2024-2025 学年秋季学期

《计算思维实训（1）》(0830A030)

课程报告

成绩	
（百分制）	

学　　号		学　　院	计算机工程与科学学院
姓　　名		手工签名	
报告题目	大语言模型在计算机视觉领域的应用研究		
实训报告成绩 50%	实训过程描述：清晰、全面。（5%）		
	报告主要部分：1、计算思维实例叙述清楚、有意义；2、分析到位；3、思路独特或有创意；4、关键实现技术描述清楚详细；5、内容丰富；6、不直接粘贴所有代码；7、代码有注释或说明。（25%）		
	实训收获体会：感受真实、深刻，建议有价值。（10%）		
	书写格式：书写规范、用词正确、无明显的错别字，图表、代码清晰规范，格式协调、效果好，参考文献书写、引用规范合理。（10%）		
工作实绩 50%	1、平时表现、进步情况（25%），2、工作实绩（25%）		

教师对该生工作实绩简要评述：

教师签名：

日期：　　年　　月　　日

Part 1. 本学期实训过程概述

1. 实训总体概况、实训过程

本学期的实训主题为“计算机视觉与大语言模型的结合”，主要任务是通过调研、学习和实践，了解大语言模型在计算机视觉领域的应用，并形成调研报告。在实训过程中，我分为几个阶段进行学习和探索。

首先，我进行了文献调研，重点研究了大语言模型（LLM）如 GPT 系列和 CLIP 在视觉任务中的应用。调研涵盖了图像描述生成、视觉问答、跨模态生成等典型任务。其次，实训过程中我通过分析现有模型架构和技术，如 Vision Transformer（ViT）和 DALL-E，理解了视觉模型与语言模型结合的原理和实现方式。

2. 对计算思维的认识

通过本次实训，我对计算思维有了更深入的理解。计算思维不仅仅是编写代码或解决具体问题的方法，它更是一种以系统化、抽象化的方式来理解问题的思维方式。在大语言模型与计算机视觉的结合中，计算思维帮助我们从宏观角度分析语言与视觉信息的融合，从微观角度优化模型的结构与任务解决方案。计算思维使得我们能够将问题分解，抽象出核心要素，进而找到最佳的解决路径。

3. 实训体会及建议

此次实训使我深刻体会到大语言模型的强大之处以及它与计算机视觉结合的广阔前景。在项目调研和实践过程中，我感受到了理论与实践结合的重要性。通过调研文献，我了解到跨模态学习的前沿技术和发展趋势；在实训实践中，我进一步巩固了这些理论知识，并提升了动手能力。

然而，整个实训过程中也有一些挑战。例如，大语言模型的计算资源需求较大，许多高精度的实验在我们现有的条件下无法完全实现。对此，我建议在未来的实训中，可以增加关于云计算或分布式计算的实践内容，以弥补硬件资源的不足。此外，未来的实训可以考虑更多的实验环节，通过提供更多的小型实践任务，使同学们能够更好地巩固和消化所学知识。

Part 2. 综合实训报告

大语言模型在计算机视觉领域的应用研究

23121538 郭咏钦

计算机工程与科学学院

摘要：大语言模型（LLM）的发展重塑了自然语言处理的格局，而随着多模态技术的进步，LLM 也逐渐进入了计算机视觉领域。通过将图像和文本信息进行有机融合，这些模型在图像描述生成、视觉问答、跨模态生成等任务中展现了强大的潜力。本文深入探讨了 LLM 在计算机视觉中的主要应用及其关键技术，并展望了其在未来的潜在发展方向。

关键字：大语言模型，计算机视觉，图像描述生成，视觉问答，跨模态学习，Transformer

一、引言

计算机视觉与自然语言处理（NLP）曾经是两个相对独立的领域，前者主要处理图像、视频等视觉数据，后者则致力于理解和生成语言。然而，随着多模态学习的崛起，视觉与语言的结合日益成为人工智能领域的研究热点。大语言模型（LLM）如 GPT 系列，在处理大规模自然语言数据上取得了显著成功，而通过与视觉模型的结合，LLM 开始在图像描述生成、视觉问答等视觉任务中发挥重要作用。[1]

通过结合视觉和语言数据，LLM 不仅能够生成高质量的自然语言描述，还能够进行复杂的跨模态推理。这使得大语言模型在诸如增强现实、智能医疗、自动驾驶等多个领域展现出广泛的应用前景。[2]

二、实例分析

1. 图像描述生成

图像描述生成任务要求模型从图像中提取视觉特征，并将这些特征转化为自然语言。传统方法依赖于图像分类模型与预定义的词汇表，而 LLM 的引入使得生成的描述更加灵活、自然[3]。

以 CLIP 为代表的多模态模型通过对大量的图像-文本对进行训练，实现了图像与文本的双向关联。这种方法不仅提高了生成描述的准确性，还增强了模型的推理能力[1][3]。例如，给定一张复杂的公园风景图，

图表 1 公园风景图

模型不仅能识别出图像中的物体，还能生成关于场景的全局描述，捕捉到不同元素之间的关系。

图表 2GPT 描述图片

2. 视觉问答系统

视觉问答任务是一种要求模型结合图像和语言信息来回答关于图像问题的任务。在这种场景中，模型需要同时具备图像理解和语言推理的能力。这不仅考验视觉模型的特征提取能力，也对大语言模型的上下文理解和生成提出了高要求。

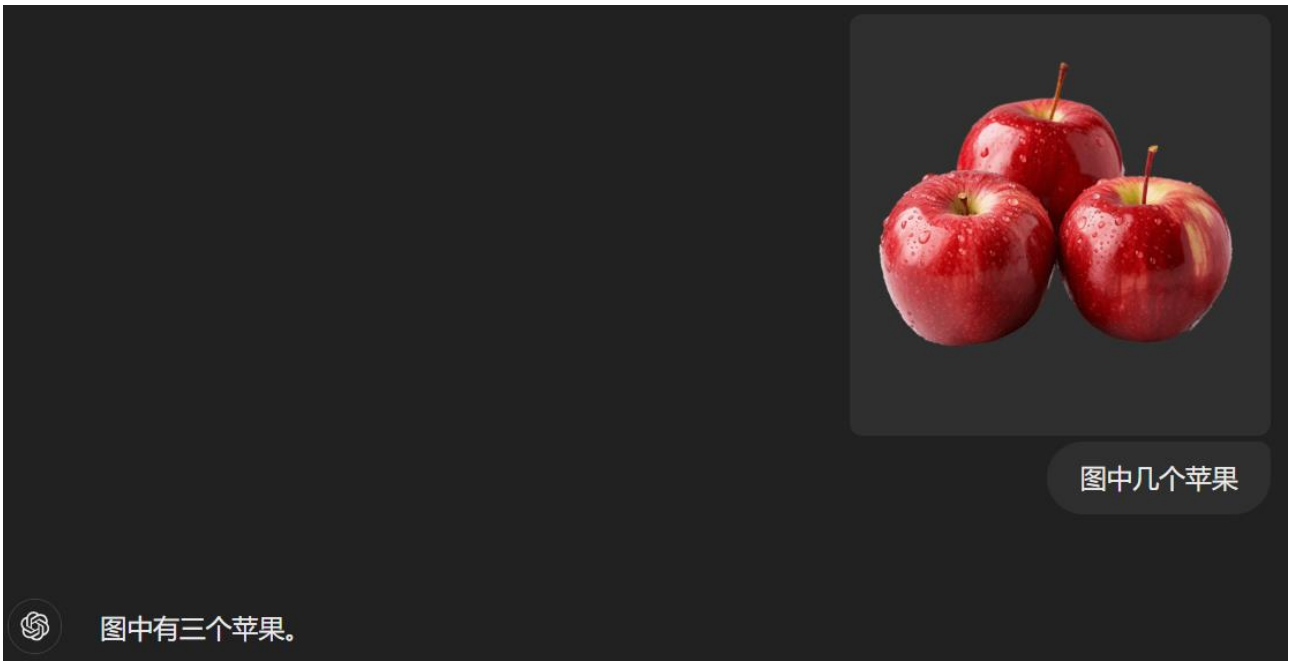
在现代视觉问答系统中，视觉特征提取往往依赖于 Transformer 架构的视觉模型，如 ViT（Vision Transformer），而问题的理解与回答生成则由大语言模型完成。通过共同的嵌入空间，

视觉和语言信息得以交互，模型可以在面对多样化问题时作出合理的回答[3]。这种方法不仅提升了视觉任务的语义理解深度，还使得生成的答案更加精确和灵活。

如我们可以测试一下对 AI 模型的图片思考能力。将此图片发给 AI，考考他图中有几个苹果。

图表 3 苹果测试图

AI 也能迅速的反应过来，并回答到图中有三个苹果。



图表 4 对 AI 的测试

三、关键技术分析

1. 跨模态对比学习（CLIP）

CLIP 模型通过将图像和文本映射到同一个嵌入空间，学习它们之间的关联性。这种跨模

态对比学习的思路摒弃了传统的分类方式，而是通过对比图像和文本的相似性，帮助模型建立更深层次的多模态关联。

CLIP 的核心创新在于它能够进行零样本学习（zero-shot learning），即在没有明确标注的情况下，模型可以基于学习到的图像-文本对的关联，对从未见过的图像进行准确的描述。这种能力极大拓展了模型的适用场景，也为图像描述生成等任务提供了新的解决思路[3]。

2. Vision Transformer (ViT)

Transformer 最早在 NLP 领域取得了巨大成功，而 ViT 模型将这一架构引入到图像处理中。与传统的卷积神经网络（CNN）不同，ViT 直接将图像分割为多个小块（patches），并使用 Transformer 对这些小块进行特征提取。ViT 的优势在于其全局注意力机制，能够捕捉到图像的全局特征，而不仅仅是局部信息。

ViT 不仅可以独立用于图像分类等任务，还能与大语言模型结合，形成强大的多模态架构。这种结合极大提高了模型在视觉问答、图像描述生成等任务中的表现，使得模型能够同时理解视觉与语言的语义信息。

3. DALL-E 和 Stable Diffusion

DALL-E 和 Stable Diffusion 是生成模型中的代表性技术，能够根据文本描述生成高质量的图像。DALL-E 在理解复杂的语言输入后，生成符合描述的图片，并能够灵活地组合各种抽象概念。而 Stable Diffusion 进一步优化了图像生成过程，通过扩散模型逐步减少噪声，在计算资源有限的情况下仍能生成高清图像。

这种从文本到图像的生成能力不仅适用于艺术创作领域，还可以应用于广告、设计等多个行业，为用户提供创意视觉内容。

四、应用场景与未来发展

1. 增强现实（AR）中的应用

增强现实（AR）技术通过叠加虚拟信息到现实世界中，帮助用户更好地理解和交互环境。在这种应用场景中，大语言模型可以提供即时的场景描述与解释功能。例如，智能眼镜可以通过摄像头捕捉周围环境，并通过大语言模型生成相关的实时语音提示，帮助用户获取更多的背景信息。

这种结合视觉与语言的信息流，将极大提升 AR 设备的智能化水平，使得用户能够更自然地与虚拟世界互动。

2. 医疗影像分析

在医疗领域，计算机视觉与大语言模型的结合为自动诊断和报告生成提供了新的可能性。例如，在放射影像的分析中，视觉模型可以识别病灶区域，而大语言模型则能够根据这些识别结果生成详细的医学报告。这种自动化流程不仅提高了医生的工作效率，也降低了人为误差的风险。

未来，随着更多医学影像数据的积累和大语言模型的优化，自动化诊断系统有望在复杂

病例中提供更为准确的诊断建议。

五、总结与展望

大语言模型的出现为计算机视觉领域带来了全新的技术路径。通过与视觉模型的结合，LLM 不仅拓展了语言生成的边界，还大大提升了视觉任务的理解与推理能力。在多模态学习的推动下，跨越语言与视觉的界限已成为可能[1][2]。

未来，随着模型规模的进一步扩大和训练数据的增加，多模态模型将更加准确地处理复杂任务，带来更多智能化、自动化的应用场景[3]。与此同时，如何有效地降低模型的计算成本、提高其对新任务的泛化能力，也将成为未来研究的重要方向。

参考文献

[1] Radford A, Kim J W, Hallacy C, et al. Learning Transferable Visual Models from Natural Language Supervision[J]. arXiv preprint arXiv:2103.00020, 2021.

[2] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[3] Ramesh A, Pavlov M, Goh G, et al. Zero-shot Text-to-Image Generation[J]. arXiv preprint arXiv:2102.12092, 2021.