

**北京航空航天大学计算机学院**

**硕士研究生学位论文**

**文献综述**

**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

摘 要

近年来，基础模型方面的研究取得了显著成效，它们在各种应用中显示出了巨大的前景。随着基础模型在自然语言处理领域取得突破性进展，计算机视觉领域也逐步走上从特定任务视觉模型到视觉基础模型的道路。

**关键词**：对抗样本

Abstract

In recent years, research on foundation models has yielded remarkable results, and they show great promise in a variety of applications. With the breakthrough of foundation models in the field of natural language processing, the field of computer vision has gradually embarked on the road from task-specific vision models to vision foundation models.

**Keywords：**Adversarial Sample;

目 录

1 视觉语言模型的研究现状 1

1.1 基于对比学习的VLMs 1

1.2 基于掩码图像建模的VLMs 1

1.3 基于大语言模型的VLMs 1

1.4 生成式VLMs 1

2 传统视觉模型对抗样本生成 2

2.1 对抗样本生成方法研究现状 2

2.2 对抗样本迁移性研究现状 2

3 结论 3

参考文献 4

图 目

[**图1 视觉基础模型体系结构** 3](#_Toc152589826)

[**图2 FLAVA架构**[7] 5](#_Toc152589827)

[**图3 FGSM攻击深度视觉模型导致高置信的错误预测**[30] 6](#_Toc152589828)

[**图4 以扰动为中心的攻击框架**[62] 13](#_Toc152589829)

[**图5 AdvCLIP攻击框架**[64] 14](#_Toc152589830)

[**图6 Co-Attack优化扰动原理**[69] 15](#_Toc152589831)

[**图7 块相似性攻击（BSA）**[77] 17](#_Toc152589832)

# 视觉语言模型的研究现状

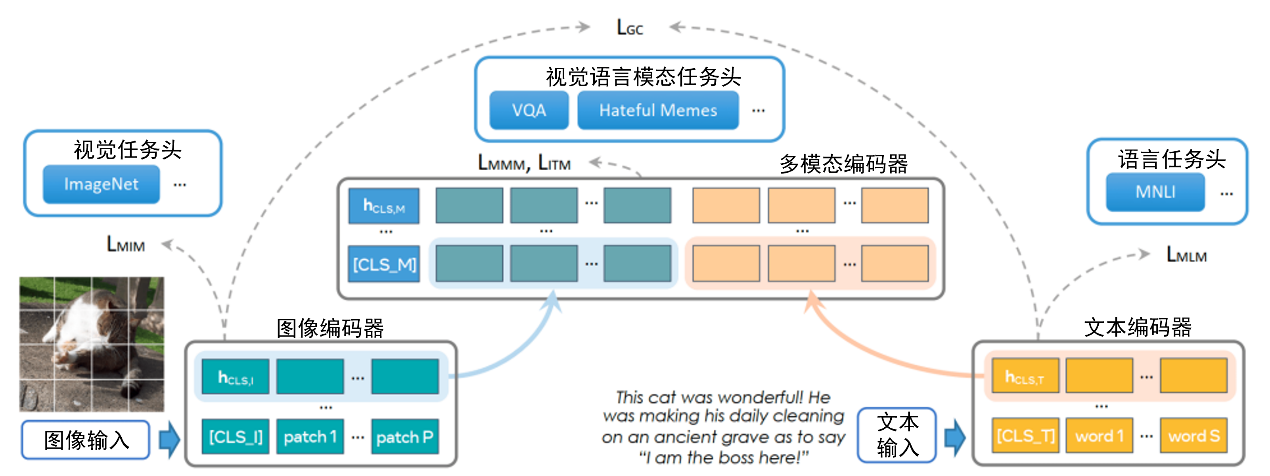
视觉语言模型（Visual-Language Models, VLMs）结合了计算机视觉与自然语言处理技术，旨在实现图像与文本的多模态信息理解。它们在图像描述生成（Image Captioning, IC）、视觉问答（Visual Question Answering, VQA）和视觉定位（Visual Grounding, VG）等跨模态任务中展现了巨大的潜力。近年来，Transformer架构在自然语言处理和计算机视觉领域的广泛应用，为视觉语言模型的发展提供了强大支持。得益于Transformer的自注意力机制，模型能够同时处理图像与文本数据，为多模态信息理解提供了创新解决方案。因此，基于Transformer的视觉语言模型已逐渐成为多模态任务研究的主流，并在多个应用领域取得了显著进展。

## 基于对比学习的VLMs

## 基于掩码图像建模的VLMs

## 基于大语言模型的VLMs

## 生成式VLMs



**图2 FLAVA架构**

# 传统视觉模型对抗样本生成

## 对抗样本生成方法研究现状

## 对抗样本迁移性研究现状

# 结论

# 参考文献

3. Vaswani A., Shazeer N., Parmar N., et al. Attention is All you Need[C]. Conference on Neural Information Processing Systems. 2017: 5998-6008.
4. V