

**北京航空航天大学计算机学院**

**硕士研究生学位论文**

**开题报告**

**论文题目**： 面向视觉语言模型的对抗攻击方法研究

**专 业**： 计算机技术

**研究方向**： 智能安全

**研 究 生**： 邓彬

**学 号**： ZY2306335

**指导教师**： 郭园方

**北京航空航天大学计算机学院**

2024年11月24日

目 录

1 论文研究背景与意义 1

1.1 论文选题背景 1

1.2 研究现状概述 2

1.2.1视觉模态对抗样本生成方法 2

1.2.2视觉语言模态对抗样本生成方法 4

1.3 研究目标与创新性 6

2 研究内容与技术路线 7

2.1 研究内容 7

2.2 基于集成模型训练的对抗样本生成方法技术路线 8

2.3 基于通用扰动表征对比学习的对抗样本生成方法技术路线 9

3 论文工作安排计划 10

3.1 工作进度安排 10

3.2 关键技术及难点 11

3.2.1如何确定集成模型构成以及各模型权重 11

3.2.2如何选择增强方法以及设计合适的对比学习损失函数 11

主要参考文献 12

图 目

[图1 FGSM攻击深度视觉模型导致高置信的错误预测[6] 2](#_Toc152072753)

[图2 现有方法存在的问题和本文的研究方案 8](#_Toc152072754)

[图3 基于通用扰动表征对比学习的对抗样本生成框架 10](#_Toc152072755)

# 论文研究背景与意义

## 论文选题背景

随着基于大规模数据训练的模型（如BERT[[[1]](#endnote-2)]、DALL-e[[[2]](#endnote-3)]、GPT-3[[[3]](#endnote-4)]）的兴起，人工智能正在经历范式转变。李飞飞等人[[[4]](#endnote-5)]提出基于一类通用模型构建人工智能系统的新兴范式，并将这类通用模型统称为基础模型（Foundation Model）。基础模型定义为在大规模的广泛数据上训练的任何模型（通常使用大规模的自监督训练），可以适应（例如通过微调）广泛的下游任务。从技术角度来看，基础模型并不新鲜，其基本组成部分，如深度神经网络（Deep Neural Network，DNN）和自监督学习，都已经存在多年。然而，基础模型在自然语言处理（Natural Language Processing，NLP）领域取得了突破性的进展，例如，有1750亿个参数的GPT-3已被有效地用于零样本/少样本学习，在不需要大规模任务相关数据或模型参数更新的情况下实现了令人印象深刻的性能。此外，最近有不少研究也对不同感知任务的基础模型进行了探索。例如，CLIP[[[5]](#endnote-6)]等预训练的视觉语言模型在不同的下游视觉任务上以零样本方式表现出了绝佳的性能。这意味着计算机视觉（Computer Vision，CV）领域将会效仿NLP领域，走上从特定任务视觉模型到视觉基础模型的道路。

为了与应用于NLP领域的基础模型有所区分，本文将应用于CV领域的基础模型统称为视觉基础模型。本文认为按照能够适应的下游任务所涉及的模态，可将现有的视觉基础模型划分为能适应于多个视觉模态任务的视觉基础模型和能适应于多个视觉语言模态任务的视觉基础模型。视觉模态任务涵盖了语义分割、目标检测等多种视觉任务，视觉语言模态任务包括视觉问答、图像文本检索、图像描述等涉及视觉、语言两种模态的任务。

在基础模型出现之前，CV领域中以深度模型的应用最为广泛。然而，虽然深度模型在各种各样的计算机视觉任务上能够取得惊人的准确率，有时其表现甚至能够超过人类，但是一些研究表明，基于深度学习的视觉模型在安全方面存在致命缺陷。如图1，通过在输入图像上添加非常微小的扰动所构造的对抗样本（Adversarial Examples），能够影响深度神经网络的预测结果。这些扰动可以被限制在人类视觉系统不可察觉的范围内，但它们可以完全改变深度视觉模型的输出预测。

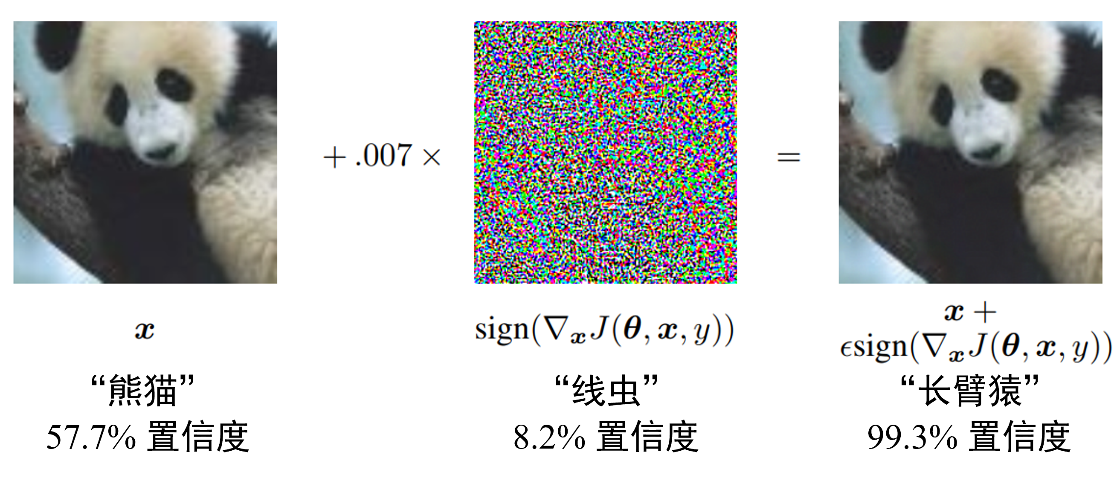


图1 FGSM攻击深度视觉模型导致高置信的错误预测[[[6]](#endnote-7)]

与传统深度视觉模型类似，视觉基础模型同样也存在安全性问题。已有研究表明，在图像任务上，通过在测试样本上添加微小的、几乎不能被肉眼识别的扰动所构造出的对抗样本，能够使视觉基础模型受到欺骗，输出完全不同的分类结果。例如，Fort[[[7]](#endnote-8)]通过实验发现仅通过简单梯度下降优化生成的对抗样本就可以使得CLIP在零样本分类任务上出现明显错误。这一发现使得视觉基础模型的安全性开始逐步受到重视。

最近，OpenAI等研究团队出于安全性考虑，建议在实际应用场景中应当采用闭源的基础模型。这为视觉基础模型的对抗鲁棒性研究带来了新的挑战，亟待通过对抗样本研究视觉基础模型在强对抗环境中的实际性能，以确保其在复杂场景下的安全性和可靠性。

## 研究现状概述

传统对抗样本生成方法主要针对特定任务设计，而在能够适应多种下游任务的视觉基础模型上不适用。因此，一些研究致力于面向视觉基础模型设计一种具有良好迁移性和泛用性的对抗样本生成方法。

对于面向视觉基础模型的对抗样本生成方法，目前已有的工作可以归纳为以下两个方面：

1. 仅扰动图像的视觉模态对抗样本生成；
2. 对图像和文本均进行扰动的视觉语言模态对抗样本生成。

本节将从这两个方面分别介绍，并分析发展趋势。

### 1.2.1视觉模态对抗样本生成方法

前面提到，现有的视觉基础模型中有一类为适应于多个视觉模态任务的模型（如SAM[[[8]](#endnote-9)]等），有一部分工作针对这类模型设计适用的对抗样本生成方法。Zhang等人[[[9]](#endnote-10)]首次针对SAM的对抗鲁棒性进行了研究。他们提出了一个称为Attack-SAM的攻击框架，并设计了一个简单有效的ClipMSE损失，在基于提示的掩膜预测任务中能够使SAM被攻击以生成任何所需的掩膜。这项研究表明视觉基础模型在白盒场景中容易受到对抗样本的影响。然而，Attack-SAM在生成对抗样本时需要知道输入的提示位置，而事先了解提示内容是不现实的，并且会使得对抗样本对固定的提示位置过拟合而降低迁移性。

Zheng等人[[[10]](#endnote-11)]在Attack-SAM基础上采用随机化提示点的方法实现对抗样本对提示不可知。此外，他们受Ilyas等人[[[11]](#endnote-12)]启发，提出通过计算对抗样本与其他干净样本的余弦相似度来测量对抗样本特征的相对强度，并将其作为正则项来提升对抗样本特征的强度，进而实现跨模型的黑盒对抗攻击。但是他们提出的方法只是对采用不同图像编码器的SAM具有迁移性。

与Zhang等人[9]和Zheng等人[10]不同，Han等人[[[12]](#endnote-13)]致力于研究是否可以使用UAP[[[13]](#endnote-14)]方法攻击视觉基础模型SAM。他们认为传统的对抗样本生成方法（如PGD[[[14]](#endnote-15)]、DeepFool[[[15]](#endnote-16)]等）都是以图像为中心，而这些方法只适用于针对特定图像的攻击，需要在图像发生变化时生成新的扰动。因此，他们提出以扰动为中心的攻击框架，基于自监督的对比学习方法生成UAP。具体而言，他们先采用类似PGD的方法基于梯度下降生成对抗扰动，然后再以对抗扰动作为对比视图，将对抗扰动增强后作为正样本，选择随机图像作为负样本，采用InfoNCE[[[16]](#endnote-17)]损失优化学习UAP的独立特征。实验表明，最终生成的UAP不仅能够在图像不可知的情形下成功攻击SAM，而且还能够在提示类型不可知的情形下影响SAM输出的掩膜。然而，以上的工作均基于SAM训练生成对抗样本，且仅适用于基于提示的掩膜预测任务，使得这些方法跨任务迁移性差。

此外，还有一部分工作在能够适应于视觉语言模态任务的视觉基础模型上进行，但是仅对图像进行扰动，因此将其归为视觉模态对抗样本。Zhou等人[[[17]](#endnote-18)]首次基于多模态对比学习方法提出了面向视觉基础模型的攻击框架AdvCLIP，能够生成下游不可知的对抗样本。该框架由一个生成器、一个判别器和一个跨模态编码器组成，其中跨模态编码器包含一个图像编码器和一个文本编码器。生成器通过输入的随机噪声生成对抗扰动，进而与图像组成对抗样本，并输入判别器和跨模态编码器。损失函数由四部分构成：对抗损失，确保对抗样本的特征远离干净图像和文本；拓扑偏差损失，最大化对抗样本与正样本之间的拓扑距离来破坏两者间的拓扑相似性；扰动约束，确保生成的扰动难以被人类视觉系统察觉；GAN损失，使对抗样本与干净图像在判别器上趋于一致，确保对抗样本在视觉上更自然。然而，AdvCLIP要求视觉基础模型的编码器是开源可访问的，即跨模态编码器中的图像编码器和文本编码器来自目标基础模型，同时还要求用户只会微调线性层，这导致AdvCLIP在黑盒场景下不适用。此外，实验结果表明，AdvCLIP仅在图像分类数据集上具有较高的攻击性能，跨任务迁移性差。且由于其仅考虑了对图像的扰动，使得AdvCLIP在图像文本检索任务上的攻击成功率不高。

Zhao等人[[[18]](#endnote-19)]将在视觉基础模型上的黑盒对抗攻击分为基于迁移和基于查询两类，并从这两个方向研究对抗样本生成。在基于迁移的攻击方向，他们提出了两种设计，一种是匹配图像-文本特征（MF-it，Matching image-text features），基于PGD生成能够匹配目标文本的对抗样本；另一种是匹配图像-图像特征（MF-ii，Matching image-image features），他们受Yuksekgonul等人[[[19]](#endnote-20)]的启发，先采用文本生成图像的扩散模型（如Stable Diffusion[[[20]](#endnote-21)]）来生成与目标文本相关的目标图像，然后通过所生成的目标图像优化对抗扰动生成对抗样本。此外，由于优化对抗样本时不需要扩散模型的梯度信息。因此，对于扩散模型只要求黑盒可访问，即可以使用公开提供的API，这大大降低了生成对抗样本的开销。在基于查询的攻击方向，他们提出了匹配文本-文本特征（MF-tt，Matching text-text features）。他们假设在黑盒场景下可对目标基础模型执行重复查询的操作，即通过提供图像和文本输入来获得文本输出，进而可以通过多次查询获得的文本输出来迭代优化对抗扰动。此外，他们还发现所提出的基于迁移的攻击方法和基于查询的攻击方法可以组合使用。实验表明，在生成对抗图像过程中加入文本信息能够一定程度上提升其在视觉语言模态任务上的攻击性能，且组合后的方法所生成的对抗样本相比于组合前提升了在不同视觉基础模型上的迁移性。然而，他们只考虑了在适应于同类任务的视觉基础模型上的迁移性，未考虑在不同类任务上的迁移性，且所生成的对抗图像在所测试的大部分视觉基础模型上的攻击成功率仅有50%~60%，这表明在生成视觉模态对抗样本时加入文本信息并不能从根本上改变其在视觉语言模态任务上迁移性差的问题。

以上提到的视觉模态对抗样本生成方法都是将视觉基础模型作为替代模型进行训练，而相比于传统深度视觉模型，视觉基础模型参数规模更大、网络更深，需要足够的内存空间存储模型参数等信息来支持基础模型的运行，且在迭代优化生成对抗样本时，由于网络更深，导致收敛速度变慢，训练周期变长，需要更多的时间和计算资源，这使得生成对抗样本的开销更大。

### 1.2.2视觉语言模态对抗样本生成方法

虽然Zhou等人[17]和Zhao等人[18]的工作证明了在视觉模态对抗样本生成过程中加入语言模态的信息，能够一定程度上提升其在视觉语言模态任务上的攻击性能，但是提升仍然有限。从实验数据上来看，现有的视觉模态对抗样本生成方法在视觉语言模态任务上的迁移性仍较差。因此，一些工作尝试在图像和文本上均进行扰动，研究设计面向视觉基础模型的视觉语言模态对抗样本生成方法。

Zhang等人[[[21]](#endnote-22)]首次在视觉基础模型上针对视觉语言模态任务提出了一种多模态交互的攻击方法，称作协同多模态对抗攻击（Co-Attack，Collaborative Multimodal Adversarial Attack）。对于采用融合策略的视觉基础模型，Co-Attack通过使扰动后的多模态嵌入远离原始多模态嵌入来实现协同扰动文本和图像，而对于采用对比学习来使视觉与语言模态对齐的视觉基础模型，Co-Attack通过使扰动后的图像嵌入远离扰动后的文本嵌入来实现对单模态嵌入的攻击。具体而言，他们提出先通过BERT-Attack[[[22]](#endnote-23)]生成文本扰动，然后将文本扰动作为文本输入，采用类似PGD的方式生成图像扰动，采用KL（Kullback-Leibler）发散损失[[[23]](#endnote-24)]来最大化嵌入表示。他们认为Co-Attack相对于不考虑多模态交互的方法能够产生更大的扰动。然而，Co-Attack以单一图像-文本对为输入，对抗样本会对训练的图像-文本对过拟合。

Co-Attack只考虑了白盒场景，为此，Lu等人[[[24]](#endnote-25)]提出了一种集合级引导攻击（SGA，Set-level Guidance Attack），并首次探索了适应于视觉语言模态任务的视觉基础模型在黑盒场景下的对抗鲁棒性。他们在Co-Attack基础上将单一的图像-文本对扩展为图像集-文本集，并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。具体而言，SGA通过数据增强将输入的图像扩展为图像集，接着为图像集中的每一张图像匹配相近的多个文本描述构成文本集，然后为文本集中每个文本描述生成对应的对抗文本，形成对抗文本集，再通过对抗文本集优化生成对抗图像。此外，在迭代优化对抗图像和对抗文本的过程中，SGA会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，以提升对抗样本的迁移性。SGA虽然实现了出色的跨模型迁移性，但是其生成对抗样本的开销非常大，从原本的单个图像-文本对扩展为图像集-文本集，计算量会成倍增加。此外，虽然SGA通过增强方法将单一的图像-文本对扩展为图像集-文本集，但是仍然是以单一的图像-文本作为输入，且采用的数据增强方法并不是生成新信息，而是混合现有信息，这会使得图像集与输入图像更加相关，进而导致生成的对抗样本对输入图像过拟合。

Wang等人[[[25]](#endnote-26)]从增强迁移性的角度同样基于对比学习提出了一种在视觉基础模型上生成视觉语言模态对抗样本的方法。受Guo等人[[[26]](#endnote-27)]启发，他们使用Gumble-softmax分布[[[27]](#endnote-28)]将离散分布的文本转化为一个可微的采样过程，使得能够通过梯度下降优化生成文本扰动。在生成视觉语言模态对抗样本过程中，他们结合了MI-FGSM[[[28]](#endnote-29)]和DI-FGSM[[[29]](#endnote-30)]方法来优化图像扰动，不同于Co-Attack方法，他们基于梯度同时生成对抗文本和对抗图像。为了提高视觉语言模态对抗样本的可迁移性，他们采用对比学习，包括模态内对比学习和图像-文本对比学习，在不同模态使对抗样本特征远离原始样本特征。此外，他们还采用了负对数似然损失来提高对抗文本的流畅性，采用BERT分数[[[30]](#endnote-31)]来约束对抗文本中的语义变化，以避免与原始文本产生较大的语义差异。他们将上述损失合并作为生成多模态对抗样本的最终损失函数，并通过超参分别控制不同损失的约束强度。实验表明，所该方法优于Co-Attack，能够在视觉模态和视觉语言模态任务中都具有更高的攻击性能，且在不同视觉基础模型具有良好的迁移性。然而，由于基于单个图像-文本对生成对抗样本，同样会出现对训练数据过拟合的问题。

Yin等人[[[31]](#endnote-32)]提出了一种任务不可知且能够在黑盒场景下攻击视觉基础模型的对抗样本生成方法VLAttack，能够适应于视觉模态和视觉语言模态的多种任务，如图像分类、视觉问答、图像描述、视觉蕴涵等。不同于Co-Attack，VLAttack先对图像进行攻击，他们提出了一种新的块相似性攻击（BSA，Block-wise Similarity Attack）策略。他们通过研究发现，如果预训练模型从干净输入和扰动输入中学习的特征表示显著不同，则这种扰动有可能可以迁移到微调的模型上，以改变其对下游任务的预测。他们通过最大化预训练模型的图像编码器和Transformer编码器的基于块的中间表示之间的余弦距离来优化图像扰动。此外，他们认为BERT-Attack[22]方法生成的语言模态对抗样本已足够强大，因此直接采用BERT-Attack优化文本扰动。为了充分学习不同模态扰动之间的相互联系，他们提出了一种新的迭代交叉搜索攻击（ICSA，Iterative Cross-Search Attack）策略。具体而言，VLAttack通过多次迭代来优化对抗图像，在第一次对抗图像生成中使用干净文本作为BSA的文本输入，在之后的每次迭代中使用多个对抗文本作为BSA的文本输入。然而，VLAttack过分关注输入图像的特征，并通过迭代加强了对图像整体特征的学习，这会导致对抗样本过拟合到训练数据。当输入图像改变时，需要重新训练来生成新的对抗样本。

综上所述，现有的视觉语言模态对抗样本生成方法均存在对训练数据过拟合的问题，当输入数据发生改变时，需要重新计算扰动来生成与之对应的对抗样本。此外，视觉语言模态对抗样本生成方法大都基于视觉基础模型训练生成对抗样本，同样存在训练开销大的问题。

总体而言，虽然已有一些工作在视觉基础模型上针对对抗样本的迁移性进行了研究，但面向视觉基础模型的对抗样本生成的研究还处于早期阶段，仍有许多问题尚未得到探索。

## 研究目标与创新性

本文的研究目标是针对对抗样本的迁移性，结合现有视觉基础模型，探究现有方法中的不足，并提出更加高效可用的解决方案，最后加以综合，形成一套完善的、面向视觉基础模型的对抗样本攻击系统。

本文的创新性首先在于考虑将集成模型训练应用于面向视觉基础模型的对抗样本生成，在不降低对抗样本攻击性能的情况下，提升对抗样本在不同任务上的迁移性，并降低训练开销。其次，本文在现有的视觉语言模态对比学习方法中引入对通用扰动表征的学习，一方面提升在视觉语言模态任务上的迁移性，另一方面强化对视觉模态和语言模态的通用扰动表征的学习，提升对抗样本对输入图像和文本的鲁棒性，避免对训练数据过拟合。

# 研究内容与技术路线

## 研究内容

现有的面向视觉基础模型的对抗样本生成方法主要存在以下几个方面的不足：（1）基于单个视觉基础模型训练，使得对抗样本只在某一类任务上具有迁移性，即跨任务迁移性差，且视觉模态对抗样本在视觉语言模态任务上迁移性较差；（2）当前面向视觉基础模型的对抗样本生成方法均以视觉基础模型作为替代模型，而视觉基础模型参数规模大、网络深，这使得需要耗费更多的时间和计算资源来训练，且需要硬件具有足够的内存空间来存储模型参数等信息，以支持视觉基础模型的运行，整体开销过大；（3）现有的视觉语言模态对抗样本生成方法均基于输入的单一图像和文本优化扰动，这会使得对抗样本对训练数据过拟合，影响攻击性能。

针对以上问题，本文制定了以下三个方面的研究内容（如图2）：

1. 基于集成模型训练的对抗样本生成方法：本文拟通过集成多个参数规模较小的深度模型作为替代模型，以在不降低对抗样本攻击性能的情况下减小训练的开销。同时本文考虑集成不同任务的不同深度模型，以提升对抗样本的跨任务迁移性。
2. 基于通用扰动表征对比学习的对抗样本生成方法：本文拟采用以通用扰动为对比视图的对比学习方法，来强化对抗样本对通用扰动表征的学习。结合模态内和模态间的通用扰动对比学习来提升对抗样本对输入图像和文本的鲁棒性，避免对训练数据过拟合。此外，在训练过程中借鉴集成模型训练方法，减小视觉语言模态对抗样本生成过程中的训练开销，并提升通用对抗扰动的跨任务迁移性。
3. 面向视觉基础模型的跨模态跨任务对抗样本攻击系统：本文将上述两种对抗样本生成方法加以综合，拟构建一套面向视觉基础模型的跨模态跨任务对抗样本攻击系统。该系统能够以相对更小的训练成本实现对适应于视觉模态、视觉语言模态不同任务的视觉基础模型的对抗样本攻击，并能够在任意图像和文本输入上欺骗视觉基础模型。

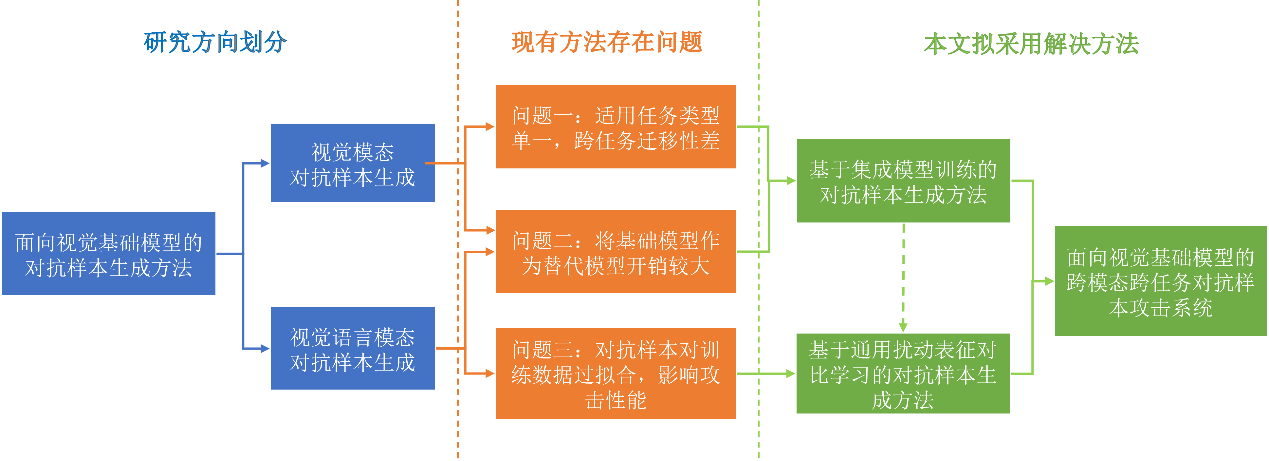


图2 现有方法存在的问题和本文的研究方案

## 基于集成模型训练的对抗样本生成方法技术路线

集成模型训练的思想最早用于提升模型在任务上的性能和模型的鲁棒性。Liu等人[[[32]](#endnote-33)]首次提出采用多个模型训练生成对抗样本，他们提出一个假设，即对抗样本如果能够对多个模型保持对抗性，那么它更有可能迁移到其他模型。这个假设首先由Goodfellow等人[6]进行了验证，之后基于FGSM方法改进的MI-FGSM[28]、DI-FGSM[29]以及TI-FGSM[[[33]](#endnote-34)]方法均进行了基于集成模型攻击的实验，并验证了其能够有效提升攻击迁移性。但是现有的集成模型攻击方法都是针对传统视觉模型，且均针对分类任务，在适应多个任务的视觉基础模型方面，还没有相关的研究。采用集成模型训练是否能够提升对抗样本在视觉基础模型上的攻击性能，以及采用什么样的集成模型训练能更大程度提升对抗样本的迁移性，都是还未得到探索和验证的问题。

根据现有研究，对抗样本的迁移性可以分为三个层面：一是跨数据迁移性，即在不同类数据集上具有迁移性；二是跨模型迁移性，即在同类任务上的不同模型上具有迁移性；三是跨任务迁移性，即在不同类任务的模型上具有迁移性。根据现有方法存在的问题，本文重点关注对抗样本的跨任务迁移性。本文认为对抗样本跨任务迁移性差可能出于三个原因，一是目标模型决策边界与训练模型存在一定差异，并不完全重合，导致能够改变训练模型决策边界的对抗样本可能在目标模型上失效；二是因为在特定模型上训练生成的对抗样本过拟合到该模型上，导致其在其他模型上迁移性差；三是不同任务所需要的特征存在差异，在一类任务上训练生成的对抗样本过拟合到该类任务，在其他任务上迁移性差。此外，现有的面向视觉基础模型的对抗样本生成方法均以视觉基础模型作为替代模型，而相比于深度模型，基础模型的参数规模更大、网络更深，这要求硬件必须具有足够内存空间存储模型的参数等信息以支持基础模型的运行，且由于网络规模大且深，在训练过程中需要耗费更多的时间和计算资源，这使得生成对抗样本的整体开销较大。

基于以上分析，本文认为可以通过集成不同任务的不同模型作为替代模型来生成对抗样本，使对抗样本能适应于不同的情形，并尽可能避免对抗样本对模型或任务过拟合的问题。此外，本文认为通过采用规模较小、网络较浅的深度模型进行集成，且集成模型的整体规模小于视觉基础模型，可以减小训练开销。本文认为可以通过优化以下问题来生成扰动：

(2.1)

其中表示在第个模型上计算得到的损失，表示第个模型的权重，表示某种特征相似度度量。受Zheng等人[10]的实验启发，这里将对抗样本与干净样本的相似性作为正则项引入，以提升模型对对抗样本特征的学习。

此外，本文提议参考Chen等人[[[34]](#endnote-35)]提出的AdaEA方法设计一套自适应的权重优化策略，通过监测每个模型对对抗样本的贡献差异，自适应地控制每个模型输出的融合，即上式中的。

## 基于通用扰动表征对比学习的对抗样本生成方法技术路线

在面向视觉基础模型的视觉语言模态对抗样本生成方面，已有一些研究基于多模态对比学习来使对抗扰动充分学习视觉模态和语言模态之间的内在联系。然而，现有的研究大都以图像为中心，或以图像-文本对为中心，即采用最小化对抗样本与干净样本之间的相似性，来使对抗样本朝着远离原始样本的方向优化。这些方法会导致对抗样本对输入的图像或图像-文本对过拟合，当面对不同图像时，需要重新计算扰动，生成与之对应的对抗样本，这不仅影响攻击性能，而且还增加了生成对抗样本的开销。

Moosavi-Dezfooli等人[13]首次针对模型输入研究对抗样本的可迁移性，提出了一种能够在任何图像上以较高概率欺骗模型的扰动，并称作通用对抗扰动（UAP，Universal Adversarial Perturbation）。随后，Ilyas等人[11]关于对抗样本可解释性进行分析，并提出对抗样本属于一种特征，而不是模型中存在的bug。他们认为对抗样本主要利用了数据中不稳定的特征，而这些特征是导致模型对对抗样本表现出脆弱性的原因。在面向视觉基础模型的研究中，Han等人[12]在SAM上提出了一种基于对比学习的UAP生成框架，他们证明了在视觉基础模型上也存在这样的通用对抗扰动。综合以上研究，本文认为通用对抗扰动具有其独有的特征，使其能够在任何图像上欺骗模型。因此，本文拟采用以通用对抗扰动为中心的方法来解决对训练数据过拟合的问题。

本文将采用对比学习方法来充分学习通用扰动的特征表示，即先基于替代模型生成通用对抗扰动，然后将通用对抗扰动作为对比视图，将增强的通用对抗扰动作为正样本，将随机采样的干净样本作为负样本，使对抗样本朝着通用扰动独有特征的方向优化。同时，本文还将借鉴集成模型训练方法，将集成模型作为替代模型，以减小视觉语言模态对抗样本的训练开销。本文计划采用模态内和模态间的通用扰动表征对比学习，以加强对视觉模态和语言模态通用扰动表征的学习，并充分利用不同模态间的作用关系，如图3。具体而言，在视觉模态和语言模态分别以通用图像扰动和通用文本扰动为对比视图，进行模态内的通用扰动表征对比学习，以避免对输入的图像或文本过拟合。优化生成阶段性通用对抗扰动后，再将通用图像扰动及其增强作为通用文本扰动对比学习的正样本，以通用文本扰动及其增强作为通用图像扰动对比学习的正样本，以随机采样的干净样本作为负样本，进行模态间的通用扰动表征的对比学习，以避免对输入的图像-文本对过拟合。通过在模态内和模态间基于通用扰动表征的对比学习，避免对抗样本对训练数据的过拟合，实现在任何图像文本输入上欺骗视觉基础模型。

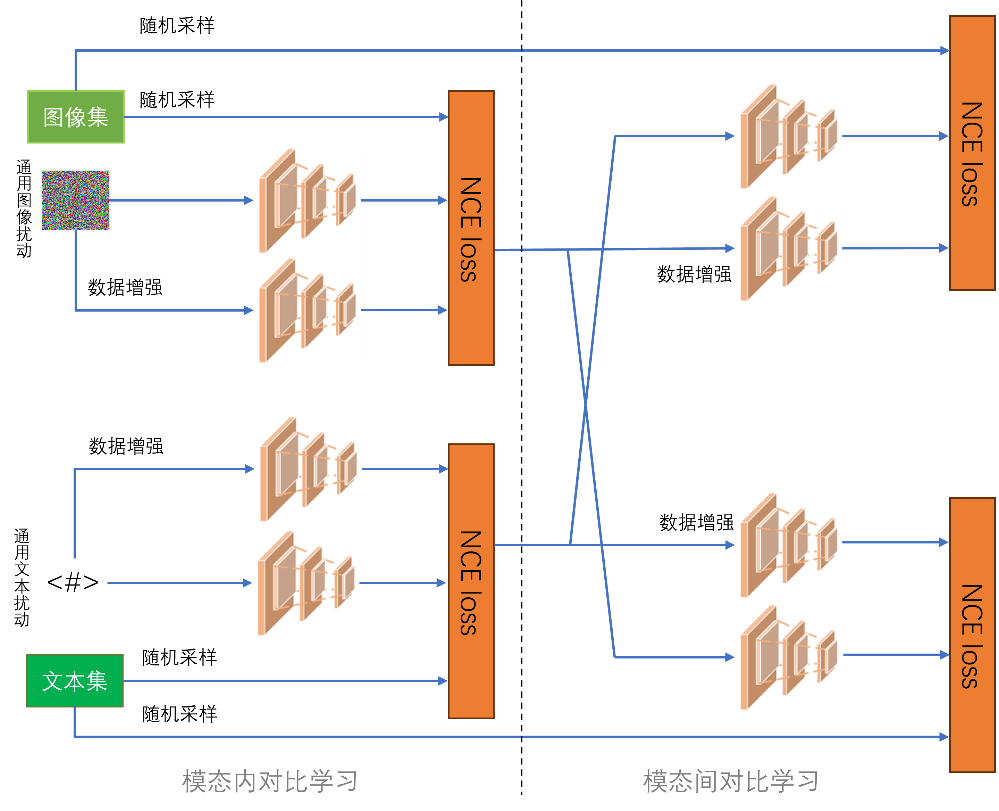


图3 基于通用扰动表征对比学习的对抗样本生成框架

# 论文工作安排计划

## 工作进度安排

本毕业论文拟定的研究计划如下：

（1）2023年11月至2023年12月，查阅文献，确定问题，完成开题。

（2）2023年12月至2024年3月，针对集成模型训练方法展开研究，通过实验探索不同集成模型对提升对抗样本在视觉基础模型上攻击迁移性的作用，并验证集成模型在训练上的开销。

（3）2024年3月至2024年6月，针对基于通用扰动表征的对比学习方法展开研究，分析基于视觉模态和文本模态通用扰动表征对比学习对提升对抗样本在视觉基础模型上攻击迁移性的作用并进行实验验证。

（4）2024年6月至2024年9月，对上述方法进行整合，构建一个面向视觉基础模型的跨模态跨任务对抗样本攻击系统，并在多种视觉基础模型上进行实验验证。

（5）2024年9月至2024年11月，完善实验，撰写毕业论文，准备毕业答辩。

## 关键技术及难点

### 3.2.1如何确定集成模型构成以及各模型权重

在2.2小节中，本文提出了基于集成模型训练的对抗样本生成方法。如何选择模型进行集成，以及如何设置所选的模型在集成模型中的权重，是设计一个基于集成模型训练的对抗样本生成方法所要面临的重要命题。现有的方法都是针对传统视觉模型，而视觉基础模型得益于其庞大的规模，能够适应多个下游任务，现有的方法在视觉基础模型上并不适用。那么如何通过集成多个深度模型实现在降低替代模型整体规模的情况下，达到甚至超过在视觉基础模型上训练生成的对抗样本的攻击性能，是本文面临的一个重要难点。此外，不同任务的模型在集成模型中的权重对所生成的对抗样本跨任务迁移性的影响，同样是本文需要讨论和验证的问题。

### 3.2.2如何选择增强方法以及设计合适的对比学习损失函数

在2.3小节中，本文提出了基于通用扰动表征对比学习的对抗样本生成方法。在对比学习方法中，基本思想是以关注的对象作为对比视图，然后采用正样本和负样本进行优化。本文拟采用通用扰动作为对比视图，然后采用增强的通用扰动作为正样本，干净样本作为负样本。然而，不同的增强方法可能会对通用扰动表征学习产生不同的影响，以及对文本扰动的增强方法是否能够带来收益也是需要实验验证的问题。此外，在视觉模态内的对比学习以及视觉语言模态间的对比学习中所采用的对比学习损失函数也会对通用扰动表征学习产生影响。采用什么增强方法以及如何设计对比学习损失函数能够更好的学习到通用扰动表征，是本文的重点工作之一。

# 主要参考文献

1. [] Devlin J, Chang M, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C] // North American Chapter of the Association for Computational Linguistics (2019). NAACL, 2019. [↑](#endnote-ref-2)
2. [] Ramesh A, Pavlov M, Goh G, et al. Zero-Shot Text-to-Image Generation[C] // Proceedings of the 38th International Conference on Machine Learning, PMLR 139:8821-8831, 2021. [↑](#endnote-ref-3)
3. [] Brown T, Mann B, et al. Language Models are Few-Shot Learners[C] // 34th Conference on Neural Information Processing Systems (NeurIPS). NIPS, 2020, pp. 1877-1901. [↑](#endnote-ref-4)
4. [] Bommasani R, Hudson D, Adeli E, Altman R, et al. On the opportunities and risks of foundation models[R/OL]. https://crfm.stanford.edu/report.html. 2021. [↑](#endnote-ref-5)
5. [] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners[J]. OpenAI blog, 1(8):9, 2019. [↑](#endnote-ref-6)
6. [] Goodfellow I, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Ex-amples[EB/OL]. arXiv:1412.6572, 2014. [↑](#endnote-ref-7)
7. [] Fort S. Adversarial examples for the OpenAI CLIP in its zero-shot classification regime and their semantic generalization[DB/OL]. https://stanislavfort.com/blo-g/OpenAI\_CLIP\_adversarial\_examples. 2021. [↑](#endnote-ref-8)
8. [] Kirillov A, Mintun E, Ravi N, et al. Segment anything[EB/OL]. arXiv:2304.02643, 2023. [↑](#endnote-ref-9)
9. [] Zhang C, Zhang C, Kang T, Kim D. Attack-SAM: towards attacking segment anything model with adversarial examples[EB/OL]. arXiv:2305.00866, 2023. [↑](#endnote-ref-10)
10. [] Zheng S, Zhang C. Black-box targeted adversarial attack on segment anything (SAM)[EB/OL]. arXiv:2310.10010, 2023. [↑](#endnote-ref-11)
11. [] Ilyas A, Santurkar S, Tsipras D, et al. Adversarial examples are not bugs, they are features[C]//33th Conference on Neural Information Processing Systems (NeurIPS). NIPS, 2019:125-136. [↑](#endnote-ref-12)
12. [] Han D, Zheng S, Zhang C. Segment Anything Meets Universal Adversarial Per-turbation[EB/OL]. arXiv:2310.12431, 2023. [↑](#endnote-ref-13)
13. [] Moosavi-Dezfooli S, Fawzi A, Fawzi O, et al. Universal Adversarial Perturbations [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017, pp. 1765-1773. [↑](#endnote-ref-14)
14. [] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[C] // 6th International Conference on Learning Representations (ICLR). ICLR, 2018. [↑](#endnote-ref-15)
15. [] Moosavi-Dezfooli, Fawzi A, Frossard P. DeepFool: A Simple and Accurate Me-thod to Fool Deep Neural Networks[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016, pp. 2574-2582. [↑](#endnote-ref-16)
16. [] Oord A, Li Y, Vinyals O. Representation Learning with Contrastive Predictive Co-ding[EB/OL]. arXiv:1807.03748, 2019. [↑](#endnote-ref-17)
17. [] Zhou Z, Hu S, Li M, et al. AdvCLIP: Downstream-agnostic Adversarial Examples in Multimodal Contrastive Learning[C] // Proceedings of the 31th ACM International Conference on Multimedia (MM). ACM MM, 2023, pp. 6311-6320. [↑](#endnote-ref-18)
18. [] Zhao Y, Pang T, Du C, et al. On Evaluating Adversarial Robustness of Large Vision-Language Models[EB/OL]. arXiv:2305.16934, 2023. [↑](#endnote-ref-19)
19. [] Yuksekgonul M, Bianchi, et al. When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It?[C] // 11th International Conference on Learning Representations (ICLR). ICLR, 2023. [↑](#endnote-ref-20)
20. [] Rombach R, Blattmann A, Lorenz D, et al. High-Resolution Image Synthesis With Latent Diffusion Models[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022, pp. 10684-10695. [↑](#endnote-ref-21)
21. [] Zhang J, Yi Q, Sang J. Towards Adversarial Attack on Vision-Language Pre-tra-ining Models[C] // Proceedings of the 30th ACM International Conference on Multimedia (MM). ACM MM, 2022, pp. 5005-5013. [↑](#endnote-ref-22)
22. [] Li L, Ma R, Guo Q, et al. BERT-ATTACK: Adversarial Attack Against BERT Using BERT[C] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). EMNLP, 2020, pp. 6193-6202. [↑](#endnote-ref-23)
23. [] Zhang H, Yu Y, Jiao J, et al. Theoretically Principled Trade-off between Robustness and Accuracy[C] // Proceedings of the 36th International Conference on Machine Learning, PMLR 97:7472-7482, 2019. [↑](#endnote-ref-24)
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 (ICCV). IEEE, 2023, pp. 102-111. [↑](#endnote-ref-25)
25. [] Wang Y, Hu W, Dong Y, et al. Exploring Transferability of Multimodal Adversarial Samples for Vision-Language Pre-training Models with Contrastive Learning[EB/OL]. arXiv:2308.12636, 2023. [↑](#endnote-ref-26)
26. [] Guo C, Sablayrolles A, et al. Gradient-based Adversarial Attacks against Text Transformers[EB/OL]. arXiv:2104.13733, 2021. [↑](#endnote-ref-27)
27. [] Jang E, Gu S, Poole B. Categorical Reparameterization with Gumbel-Softmax [EB/OL]. arXiv:1611.01144, 2016. [↑](#endnote-ref-28)
28. [] Dong Y, Liao F, Pang T, et al. Boosting Adversarial Attacks With Momentum[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018, pp. 9185-9193. [↑](#endnote-ref-29)
29. [] Xie C, Zhang Z, Zhou Y, et al. Improving Transferability of Adversarial Examples With Input Diversity[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019, pp. 2730-2739. [↑](#endnote-ref-30)
30. [] Zhang T, Kishore V, Wu F, et al. BERTScore: Evaluating Text Generation with BERT[EB/OL]. arXiv:1904.09675, 2020. [↑](#endnote-ref-31)
31. [] Yin Z, Ye M, Zhang T, et al. VLAttack: Multimodal Adversarial Attacks on Vision-Language Tasks via Pre-trained Models[EB/OL]. arXiv:2310.04655, 2023. [↑](#endnote-ref-32)
32. [] Liu Y, Chen X, Liu C, Song D. Delving into Transferable Adversarial Examples and Black-box Attacks[C] // 5th International Conference on Learning Representations (ICLR). ICLR, 2017. [↑](#endnote-ref-33)
33. [] Dong Y, Pang T, Su H, Zhu J. Evading Defenses to Transferable Adversarial Ex-amples by Translation-Invariant Attacks[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019, pp. 4312-4321. [↑](#endnote-ref-34)
34. [] Chen B, Yin J, Chen S, et al. An Adaptive Model Ensemble Adversarial Attack for Boosting Adversarial Transferability[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023, pp. 4489-4498. [↑](#endnote-ref-35)