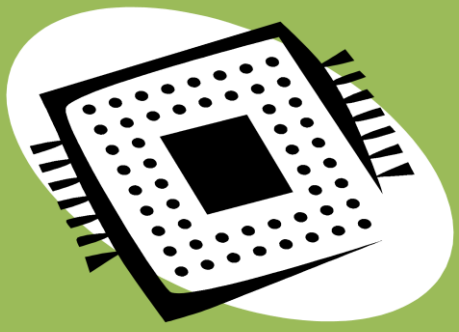
华中科技大学

2023

基于 VLP 的对抗鲁棒性研究

专　　业：	计算机科学与技术
班　　级：	CS2106
学　　号：	U202115514
姓　　名：	杨明欣
电　　话：	13390396012
邮　　件：	ymx@hust.edu.cn
指导老师：	刘康
完成日期：	2024-01-16



计算机科学与技术学院

摘要

本综述聚焦于基于视觉语言预训练（VLP）的对抗鲁棒性，旨在深入了解 VLP 模型在面对对抗攻击时的表现、对抗攻击方法的应用，以及探讨有效的防御方法。通过对 VLP 模型的鲁棒性进行调研，本研究揭示了这些模型在多模态场景下面临的挑战，特别是在应对不同对抗攻击时的表现。本研究涵盖了对抗攻击方法的深入研究，旨在了解如何通过特定逻辑设计的输入样本来欺骗 VLP 模型，以及这些攻击方法的影响。同时，本研究还关注对抗防御方法，试图探索有效的策略，提高 VLP 模型的对抗鲁棒性。

关键词：视觉语言模型、对抗鲁棒性、对抗攻击、对抗防御

Abstract

This review focuses on the adversarial robustness of Vision-Language Pre-trained (VLP) models, aiming to gain a deep understanding of the performance of VLP models when facing adversarial attacks, the application of adversarial attack methods, and the exploration of effective defense methods. Through an investigation into the robustness of VLP models, this study reveals the challenges these models face in multimodal scenarios, especially in their performance against various adversarial attacks. The research encompasses an in-depth examination of adversarial attack methods, seeking to understand how to deceive VLP models through input samples designed with specific logic and the impact of these attack methods. Simultaneously, this study also pays attention to adversarial defense methods, attempting to explore effective strategies to enhance the adversarial robustness of VLP models.

Keywords: Vision-Language Model, Adversarial Robustness, Adversarial Attack, Adversarial Defense

目录

1 基本概念

1.1 概述

深度学习的迅猛发展在人工智能领域掀起了一场革命，尤其是生成式人工智能（如GPT-3）的崛起[1]，为自然语言处理任务提供了前所未有的性能。这些模型通过在大规模文本语料库上进行预训练，成功地捕捉到语言的复杂结构和语义信息，取得了很好的效果。

与此同时，计算机视觉领域也在以前所未有的速度发展，大规模图像处理上取得了显著的进展。其中之一是 "Segment Anything" 的方法，它迈出了分割任务的新步伐。传统的图像分割方法通常面临着特定物体类别的局限性，而 "Segment Anything" 方法旨在突破这一限制，使模型能够有效地分割图像中的任何物体，提高了分割任务的通用性。[2]

在两个领域相辅相成推动下，多模态研究变得日益重要，多模态方法逐渐成为研究焦点。在多模态研究中，视觉语言预训练（Vision-Language Pre-training，VLP）模型崭露头角，这些模型通过联合学习视觉和语言表示，不仅在自然语言理解任务中表现出色，还在处理图像相关信息方面展现了巨大潜力[3]。CLIP（Contrastive Language-Image Pretraining）等代表性 VLP 模型成功地将语言和图像信息紧密结合，推动了多模态应用的发展[4]。

然而，随着这些技术在实际应用中的广泛应用，研究者们开始关注 VLP 模型在面对对抗攻击时的鲁棒性[5]。对抗攻击是指通过特定逻辑故意设计的输入样本，旨在欺骗模型并导致错误的输出[6]，因此如何在多模态场景下提高 VLP 模型的对抗鲁棒性成为一个关键问题。在本部分中，我们将深入研究 VLP 模型的兴起，并着重讨论它们在面对不同对抗攻击时的鲁棒性挑战。

1.2 VLP 模型的定义

在 VLP 中，模型通过在大规模跨模态数据集上进行预训练，学习了语义上的对应关系[3]。以图像文本预训练为例，模型被期望能够将文本中描述的物体或场景与图像

中相应的视觉信息关联起来。为实现这一目标，VLP 的对象（例如，图像、文本、视频）以及模型架构必须经过精心设计。这样的设计使得模型能够深入理解不同模态之间的语义内容，并有效地捕捉它们之间的关系。这种预训练策略为 VLP 模型提供了跨模态任务上优越性能的基础。

总的来说，VLP 通过巧妙设计的对象和模型架构，在大规模数据上进行预训练，从而使模型能够学习和理解不同模态之间的语义对应关系。

1.3 VLP 模型的分类

目前主流的 VLP 模型主要分为融合型（fused VLP models）和对齐型（aligned VLP models）两类，这两种模型在处理多模态信息时采用了不同的策略。融合型 VLP 模型注重将不同模态信息融合到一个共享的空间，以便于跨模态理解；而对齐型 VLP 模型则更注重学习不同模态之间的语义对应关系，使它们在语义上保持一致性，这两种模型各自有其独特的优势。

1.3.1 融合型（fused VLP models）

融合型 VLP 模型通过将不同模态的信息融合到一个共同的语义空间中，以实现跨模态的语义理解。这种模型通常使用共享的嵌入空间，使得文本、图像或其他模态的表示能够在同一空间中相互影响。下面我们以 ALBEF[7]为例介绍这一类模型的特点。

早期的 VLP 主要是靠一个预训练的图像检测器来提取图像特征，并用多模态的 encoder 编码器将图像特征和文本特征融合，然后使用“完形填空（MLM）、图文匹配（Match）”等下游任务来训练多模态编码层[9]。但是这样就会存在一个问题，图像特征和文本特征不在同一个特征空间，这也就是 fused VLP models 要解决的核心问题。

为此，ALBEF 设计了一个图像 12 层 transformer、文本 6 层 transformer、多模态 6 层 transformer 的多 transformer 架构，架构图如图 1-1 所示。

其核心就是 Image-Text Contrastive Learning（ITC）模块，通过 ITC 任务使得图像编码器的向量空间与文本编码器的向量空间一致，具体方法就是通过对比学习[10]。在 ITC 任务中，获取图像编码器的[CLS]编码作为图像的特征，获取文本编码器的[CLS]特征作为整个文本的特征信息，计算两个特征之间的相似度，其相似度计入最终目标的损失函数中，通过训练使配对的图像文本特征相似度越来越高，不配对的图像文本特

征相似度越来越低。

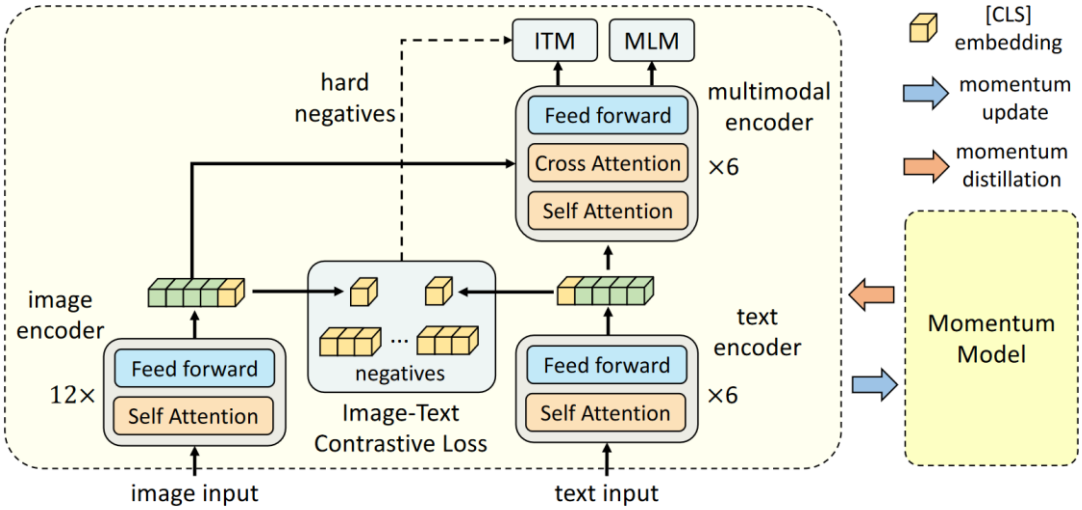


图 1-1 ALBEF 模型架构图

近年来，这一类模型得到了很大的发展和进步。如西湖大学提出的 CVT-SLR 充分利用视觉和语言模态的预训练知识，解决 SLR（手语识别）问题，取得了非常好的效果，获得了 CVPR 2023 Highlight 评价。

1.3.2 对齐型（aligned VLP models）

对齐型 VLP 模型侧重于学习不同模态之间的对应关系，而不是将它们融合到一个共同的空间中。这种模型通过引入对齐机制，使得不同模态的表示在语义上保持一致，而不一定要映射到相同的嵌入空间。在训练过程中，对齐型 VLP 模型注重通过对齐损失函数等手段，促使不同模态的表示在语义上保持一致性。这种方法的优势在于能够更加灵活地处理不同模态的异构性，适用于多样化的多模态任务。下面我们以 CLIP[4] 为例介绍这一类模型的特点。

2021 年，OpenAI 这家致力于推动强人工智能领域发展的公司开创了一条引人注目的道路，通过提出 Contrastive Language-Image Pre-training（CLIP）的方法，彻底颠覆了传统对计算机视觉领域的认知。Alec Radford 等人在该研究中推崇了一种不同寻常的思路，成功地打破了文本和图像之间的局限性。CLIP 通过大规模的文本-图像配对预训练，具有直接迁移到 Imagenet 上的能力，而且令人惊叹的是，它完全不需要对图像标签进行微调，便能够实现零样本分类。

CLIP 的模型结构相对于上面介绍的 ALBEF 更为简单，包括两个部分，即文本编

码器（Text Encoder）和图像编码器（Image Encoder)。Text Encoder 选择的是 Text Transformer 模型；Image Encoder 选择了两种模型，一是基于 CNN 的 ResNet，二是基于 Transformer 的 ViT，架构图如图 1-2 所示。

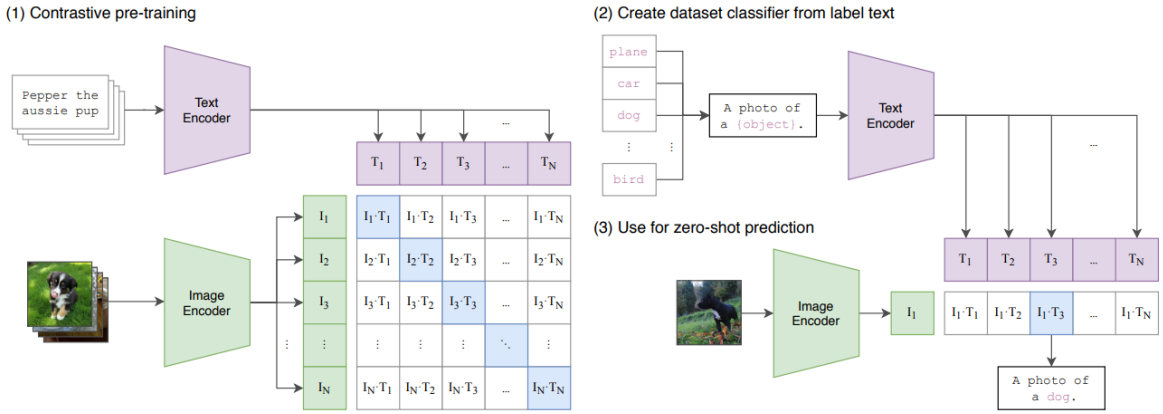


图 1-2 CLIP 模型架构及工作流程

CLIP 能够实现语义一致的核心要点也是基于对比学习，通过大规模的文本-图像配对预训练，使视觉和语义信息在共享的语义空间中进行深度整合。通过对比损失，模型被迫学习将相关的文本-图像对编码为相近（矩阵中对角的部分，数量为 N，即 embedding 的维度）的表示，而将不相关的对编码为远离（矩阵中除去对角的部分，数量为 N^2 - N，）的表示。

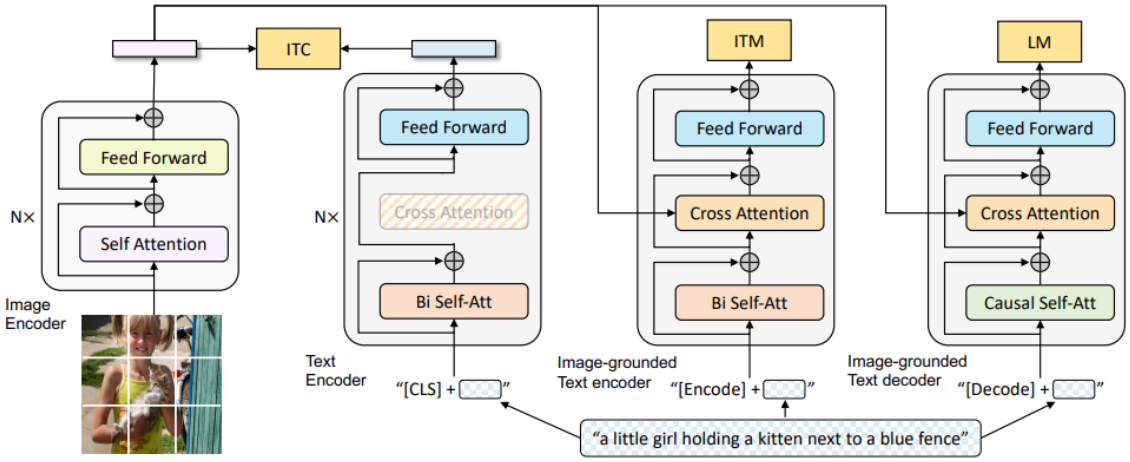


图 1-3 BLIP 模型架构

进一步随着生成式人工智能的发展，多模态模型不仅满足于单一任务，需要同时具备理解能力和生成能力。BLIP[12]则在 CLIP 基础上进行改进和发展，将自然语言理解和自然语言生成任务进行了融合形成了多模态通用模型，模型架构如图 1-3 所示，

通过图像-文本对比损失 ITC、图像-文本匹配损失 ITM、语言建模损失 LM 三个损失函数联合进行预训练，具备通用能力。BLIP-2[13]则进一步发展，形成通用且计算效率高的视觉语言预训练方法，它利用了冻结的预训练图像编码器和 LLM，使得多模态模型的发展可以向更庞大的方向发展，MiniGPT4[14]就是一个重要的例子。

1.4 对抗攻击

对抗攻击是指通过对机器学习模型输入进行微小的、故意设计的修改，以欺骗模型、使其产生误判的技术。这些修改通常是不可察觉的，但足以引起模型错误的分类。有很多分类方式，包括根据对于被攻击模型的了解程度有黑盒攻击和白盒攻击的分类、根据攻击的效果有目标攻击和无目标攻击的分类、根据攻击的模态有文本对抗攻击和视觉对抗攻击的分类。下面我们列出对抗攻击的核心范式。

$$\max_{\delta \in S} L(\theta, x+\delta, y) \tag{1}$$

在上面的式子中，x 代表特征向量，y 代表标签，L 函数为目标损失函数的计算，θ 代表模型参数，δ 代表微小扰动，上面的式子很好地解释了对抗攻击的目的就是在可控的扰动范围 S 内，即不改变人类或客观评判的情况下，使得模型的损失函数变大，即是模型判断错误。

1.4.1 视觉领域的对抗攻击

对抗现象最早是从视觉领域产生的，图像因为其连续性，因此很容易进行扰动，产生对抗样本[6]。

生成对抗样本进行对抗攻击的方法很多，其中比较出色的包括 FGSM[16]和 PGD[17]等。FGSM（Fast Gradient Sign Method）是一种简单而高效的对抗攻击算法，通过计算模型损失函数对输入图像的梯度，然后按照梯度的方向对图像进行微小扰动，生成对抗样本。这一方法在白盒攻击场景中表现出色，适用于快速生成对抗样本。相比之下，PGD（Projected Gradient Descent）是一种更为强大的对抗攻击方法，通过进行多次迭代的梯度下降来生成对抗样本，每次迭代对图像进行微小扰动，并在每次迭代之后将扰动投影回一个预定义的范围，以提高攻击的强度和灵活性。这两种算法都挑战深度学习模型的鲁棒性，为对抗攻击领域提供了重要的研究基础。

1.4.2 文本领域的对抗攻击

在文本领域，对抗攻击旨在通过对文本输入进行微小的、精心设计的修改，欺骗自然语言处理（NLP）模型，由于文本是离散的，因此视觉领域的一些攻击方法是不适用的[15]。

文本对抗攻击的主要方法包括修改关键词、改变语法结构、插入噪声等，以生成能够误导模型的对抗性文本。例如 TextFooler[18]，一种基于替换和插入的对抗攻击方法，通过替换文本中的关键词或插入噪声，使模型产生错误的分类。该方法在迁移性和攻击性能上取得了显著的成功。进一步提出的 PWWS[19]方法，在生成对抗样本上取得了更好的效果。

文本领域的对抗攻击主要包括字符级攻击，词级攻击以及句子级攻击。

2 VLP 的对抗鲁棒性

2.1 概述

随着视觉语言模型在多模态智能应用中的广泛应用，对其鲁棒性的关注日益增加。这关切的焦点之一是对抗鲁棒性，即模型在面对精心设计的对抗性攻击时的表现。在视觉语言预训练（VLP）领域，对抗鲁棒性的研究旨在深入理解模型的弱点，探索提升模型对对抗性输入的稳健性的方法。

2.2 对抗鲁棒性评估

目前的工作已经有对于 VLP 模型鲁棒性的详细评估，进行了大量的实验，提出了完善的评测基准。现有研究表明，视觉语言模型的对抗鲁棒性存在显著脆弱性。尽管一些模型采用了对抗训练等技术，但实证研究显示，常规训练的视觉语言模型在面对新型对抗攻击时依然容易受到影响[20]。

此外，攻击的迁移性研究强调了不同模型在面对相似对抗性样本时仍然共享的弱点，进一步凸显了对抗鲁棒性的局限性。模型解释性与对抗攻击之间的关系也表明，缺乏足够解释性可能使模型更容易受到攻击。

2.3 预训练方法对鲁棒性的影响

同时，目前的研究对于 VLP 模型不同的适应方式对应模型的对抗鲁棒性进行了评估。在一个包含 96 种不同的视觉损坏，包括脉冲噪声、雪等，以及 87 种文本损坏，包括文本添加、反向翻译等大规模鲁棒性基准数据集，通过大量实验，评估了 11 种适应下游任务的方法在视觉语言模型上的鲁棒性[21]。其中部分的实验结果如图 2-1 所示。

Adaptation method *Image Corruptions*	Updated Params	VQAv2 Acc (%)	VQAv2 RR (%)	GQA Acc (%)	GQA RR (%)	NLVR^2 Acc (%)	NLVR^2 RR (%)	MSCOCO Caption CIDEr	MSCOCO Caption RR (%)
Full Fine-tuning	100%	66.75	84.86±5.17	55.04	89.20±0.04	73.01	90.34±0.04	115.03	68.40±0.14
Multiple Adapters	12.22%	65.30	85.33±4.90	53.39	86.16±0.04	69.41	**92.02**±0.04	114.47	68.72±0.14
Half-shared Adapters	8.36%	65.20	85.18±5.01	52.96	89.37±0.04	70.03	91.72±0.04	114.50	68.45±0.14
Single Adapter	4.18%	65.35	**85.76**±5.32	54.14	82.49±0.04	73.89	90.04±0.05	115.04	68.68±0.14
Hyperformer	5.79%	65.38	85.38±4.84	52.52	90.05±0.04	72.21	90.13±0.05	114.89	68.74±0.14
Multiple Compacters	7.05%	64.91	85.65±4.81	52.75	88.89±0.04	69.45	91.33±0.04	115.16	68.67±0.13
Single Compacter	2.70%	64.47	85.47±4.96	52.90	82.62±0.04	69.94	92.04±0.04	113.06	**69.92**±0.13
Multiple LoRA	17.72%	65.44	84.78±4.86	52.05	**91.15**±0.04	51.32	–	115.41	68.47±0.14
Single LoRA	5.93%	65.34	84.78±4.81	53.19	82.58±0.04	73.58	90.05±0.04	114.54	69.26±0.13
Multiple Prompts	4.53%	46.81	–	34.01	–	49.87	–	108.62	67.70±0.14
Single Prompt	2.00%	44.00	–	37.54	–	51.95	–	103.70	68.56±0.13

Adaptation method *Text Corruptions*	Updated Params	VQAv2 Acc (%)	VQAv2 RR (%)	GQA Acc (%)	GQA RR (%)	NLVR^2 Acc (%)	NLVR^2 RR (%)
Full Fine-tuning	100%	66.75	73.65±22.38	55.04	66.92±24.14	73.01	87.06±11.00
Multiple Adapters	12.22%	65.30	76.62±20.66	53.39	66.93±22.43	69.41	**90.14**±10.19
Half-shared Adapters	8.36%	65.20	76.78±20.79	52.96	68.20±24.78	70.03	89.16±10.12
Single Adapter	4.18%	65.35	**77.64**±21.09	54.14	67.47±20.03	73.89	88.49±10.87
Hyperformer	5.79%	65.38	75.06±21.29	52.52	**70.30**±23.13	72.21	87.27±11.27
Multiple Compacters	7.05%	64.91	77.10±20.85	52.75	67.39±23.29	69.45	90.00±9.76
Single Compacter	2.70%	64.47	77.17±20.40	52.90	67.90±20.33	69.94	90.10±9.81
Multiple LoRA	17.72%	65.44	74.04±21.97	52.05	68.77±22.76	51.32	–
Single LoRA	5.93%	65.34	74.50±21.42	53.19	63.94±20.99	73.58	87.64±11.04
Multiple Prompts	4.53%	46.81	–	34.01	–	49.87	–
Single Prompt	2.00%	44.00	–	37.54	–	51.95	–

图 2-1 不同适应方法的鲁棒性（RR 越大，代表越鲁棒）

在大量实验的分析中发现了一些引人注目的结果：首先，相比于视觉攻击，适应方法对文本攻击表现出更高的敏感性。其次，全面微调并不始终产生最佳的相对鲁棒性，相反，使用适配器在性能可比的情况下能够取得更好的鲁棒性。令人惊讶的是，实验结果显示大量的适应数据和模型参数并不能保证提高鲁棒性，事实上，增加适应数据的量甚至可能导致鲁棒性降低。

最后，我们发现没有一种适应方法能够在所有任务和攻击情景下都超越其他方法，强调了在不同应用场景下选择适当的适应方法的重要性。这些发现为深入理解多模态适应方法的鲁棒性提供了有益的见解。

2.4 对抗鲁棒的现实场景

视觉语言模型的对抗鲁棒性在现实场景中具有重要意义。在实际应用中，模型常常面临来自于恶意攻击或者自然环境变化的挑战，这些挑战可能导致模型性能下降或者产生误导性的输出。对抗鲁棒性成为保障模型可靠性和稳定性的必要条件。

在视觉语言交互中，模型需要在处理多模态输入时保持鲁棒性。例如，在智能辅助系统中，用户可能通过语言和图像共同传达信息，而对抗攻击可能导致模型误解用户

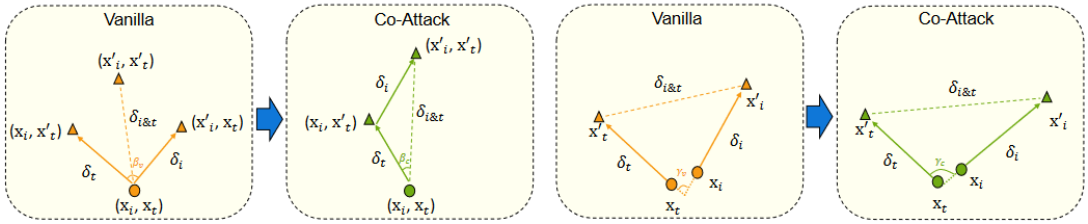
意图或提供不准确的反馈，从而影响系统的实用性和用户体验。此外，在安全监控和自动驾驶等领域，模型对于视觉和语言信息的正确解释至关重要。对抗攻击可能通过操纵输入，使模型无法正确识别对象、理解场景或执行指令，进而威胁到系统的安全性和可靠性。

因此，提高视觉语言模型的对抗鲁棒性，使其在面对各种挑战时能够保持稳定的性能，对于确保人工智能系统在实际应用中的可靠性至关重要。这需要深入研究并制定相应的对抗防御策略，以适应多变的实际场景需求。

3 VLP 对抗攻击方法和技术

3.1 概述

对于视觉语言模型的攻击方法在近些年得到了很大的发展，早期的对抗攻击手段主要针对于视觉语言模型中的单一视觉模块或者单一文本模块，具有一定的效果。但是更深入的研究发现，只对单一模块进行攻击或者不考虑模态之间的联系同时对于两个模态进行攻击时，可能不会产生作用甚至产生反作用，具体示意如图 3-1 协同对抗攻击示意图所示。因此，对于视觉语言模型的攻击逐步转换为对于视觉语言之间联系的攻击，经实证取得了更好的效果。



(a) multimodal embedding space (for fused VLP model) (b) unimodal embedding space (for aligned VLP model)

图 3-1 协同对抗攻击示意图

3.2 单一攻击

3.2.1 对视觉模态的攻击

Xu 等人的研究[22]涉及对视觉问答模型进行攻击，攻击手段涉及对图像模态进行微扰。这种攻击方法可能旨在通过对输入的图像进行变化，来观察模型在处理扰动后的图像时的鲁棒性以及其对视觉问答任务的影响。这类研究有助于深入理解视觉问答模型在面对对抗性攻击时的脆弱性，并推动开发更鲁棒的模型或者对抗性训练策略。

3.2.2 对语言模态的攻击

[23]针对视觉问答模型在面对问题文本变异时的不足，提出了一种创新的循环一致性训练策略。通过引入问题一致性和答案一致性的机制，该策略利用问题生成模块和门机制来训练视觉问答模型，以提高其对于问题文本变异的鲁棒性。具体而言，问题生

成模块利用 VQA 模型的预测答案和图像特征生成新问题，并通过循环一致性损失保持生成问题与原问题的一致性。引入的门机制用于过滤生成的问题，确保其与原问题的一致性。通过实验证明，该循环一致性训练策略在提升 VQA 模型鲁棒性方面取得了显著效果，为视觉语言模型的对抗性训练提供了新的思路和方法。这类研究进一步推动了对于视觉语言模型对抗攻击方法的探究。

3.3 协同攻击

3.3.1 Co-Attack

这篇文章提出的攻击方法是对于视觉语言模型进行协同攻击的第一次尝试，在 VLP 模型上提出了一种新的多模态攻击方法，称为协同多模态对抗攻击(Co-Attack)[8]，它共同对图像模态和文本模态进行攻击。

Co-Attack 解决了连续图像模态和离散文本模态之间的输入表示差距的挑战。采用逐步方案，首先扰动离散文本输入，因为在离散空间中优化设计的目标很困难。文本模态扰动作为标准，然后进行图像模态扰动，协同干扰输入文本和输入图像，以引导多模态嵌入远离原始嵌入。其实验结果如图 3-2 所示。

Model	Attack	Flickr30K (1K test set)						MSCOCO (5K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ALBEF	Fooling VQA	12.80	4.70	3.10	12.72	6.62	4.40	39.68	31.50	24.82	34.23	34.73	29.70
	SSAP	66.10	55.30	49.60	63.54	64.82	60.62	54.30	58.58	55.32	46.77	60.26	60.94
	SSAP-MIM	61.20	51.50	45.70	60.52	60.94	57.42	48.72	51.64	48.56	42.95	53.95	54.29
	SSAP-SI	70.20	61.50	54.70	66.70	67.90	64.32	58.08	62.64	59.16	49.05	62.62	63.57
	Vanilla	65.70	55.40	48.10	62.92	62.80	58.28	60.54	65.78	62.72	51.08	65.86	67.09
	Co-Attack	70.60	60.50	53.50	67.22	67.10	62.76	58.84	64.18	61.40	51.13	66.17	67.73
	Co-Attack-SI	**72.20**	**63.50**	**58.20**	**69.72**	**71.00**	**67.12**	**65.90**	**71.74**	**67.92**	**54.39**	**70.16**	**71.59**
$CLIP_{ViT}$	Fooling VQA	9.90	5.60	3.40	8.26	5.54	4.20	14.52	13.68	12.10	6.58	8.55	8.02
	SSAP	56.20	49.90	42.20	43.68	47.00	42.58	41.44	52.80	53.34	26.70	41.51	45.43
	SSAP-MIM	48.10	42.20	35.30	38.40	39.42	35.52	38.40	48.10	47.70	24.37	37.47	40.74
	SSAP-SI	58.90	53.90	46.70	46.72	53.54	51.20	43.44	57.06	58.50	27.22	43.38	48.33
	Vanilla	64.00	60.90	52.90	51.08	60.34	58.54	46.34	62.10	64.34	29.75	48.78	54.88
	Co-Attack	73.80	**79.50**	74.90	**58.14**	**75.52**	**77.38**	50.98	**72.34**	**78.12**	**32.31**	**55.66**	**64.67**
	Co-Attack-SI	**74.40**	79.40	**76.00**	57.82	75.20	77.32	**51.06**	72.26	77.66	32.21	55.24	64.23

图 3-2 协同攻击效果，值为攻击成功率

可以看到其攻击效果显著高于上面提到的仅针对单一模态的攻击方法。

3.3.2 SGA

从对抗迁移性的角度出发，首次探索 VLP 模型在黑盒场景下的对抗鲁棒性。作者首先评估了现有方法在基于 VLP 模型的多模态场景下的对抗迁移性，实验结果表明，现有的单模态攻击和多模态白盒攻击（Co-Attack）方法，都不足以生成具有强迁移性的对抗样本。同样，作者认为这种现象的根本原因是缺少模态间交互和样本多样性不足。

因此提出一种集合级引导攻击（Set-level Guidance Attack, SGA）方法。该方法将单一地图像-文本对扩展为集合级别的图像-文本对，并以跨模态数据为监督信息，从而生成具有强迁移性的对抗样本。使用来自不同模态的配对数据作为监督信号来引导对抗样本的优化方向。在迭代优化对抗图像和对抗文本的过程，该策略逐步拉远图像和文本在特征空间中的距离，从而破坏跨模态交互，达到攻击效果。

同时，现有的对抗攻击方法虽然在白盒场景下能取得很好的攻击效果，但是其生成的对抗样本很难迁移到其他的黑盒模型。尽管结合不同的单模态迁移攻击方法，所生成的对抗样本的迁移性的提升依然有限。而相较于现有的对抗攻击方法，SGA 能够大幅度提升对抗样本在 VLP 模型之间的迁移性，特别是同类型的 VLP 模型之间的迁移性，例如从 ALBEF 到 TCL。

3.4 总结

视觉语言模型（VLP）的对抗攻击技术是一项复杂而关键的研究领域，旨在评估和提高这些模型在面对恶意扰动时的鲁棒性，其可应用于提高模型的安全性、评估模型的鲁棒性、推动 VLP 模型的发展和改进。

4 VLP 对抗防御方法和技术

4.1 概述

在视觉语言模型的研究中，对抗防御的问题日益凸显，其挑战性远远超过对抗攻击。对抗攻击旨在通过有意设计的扰动来欺骗模型，然而，对抗防御则涉及到在面对不断进化的攻击技巧时，保护模型免受对抗性扰动的影响。与对抗攻击相比，对抗防御更为困难，因为它需要建立鲁棒性，以抵御来自多个模态的多样化、复杂的攻击。

4.2 防御方法

视觉语言模型通常需要处理来自图像和文本的多模态输入，这使得对抗防御变得更加具有挑战性。攻击者可能通过改变图像或文本中的细微细节，甚至同时对两个模态进行干扰，来欺骗模型，因此对抗防御必须能够有效地区分真实信息和对抗性扰动。这一任务相对于对抗攻击而言更为困难，因为模型需要具备深刻的理解和对多模态输入的一致性敏感度，以便在攻击面前保持稳健性。

[25]提出了一种提高对抗鲁棒性的方案。作者发现标准的多模态融合模型容易受到单一模态的对抗攻击，因此提出了一种对抗鲁棒的融合策略来解决这个问题。该方法包括两个步骤：首先，使用预训练的模态特征提取器从每个模态中提取特征；然后，使用一个鲁棒的融合策略来将这些特征组合起来进行分类。这个鲁棒的融合策略包括两个关键组件：一个是对抗训练，用于训练模型对单一模态的对抗攻击具有鲁棒性；另一个是鲁棒融合，用于将多个模态的特征组合起来进行分类，同时保持对单一模态的对抗攻击具有鲁棒性。

实验结果表明，该方法在单一模态的对抗攻击下具有很强的鲁棒性，并且在多个基准测试中都取得了比现有方法更好的性能。这些结果表明，该方法可以为开发更鲁棒的多模态神经网络提供有用的指导。

4.3 总结

对抗防御的工作面临着多方面的挑战，包括多模态输入的复杂性、攻击技巧的不断演变以及鲁棒性评估的困难。目前取得的成果表明，虽然某些方法在特定任务和攻击场景下表现出良好的性能，但对于通用的对抗防御机制仍存在一定的局限性。

5 总结与结论

5.1 概述

通过上面对与视觉语言模型的对抗鲁棒性调研和分析，对于对抗攻击和对抗防御方法进行整理，我们发现目前在视觉语言模型的对抗性研究方面取得了一定的进展，但是还有很大的发展前景和研究空间。

5.2 对抗攻击

尽管当前的对抗攻击方法在视觉语言模型中取得了一定的成就，但对两个模态之间关系的挖掘仍存在待深入研究的空间。未来的工作可以集中在以下几个方面，以提高视觉语言模型对抗攻击的鲁棒性和性能：

1. 深化模态关系理解：现有的对抗攻击方法主要集中在单模态或简单的多模态攻击上，缺乏对模态之间深层次关系的全面理解。未来的研究可以探索如何更深入地挖掘图像和文本之间的内在关系，以更好地进行攻击。
2. 多模态关联性建模：未来的工作可以致力于开发更高效、更准确的多模态关联性建模方法。通过引入更先进的关系建模技术，模型可以更好地捕捉图像和文本之间的语义关联，从而提高在对抗场景下的鲁棒性。
3. 自适应学习策略：对抗攻击方法需要具备适应性，能够灵活应对不同类型的攻击。未来的研究可以探讨如何设计自适应学习策略，使模型在面对新型攻击时能够迅速调整，并在不损害性能的前提下提高鲁棒性。
4. 伦理和法规问题：随着对抗攻击技术的发展，也需要更多的关注伦理和法规问题。在大语言模型的攻击中，越狱攻击已经被广泛研究。同时，目前也有一些研究开始探究视觉语言模型的价值观对齐[26]，未来的研究可以探讨如何在保障安全性的同时，确保技术的合规性和道德性，以便将其应用于实际应用中。

通过在这些方向上进行深入研究，可以在未来设计更有效的对抗攻击方法，同时帮助模型更好地应对复杂多变的攻击手段，为视觉语言模型的安全性和鲁棒性开创新的可能性。

5.3 对抗防御

在多模态场景下，对抗防御需要更多的关注，以提高模型对来自图像和文本等多源的复杂攻击的鲁棒性。引入强化学习技术可以使模型在动态环境中不断学习和适应，更好地理解对抗攻击的变化。此外，关注模型的解释性和可解释性，使其更具透明性，有助于用户和开发者理解模型的决策过程。标准化和合规性问题也将成为未来研究的重要方向，以确保对抗防御技术在应用中符合伦理和法规要求。

5.4 总结

本综述全面调查了基于视觉语言预训练（VLP）的对抗鲁棒性研究，通过深入分析多模态模型的对抗性攻击性能，揭示了扰动双模态输入相较于单模态输入更为强大的特性。进一步强调了对抗性防御的挑战。总体而言，该综述深刻洞察了 VLP 模型的对抗性特性，为未来提高视觉语言模型对抗鲁棒性的研究提供了有益的启示。

6 参考文献

[1] Brown T B , Mann B , Ryder N ,et al.Language Models are Few-Shot Learners[J]. 2020.DOI:10.48550/arXiv.2005.14165.

[2] Kirillov A, Mintun E, Ravi N, et al. Segment anything[J]. arXiv preprint arXiv:2304.02643, 2023.

[3] Chen F L, Zhang D Z, Han M L, et al. Vlp: A survey on vision-language pre-training[J]. Machine Intelligence Research, 2023, 20(1): 38-56.

[4] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.

[5] Chen S, Gu J, Han Z, et al. Benchmarking Robustness of Adaptation Methods on Pre-trained Vision-Language Models[J]. arXiv preprint arXiv:2306.02080, 2023.

[6] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey[J]. Ieee Access, 2018, 6: 14410-14430.

[7] Li J, Selvaraju R, Gotmare A, et al. Align before fuse: Vision and language representation learning with momentum distillation[J]. Advances in neural information processing systems, 2021, 34: 9694-9705.

[8] Zhang J, Yi Q, Sang J. Towards adversarial attack on vision-language pre-training models[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 5005-5013.

[9] Li, X., X. Yin, C. Li, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In ECCV, pages 121–137. 2020.

[10] Jaiswal A, Babu A R, Zadeh M Z, et al. A survey on contrastive self-supervised learning[J]. Technologies, 2020, 9(1): 2.

[11] Zheng J, Wang Y, Tan C, et al. Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 23141-23150.

[12] Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International Conference on Machine Learning. PMLR, 2022: 12888-12900.

[13] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[J]. arXiv preprint arXiv:2301.12597, 2023.

[14] Zhu D, Chen J, Shen X, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models[J]. arXiv preprint arXiv:2304.10592, 2023.

[15] Zhang W E, Sheng Q Z, Alhazmi A, et al. Adversarial attacks on deep-learning models in natural language processing: A survey[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2020, 11(3): 1-41.

[16] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 9185-9193.

[17] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083, 2017.

[18] Jin D, Jin Z, Zhou J T, et al. Is bert really robust? a strong baseline for natural language attack on text classification and entailment[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(05): 8018-8025.

[19] Ren S, Deng Y, He K, et al. Generating natural language adversarial examples through probability weighted word saliency[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 1085-1097.

[20] Xu X, Chen X, Liu C, et al. Fooling vision and language models despite localization and attention mechanism[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4951-4961.

[21] Chen S, Gu J, Han Z, et al. Benchmarking Robustness of Adaptation Methods on Pre-trained Vision-Language Models[J]. arXiv preprint arXiv:2306.02080, 2023.

[22] Xu X, Chen X, Liu C, et al. Fooling vision and language models despite localization and attention mechanism[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4951-4961.

[23] Shah M, Chen X, Rohrbach M, et al. Cycle-consistency for robust visual question

answering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 6649-6658.

[24] Lu D, Wang Z, Wang T, et al. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 102-111.

[25] Yang K, Lin W Y, Barman M, et al. Defending multimodal fusion models against single-source adversaries[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3340-334

[26] Shayegani E, Dong Y, Abu-Ghazaleh N. Jailbreak in pieces: Compositional Adversarial Attacks on Multi-Modal Language Models[J]. arXiv preprint arXiv:2307.14539, 2023