与防御技术的研究; 同时也能帮助用户理解模型如何得出结论, 避免因对抗性攻击带来的信任风险.

最后, 多模态视觉语言表征学习仍缺乏标准的数据和评测体系. 视觉语言表征学习模型往往只需要丰富的图文对进行训练, 这就导致了不同研究往往在各自的数据集组合和下游任务组合上测试, 缺乏直接可比性, 如表 2 所示; 在对抗样本攻防评估方面, 虽然有统一的攻击成功率作为基本评价指标, 但针对下游任务的对抗鲁棒性评估尚未形成统一通用的标准和完整的评价体系, 评测结果碎片化且难以复现. 此外, 当前评测体系缺乏对威胁场景的覆盖, 大多数实验仅在理想条件下考察模型, 对模型在对抗攻击、分布外数据等极端场景下的表现关注不足. 与此同时, 多任务评估也相对缺失, 现有研究通常仅评估模型在单一或少数任务上的性能, 缺少对模型在多任务环境下综合能力的系统性考查.

为此, 未来的研究与实践需要在评测体系上作出改进和标准化. 一方面, 应构建统一的测试集和评测基准, 将主要的数据集和典型下游任务纳入同一平台, 提供公平可比的测试环境, 从而减少不同研究间评测方案差异对结果的影响. 另一方面, 需要制定多维度的评估标准, 即不仅评测模型在各项任务上的准确性和效果, 还应考察模型在对抗样本等威胁场景下的鲁棒性, 以及模型决策过程和输出结果的可解释性等因素, 从多个角度全面衡量模型能力. 通过建立统一完善的评测体系, 涵盖标准数据集、统一指标、多任务场景以及对抗鲁棒性和可解释性评价, 研究者将能够更客观、公平地比较不同模型的性能与安全性. 我们殷切期待未来研究能够建立完善的多模态安全评测体系, 并推动开发出更准确、更安全、更实用的多模态视觉语言表征学习模型.

作者贡献声明: 曾诚负责论文的总体规划、主要内容的研究与调研、论文撰写; 葛云洁负责论文结构的梳理和修改润色; 赵令辰和王骞对论文提出指导意见, 并在论文撰写过程中提供支持.

参 考 文 献

[1] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798−1828

[2] Guo Wenzhong, Wang Jianwen, Wang Shiping. Deep multimodal representation learning: A survey[J]. IEEE Access, 2019, 7: 63373−63394

[3] Jia Chao, Yang Yinfei, Ye Xia, et al. Scaling up visual and vision-language representation learning with noisy text supervision [C] // Proc of the 38th Int Conf on Machine Learning (ICML). New York: PMLR, 2021: 4904–4916

[4] Li Junnan, Li Dongxu, Savarese S, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models [C] // Proc of the 40th Int Conf on Machine Learning. New York: PMLR, 2023: 19730–19742

[5] Li Sui. Research on diversified image description generation methods based on multimodal representation learning [D]. Beijing: China University of Mining and Technology, 2024 (in Chinese)
(李穗. 基于多模态表征学习的多样化图像描述生成方法研究[D]. 北京: 中国矿业大学, 2024)

[6] Zhang Mengrui. Research on multimodal sentiment analysis based on representation learning and modality fusion [D]. Hangzhou: Hangzhou Dianzi University, 2024 (in Chinese)
(张梦睿. 基于表征学习及模态融合的多模态情感分析研究[D]. 杭州: 杭州电子科技大学, 2024)

[7] Yao Wenfang, Yin Kejing, Cheung W K, et al. DrFuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency [C] // Proc of the AAAI Conf on Artificial Intelligence. Washington, DC: AAAI, 2024: 16416–16424

[8] Feng Yuchao, Feng Zhen, Wei Hua, et al. Multimodal-XAD: Explainable autonomous driving based on multimodal environment descriptions[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(12): 19469−19481

[9] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [J]. arXiv preprint, arXiv: 1312.6199, 2013

[10] Zhang Jiaming, Yi Qi, Sang Jitao. Towards adversarial attack on vision-language pre-training models [C] // Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022: 5005–5013

[11] Du Pengfei, Li Xiaoyong, Gao Yali. A survey of multimodal visual-language representation learning[J]. Journal of Software, 2021, 32(2): 327−348 (in Chinese)
(杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述[J]. 软件学报, 2021, 32(2): 327−348)

[12] Yin Jiong, Zhang Zhedong, Gao Yuhan, et al. Survey on vision-language pre-training[J]. Journal of Software, 2023, 34(5): 2000−2023 (in Chinese)
(殷炯, 张哲东, 高宇涵, 等. 视觉语言预训练综述[J]. 软件学报, 2023, 34(5): 2000−2023)

[13] Li Junnan, Selvaraju R R, Gotmare A D, et al. Align before fuse: Vision and language representation learning with momentum distillation [C] // Advances in Neural Information Processing Systems. Red Hook, NY: Curran Associates, Inc., 2021: 9694–9705

[14] Yang Jinyu, Duan Jiali, Tran Son, et al. Vision-language pre-training with triple contrastive learning [C] // Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 15671–15680

[15] Li Junnan, Li Dongxu, Xiong Caiming, et al. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation [J]. arXiv preprint, arXiv: 2201.12086, 2022

[16] Kim Wonjae, Son Bokyung, Kim Ildoo. ViLT: Vision-and-language Transformer without convolution or region supervision [C] // Proc of the 38th Int Conf on Machine Learning. Virtual: PMLR, 2021: 5583–5594

[17] Dou Ziyi, Xu Yichong, Gan Zhe, et al. An empirical study of training end-to-end vision-and-language Transformers [C] // Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 18166–18176

[18] Zeng Yan, Zhang Xinsong, Li Hang, et al. Multi-grained vision language pre-training: Aligning texts with visual concepts [C] // Proc of the 39th Int Conf on Machine Learning. New York: PMLR, 2022: 25994–26009

[19] Lu Dong, Wang Zhiqiang, Wang Teng, et al. Set-level guidance attack: boosting adversarial transferability of vision-language pre-training models [C] // Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 102–111

[20] Gao Sensen, Jia Xiaojun, Ren Xuhong, et al. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory [J]. arXiv preprint, arXiv: 2403.12445, 2024

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] // Advances in Neural Information Processing Systems 30. Montreal, Canada: Curran Associates, Inc. , 2017: 5998–6008

[22] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [C/OL] //Proc of Int Conf on Learning Representations. Appleton, WI: OpenReview.net, 2021[2025-04-15]. https://openreview.net/forum?id=YicbFdNTTy

[23] Lin Tsung-Yi, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[C] //Proc of the 13th European Conf on Computer Vision (ECCV). Berlin: Springer, 2014: 740−755

[24] Sharma P, Ding Nan, Goodman S, et al. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning[C] //Proc of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018). Stroudsburg, PA: ACL, 2018: 2556−2565

[25] Ordonez V, Kulkarni G, Berg T L. Im2Text: Describing images using 1 million captioned photographs[C] //Proc of the 24th Int Conf on Neural Information Processing Systems (NeurIPS). Red Hook, NY: Curran Associates, 2011: 1143−1151

[26] Schuhmann C, Vencu R, Beaumont R, et al. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs[J]. arXiv preprint, arXiv: 2111.02114, 2021

[27] Krishna R, Zhu Yuke, Groth O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annota-

tions[J]. International Journal of Computer Vision, 2017, 123(1): 32−73

[28] Antol S, Agrawal A, Lu Jiasen, et al. VQA: Visual question answering[C] //Proc of the 2015 IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2015: 2425−2433

[29] Thomee B, Shamma D A, Friedland G, et al. YFCC100M: The new data in multimedia research[J]. Communications of the ACM, 2016, 59(2): 64−73

[30] Zhai Xiaohua, Kolesnikov A, Houlsby N, et al. Scaling vision transformers[J]. arXiv preprint, arXiv: 2106.04560, 2021

[31] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [J]. arXiv preprint, arXiv:1412.6572, 2014

[32] Papernot N, McDaniel P D, Goodfellow I J, et al. Practical black-box attacks against machine learning [C] // Proc of the 2017 ACM on Asia Conf on Computer and Communications Security. New York: ACM, 2017: 506–519

[33] Wu Lei, Zhu Zhanxing, Tai Cheng, et al. Understanding and enhancing the transferability of adversarial examples [J]. arXiv preprint, arXiv: 1802.09707, 2018

[34] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks [C/OL] // Proc of the 6th Int Conf on Learning Representations. Appleton, WI: OpenReview. net, 2018[2025-04-15]. https://openreview.net/forum?id=rJzIBfZAb

[35] Carlini N, Wagner D. Towards evaluating the robustness of neural networks [C] // Proc of 2017 IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2017: 39–57

[36] Li Linyang, Ma Ruotian, Guo Qipeng, et al. BERT-ATTACK: Adversarial attack against BERT using BERT [C] // Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 6193–6202

[37] Devlin J, Chang Mingwei, Lee Kenton, et al. BERT: Pre-training of deep bidirectional Transformers for language understanding [C] // Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsburg, PA: ACL, 2019: 4171–4186

[38] Lu Jiasen, Batra D, Parikh D, et al. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks [C] // Advances in Neural Information Processing Systems 32. Red Hook, NY: Curran Associates, 2019: 13–23

[39] Liunian H L, Yatskar M, Yin Da, et al. VisualBERT: A simple and performant baseline for vision and language [J]. arXiv preprint, arXiv: 1908.03557, 2019

[40] Su Weijie, Zhu Xizhou, Cao Yue, et al. VL-BERT: Pre-training of generic visual-linguistic representations [C/OL] // Proc of the 8th Int Conf on Learning Representations. Appleton, WI: OpenReview. net, 2020[2025-04-15]. https://openreview.net/forum?id=SygXPaEYvH

[41] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C] // Proc of the 38th Int Conf on Machine Learning. New York: PMLR, 2021: 8748–8763

[42] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C] // Proc of the 2016 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770–778

[43] Crowson K, Biderman S, Kornis D, et al. VQGAN-CLIP: Open domain image generation and editing with natural language guidance [J]. arXiv preprint, arXiv: 2204.08583, 2022

[44] Mu N, Kirillov A, Wagner D, et al. SLIP: Self-supervision meets language-image pre-training [G] // LNCS 13686: Proc of the 26th European Conf on Computer Vision. Cham: Springer, 2022: 529–544

[45] Singh A, Hu R, Goswami V, et al. FLAVA: A foundational language and vision alignment model [C] // Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 15617–15629

[46] Yu Jiahui, Wang Zirui, Vasudevan V, et al. CoCa: Contrastive captioners are image-text foundation models [J]. arXiv preprint, arXiv: 2205.01917, 2022

[47] Li Juncheng, He Xin, Wei Longhui, et al. Fine-grained semantically aligned vision-language pre-training [J]. arXiv preprint, arXiv: 2208.02515, 2022

[48] Lu Yuan, Chen Dongdong, Chen Yiling, et al. Florence: A new foundation model for computer vision [J]. arXiv preprint, arXiv: 2111.11432, 2021

[49] Yao Lewei, Huang Runhui, Hou Lu, et al. Fine-grained interactive vision-language pre-training [C] // Advances in Neural Information Processing Systems 35. Red Hook, NY: Curran Associates, 2022: 7290−7303

[50] Liunian H L, Zhang Pengchuan, Zhang Haotian, et al. Grounded language-image pre-training [C] // Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 10965−10975

[51] Zhang Haotian, Zhang Pengchuan, Hu Xiaowei, et al. GLIPv2: Unifying localization and VL understanding [C] // Advances in Neural Information Processing Systems 35. Red Hook, NY: Curran Associates, 2022: 36067−36080

[52] Zhong Yiwu, Yang Jianwei, Zhang Pengchuan, et al. RegionCLIP: Region-based language-image pretraining [C] // Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. New Orleans, LA: IEEE, 2022: 16793−16803

[53] Liu Jihao, Zheng Jinliang, Liu Boxiao, et al. Enhancing vision-language model with unmasked token alignment[J/OL]. Transactions on Machine Learning Research, 2024[2025-04-15]. https://openreview.net/forum?id=JkFEVbW6wE

[54] Ouyang Shuyi, Wang Hongyi, Xie Shiao, et al. SLViT: Scale-wise language-guided vision Transformer for referring image segmentation [C] // Proc of the 32nd Int Joint Conf on Artificial Intelligence (IJCAI). Red Hook, NY: Curran Associates, 2023: 1294−1302

[55] Gao Yuting, Liu Jinfeng, Xu Zihan, et al. PyramidCLIP: Hierarchical feature alignment for vision-language model pretraining [C] // Advances in Neural Information Processing Systems 35 (NeurIPS 2022). Red Hook, NY: Curran Associates, 2022: 35959−35970

[56] Wang Zhecan, You Haoxuan, Li Liunian Harold, et al. SGEITL: Scene graph enhanced image-text learning for visual commonsense reasoning [C] // Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 5914−5922

[57] Ye Shuquan, Xie Yujia, Chen Dongdong, et al. Improving commonsense in vision-language models via knowledge graph

riddles［C］// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 2634−2645

［58］ Khan Z, Vijay Kumar B G, Yu Xiang, et al. Single-stream multi-level alignment for vision-language pretraining［G］// LNCS 13696: Proc of the European Conf on Computer Vision (ECCV 2022). Cham: Springer, 2022: 735−751

［59］ Liu Ze, Lin Yutong, Cao Yue, et al. Swin Transformer: Hierarchical vision Transformer using shifted windows［C］// Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 10012–10022

［60］ Liu Yinhan, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach［C］// Proc of the 36th Int Conf on Machine Learning. New York: PMLR, 2019: 1831–1840

［61］ He Pengcheng, Liu Xiaodong, Gao Jianfeng, et al. DeBERTa: Decoding-enhanced BERT with disentangled attention［C/OL］// Proc of the Int Conf on Learning Representations. Appleton, WI: OpenReview.net, 2021［2025-04-15］. https://openreview.net/forum?id=XPZIaotutsD

［62］ Wang Jianfeng, Yang Zhengyuan, Hu Xiaowei, et al. GIT: A generative image-to-text Transformer for vision and language［J］. arXiv preprint, arXiv: 2205.14100, 2022

［63］ Xu Xiao, Wu Chenfei, Che Wanxiang, et al. BridgeTower: Building bridges between encoders in vision-language representation learning［C］// Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 10637–10647

［64］ Wang Tiannan, Zhou Wangchunshu, Zeng Yan, et al. EfficientVLM: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning［C］// Findings of the Association for Computational Linguistics: ACL 2023. Stroudsburg, PA: ACL, 2023: 13899–13913

［65］ Jiang Chaoya, Xu Haiyang, Ye Wei, et al. BUS: Efficient and effective vision-language pre-training with bottom-up patch summarization［C］// Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 2900–2910

［66］ Jian Yiren, Liu Tingkai, Tao Yunzhe, et al. SimVLG: Simple and efficient pretraining of visual language generative models［J］. arXiv preprint, arXiv: 2310.03291, 2023

［67］ Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces［J］. arXiv preprint, arXiv: 2312.00752, 2023

［68］ Qiao Yanyuan, Yu Zheng, Zhao Zijia, et al. VL-Mamba: Exploring state space models for multimodal learning［C］// Proc of the 4th NeurIPS Efficient Natural Language and Speech Processing Workshop. New York: PMLR, 2024: 102–113

［69］ Wang Zirui, Yu Jiahui, Yu Adamswei, et al. SimVLM: Simple visual language model pretraining with weak supervision［C/OL］// Proc of the Int Conf on Learning Representations. Appleton, WI: OpenReview. net, 2022［2025-04-15］. https://openreview.net/pdf?id=GUrhfTuf_3

［70］ Wang Peng, Yang An, Men Rui, et al. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework［C］// Proc of the 39th Int Conf on Machine Learning. New York: PMLR, 2022: 23318–23340

［71］ Zhu Xizhou, Zhu Jinguo, Li Hao, et al. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks［C］// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 16804–16815

［72］ Li Hao, Zhu Jinguo, Jiang Xiaohu, et al. Uni-Perceiver v2: A generalist model for large-scale vision and vision-language tasks［C］// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Vancouver, BC: IEEE/CVF, 2023: 2691–2700

［73］ Bao Hangbo, Wang Wenhui, Dong Li, et al. VLMo: Unified vision-language pre-training with mixture-of-modality-experts［C］// Advances in Neural Information Processing Systems 35. Red Hook, NY: Curran Associates, 2022: 32897−32912

［74］ Wang Wenhui, Bao Hangbo, Dong Li, et al. Image as a foreign language: BEiT pretraining for all vision and vision-language tasks［C］// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 19175–19186

［75］ Van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete represent-ation learning［C］// Advances in Neural Information Processing Systems 30. Red Hook, NY: Curran Associates, 2017: 6306–6315

［76］ Vishwamitra N, Hu Hongxin, Zhao Ziming, et al. Understanding and measuring robustness of multimodal learning［J］. arXiv preprint, arXiv: 2112.12792, 2021

［77］ Yin Ziyi, Ye Muchao, Zhang Tianrong, et al. VLAttack: Multimodal adversarial attacks on vision-language tasks via pre-trained models［C］// Advances in Neural Information Processing Systems 36 (NeurIPS 2023). Red Hook, NY: Curran Associates, 2023: 59062–59094

［78］ He Bangyan, Jia Xiaojun, Liang Siyuan, et al. SA-Attack: Improving adversarial transferability of vision-language pre-training models via self-augmentation［J］. arXiv preprint, arXiv: 2312.04913, 2023

［79］ Zheng Haonan, Deng Xinyang, Jiang Wen, et al. A unified understanding of adversarial vulnerability regarding unimodal models and vision-language pre-training models［J］. arXiv preprint, arXiv: 2407.17797, 2024

［80］ Jia Xiaojun, Gao Sensen, Guo Qing, et al. Semantic-aligned adversarial evolution triangle for high-transferability vision-language attack［J］. arXiv preprint, arXiv: 2411.02669, 2024

［81］ Zhang Pengfei, Bai Guangdong, Huang Zi. MAA: Meticulous adversarial attack against vision-language pre-trained models［J］. arXiv preprint, arXiv: 2502.08079, 2025

［82］ Wang Haodi, Dong Kai, Zhu Zhilei, et al. Transferable multimodal attack on vision-language pre-training models［C］// Proc of the 2024 IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2024: 1722–1740

［83］ Fu Jiyuan, Chen Zhaoyu, Jiang Kaixun, et al. Improving adversarial transferability of vision-language pre-training models through collaborative multimodal interaction［J］. arXiv preprint, arXiv: 2403.10883, 2024

［84］ Ye Jingwen, Yu Ruonan, Liu Songhua, et al. Mutual-modality adversarial attack with semantic perturbation［C］// Proc of the 38th AAAI Conf on Artificial Intelligence (AAAI 2024). Menlo Park, CA: AAAI, 2024: 6657–6665

［85］ Zhou Ziqi, Hu Shengshan, Li Minghui, et al. AdvCLIP: Downstream-agnostic adversarial examples in multimodal contrastive learning［J］. arXiv preprint, arXiv: 2308.07026, 2023

[86] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C] // Proc of the 27th Int Conf on Neural Information Processing Systems (NIPS 2014). Cambridge, MA: MIT, 2014: 2672–2680

[87] Kim H, Kim M, Kim C. Doubly-universal adversarial perturbations: deceiving vision-language models across both images and text with a single perturbation [J]. arXiv preprint, arXiv: 2412.08108, 2024

[88] Zhang Pengfei, Huang Zi, Bai Guangdong. Universal adversarial perturbations for vision-language pre-trained models [C] // Proc of the 47th Int ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR 2024). New York: ACM, 2024: 862–871

[89] Huang Hanxun, Erfani S M, Li Yige, et al. TUAP: Targeted universal adversarial perturbations for CLIP [J]. arXiv preprint, arXiv: 2505.05528, 2025

[90] Awadalla A, Gao I, Gardner J, et al. OpenFlamingo: An open-source framework for training large autoregressive vision-language models [J]. arXiv preprint, arXiv: 2308.01390, 2023

[91] Liu Haotian, Li Chunyuan, Wang Yuheng, et al. LLaVA: Large language and vision assistant via visual instruction tuning [EB/OL]. arXiv preprint, arXiv: 2304.08485, 2023

[92] Fang Hao, Kong Jiawei, Yu Wenbo, et al. One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models [J]. arXiv preprint, arXiv: 2406.05491, 2024

[93] Yang Fan, Huang Yihao, Wang Kailong, et al. Efficient and effective universal adversarial attack against vision-language pre-training models [J]. arXiv preprint, arXiv: 2410.11639, 2024

[94] Zhang Jiaming, Ye Junhong, Ma Xingjun, et al. AnyAttack: Towards large-scale self-supervised adversarial attacks on vision-language models [J]. arXiv preprint, arXiv: 2410.05346, 2024

[95] Zhang Shibo, Cheng Yushi, Zhu Wenjun, et al. CAPatch: Physical adversarial patch against image captioning systems [C] // Proc of the 32nd USENIX Security Symp (USENIX Security 2023). Berkeley, CA: USENIX Association, 2023: 679–696

[96] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models [C] // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Red Hook, NY: Curran Associates, 2020: 6840–6851

[97] Kong Dehong, Liang Siyuan, Zhu Xiaopeng, et al. Patch is enough: Naturalistic adversarial patch against vision-language pre-training models[J]. Visual Intelligence, 2024, 2: 33

[98] Wang Zhixiang, Wang Xiaosen, Wang Bo, et al. DiffPatch: Generating customizable adversarial patches using diffusion models [J]. arXiv preprint, arXiv: 2412.01440, 2024

[99] Li Chao, Gao Shangqian, Deng Cheng, et al. Cross-modal learning with adversarial samples [C] // Advances in Neural Information Processing Systems 32 (NeurIPS 2019). Red Hook, NY: Curran Associates, 2019: 10791–10801

[100] Li Chao, Gao Shangqian, Deng Cheng, et al. Adversarial attack on deep cross-modal Hamming retrieval [C] // Proc of 2021 IEEE/CVF Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE: 2198–2207

[101] Chen Hongge, Zhang Huan, Chen Pinyu, et al. Attacking visual language grounding with adversarial examples: A case study on neural image captioning[C] // Proc of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018). Stroudsburg, PA : ACL, 2018: 2587–2597

[102] Xu Yan, Wu Baoyuan, Shen Fumin, et al. Exact adversarial attack to image captioning via structured output learning with latent variables[C] //Proc of 2019 IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 4135–4144

[103] Zhang Shaofeng, Wang Zheng, Xu Xing, et al. Fooled by imagination: Adversarial attack to image captioning via perturbation in complex domain[C] // Proc of 2020 IEEE Int Conf on Multimedia and Expo (ICME). Piscataway, NJ: IEEE, 2020: 1–6

[104] Xu Xiaojun, Chen Xinyun, Liu Chang, et al. Fooling vision and language models despite localization and attention mechanism[C] // Proc of 2018 IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2018: 4951–4961

[105] Sharma Vasu, Kalra Ankita, Vaibhav, et al. Attend and attack: Attention guided adversarial attacks on visual question answering models[J]. arXiv preprint, arXiv: 1807.09956, 2018

[106] Zhang Boyuan, Li Jiaxu, Shi Yucheng, et al. VADS: Visuo-adaptive DualStrike attack on visual question answer[J]. Computer Vision and Image Understanding, 2024, 249: 104137

[107] Yin Ziyi, Ye Muchao, Zhang Tianrong, et al. VQAttack: Transferable adversarial attacks on visual question answering via pre-trained models [C] // Proc of the AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2024: 6755–6763

[108] Li Siyuan, Xu Xxing, Zhou Zailei, et al. ARRA: Absolute-relative ranking attack against image retrieval [C] // Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022: 610–618

[109] Xu Weilin, Evans D, Qi Yanjun. Feature squeezing: Detecting adversarial examples in deep neural networks [C] // Proc of the 25th Network and Distributed System Security Symp (NDSS). Reston, VA: Internet Society, 2018: 1–15

[110] Guo Chuan, Rana M, Cissé M, et al. Countering adversarial images using input transformations [C] // Proc of the 6th Int Conf on Learning Representations (ICLR). Vancouver: ICLR, 2018: 1–12

[111] Naseer M, Khan S, Hayat M, et al. A self-supervised approach for adversarial robustness [C] // Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 259–268

[112] Fang A, Ilharco G, Wortsman M, et al. Data determines distributional robustness in contrastive language-image pre-training (CLIP) [C] // Proc of the 39th Int Conf on Machine Learning (ICML). New York: PMLR, 2022: 6216–6234

[113] Nguyen T, Ilharco G, Wortsman M, et al. Quality not quantity: On the interaction between dataset design and robustness of CLIP [C] // Proc of the 36th Conf on Neural Information Processing Systems (NeurIPS). Red Hook, NY: Curran Associates, 2022: 21455−21469

[114] Tu Weijie, Deng Weijian, Gedeon T. A closer look at the robustness of contrastive language-image pre-training (CLIP) [C] // Proc of the 37th Conf on Neural Information Processing Systems (NeurIPS). Red Hook, NY: Curran Associates, 2023: 13678−13691

[115] Mao Chengzhi, Geng S, Yang Junfeng, et al. Understanding zero-

shot adversarial robustness for large-scale models［C/OL］// Proc of the 11th Int Conf on Learning Representations (ICLR). Appleton, WI: ICLR, 2023［2025-04-15］. https://openreview.net/forum?id=P4bXCawRi5J

［116］ Deng Jia, Dong Wei, Socher R, et al. ImageNet: A large-scale hierarchical image database［C］// Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 248–255

［117］ Schlarmann C, Singh N D, Croce F, et al. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models［C］// Proc of the 41st Int Conf on Machine Learning (ICML). New York: PMLR, 2024: 43685−43704

［118］ Hossain M Z, Imteaj A. Sim-CLIP: Unsupervised siamese adversarial fine-tuning for robust and semantically-rich vision-language models［J］. arXiv preprint, arXiv: 2407.14971, 2024

［119］ Li Xiao, Zhang Wei, Liu Yining, et al. Language-driven anchors for zero-shot adversarial robustness［C］// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 24686–24695

［120］ Zhou Wanqi, Bai Shuanghao, Mandic D P, et al. Revisiting the adversarial robustness of vision language models: A multimodal perspective［J］. arXiv preprint, arXiv: 2404.19287, 2024

［121］ Waseda F, Tejero-de-Pablos A, Echizen I. Multimodal adversarial defense for vision-language models by leveraging one-to-many relationships［J］. arXiv preprint, arXiv: 2405.18770, 2024

［122］ Yu Lu, Zhang Haiyang, Xu Changsheng. Text-guided attention is all you need for zero-shot robustness in vision-language models［C］// Proc of the 37th Conf on Neural Information Processing Systems (NeurIPS). Red Hook, NY: Curran Associates, 2024: 96424−96448

［123］ Cui Xuanming, Aparcedo A, Jang Y K, et al. On the robustness of large multimodal models against image adversarial attacks［C］// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 24625−24634

［124］ Li Lin, Guan Haoyan, Qiu Jianing, et al. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models［C］// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 24408–24419

［125］ Zhang Jiaming, Ma Xingjun, Wang Xin, et al. Adversarial prompt tuning for vision-language models［C］// Proc of the 18th European Conf on Computer Vision (ECCV). Berlin: Springer, 2024: 56–72

［126］ Zhang Hongyang, Yu Yaodong, Jiao Jiantao, et al. Theoretically principled trade-off between robustness and accuracy［C］// Proc of the 36th Int Conf on Machine Learning. New York: PMLR, 2019: 7472–7482

［127］ Li Chao, Liao Yonghao, Ding Caichang, et al. MDAPT: Multi-modal depth adversarial prompt tuning to enhance the adversarial robustness of visual language models［J］. Sensors, 2025, 25(1): 258

［128］ Bhagwatkar R, Nayak S, Bashivan P, et al. Improving adversarial robustness in vision-language models with architecture and prompt design［C］// Findings of Association for Computational Linguistics: EMNLP 2024. Stroudsburg, PA: ACL, 2024: 17003–17020

［129］ Zhai Xiaohua, Mustafa B, Kolesnikov A, et al. Sigmoid loss for language image pre-training［C］// Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 11976–11986

［130］ Han Zhongyi, Luo Gongxu, Sun Hao, et al. AlignCLIP: Attention alignment for robust open-vocabulary classification via vision-language models［J］. Machine Learning, 2025, 114(3): 1−19

［131］ Zhang Mingkun, Bi Keping, Chen Wei, et al. CLIPure: Purification in latent space via CLIP for adversarially robust zero-shot classification［C/OL］// Proc of the 13th Int Conf on Learning Representations. Red Hook, NY: Curran Associates, 2025［2025-04-15］. https://openreview.net/forum?id=TQ2ZOy6miT

［132］ Zhang Yudong, Xie Ruobing, Chen Jiansheng, et al. PIP: Detecting adversarial examples in large vision-language models via attention patterns of irrelevant probe questions［C］// Proc of the 32nd ACM Int Conf on Multimedia. New York: ACM, 2024: 11175−11183

［133］ Fares S, Ziu K, Aremu T, et al. MirrorCheck: Cross-modal verification via text-to-image generation for detecting adversarial attacks in vision-language models［J］. arXiv preprint, arXiv: 2406.09250, 2024

［134］ Huang Youcheng, Zhu Fengbin, Tang Jingkun, et al. NEARSIDE: N-time embedding-based adversarial image detection for large vision-language models［J］. arXiv preprint, arXiv: 2410.22888, 2024

［135］ Xing Songlong, Zhao Zhengyu, Sebe N. CLIP is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of CLIP［J］. arXiv preprint, arXiv: 2503.03613, 2025

［136］ Zhou Bolei, Khosla A, Lapedriza A, et al, Torralba Antonio. Learning deep features for discriminative localization［C］//Proc of the 2016 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2921−2929

［137］ Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization［C］//Proc of the 2017 IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 618−626



Zeng Cheng, born in 2002. PhD candidate. His main research interests include artificial intelligence interpretability and artificial intelligence security.

曾　诚, 2002 年生. 博士研究生. 主要研究方向为人工智能可解释性、人工智能安全.



Ge Yunjie, born in 1997. PhD. Her main research interests include artificial intelligence security and model fairness.

葛云洁, 1997 年生. 博士. 主要研究方向为人工智能安全、模型公平性.



Zhao Lingchen, born in 1994. PhD, associate professor. His main research interests include artificial intelligence security and privacy computing.

赵令辰, 1994 年生. 博士, 副教授. 主要研究方向为人工智能安全、隐私计算.



Wang Qian, born in 1980. PhD, professor. His main research interests include artificial intelligence security, data storage, security of outsourced query and computation, privacy protection, wireless system security, and applied cryptography.

王　骞, 1980 年生. 博士, 教授. 主要研究方向为人工智能安全、数据存储、外包查询及计算安全、隐私保护、无线系统安全、应用密码学.