VQABBoxpro: Enhancing Visual Question Answering with High-Resolution Image Input and Semantic-Linked Bounding Boxes

摘要

大模型出现后一些工作在 VQA 领域也进行多模态迁移，但是模型输入图片的像素大小有限以及模型无法指回图片语义相关位置。本文在 LLaVA: Visual Instruction Tuning 的基础上做了一些改进。模型架构上，采用了利用视觉特征投影到语言模型空间的策略，并引入了语言描述的相对坐标边界框以提升模型在视觉问答任务上的表现。数据集应用方面，使用了多个在视觉问答任务中常用的数据集，并针对模型可能出现的多任务处理问题开发了混合多任务对话数据集加入模型训练。这些改进提高了模型处理效率和多任务处理性能，并增强了模型在视觉问答任务上的表现。

关键词： VQA；MLLM; Visual Grounding;

1 引言

随着人工智能技术的持续发展，多模态大型语言模型（Large Language Models, LLMs）在视觉语言领域掀起了研究的热潮 [1,8]。这些模型融合了视觉与文本信息，广泛应用于视觉 AI 助手、图片描述生成、视觉问题回答（Visual Question Answering, VQA）以及指示表达理解（Referring Expression Comprehension, REC）等任务中。多模态大型语言模型可以继承自单模态 LLMs 的高级能力，比如逻辑推理、常识理解和强大的语言表达，从而在接收到适当的视觉语言指令后，展现出生成详细图像描述、编码生成、图像中视觉对象定位乃至完成复杂视觉问题的多模态推理等强大功能 [7]。LLaVA 就是这样的工作，因此我选择 Visual Instruction Tuning [5] 这篇文章。

尽管多模态大型语言模型展现出巨大潜力，但有效学习执行多种视觉语言任务并制定相应的多模态指令仍然是一个挑战。不同任务的内在复杂性使得其操作手段和回应有着多种可能性。例如，面对“告诉我人物的位置”这一用户输入，在不同的上下文背景下可能有不同的答案。在指示表达理解任务中，答案可能是人物的一个边界框位置；在视觉问题回答任务中，模型可能会用自然语言描述他们的空间位置；而在人物检测任务中，模型可能会识别给定图像中每个人的空间位置。因此需要实现更统一处理方法的模型，提高图片输入分辨率和加入边界框语言维度来描述图片。这样的改进将有助于增强模型对视觉信息的理解和更准确地执行与语言相关的任务。

2 相关工作

2.1 大语言模型（LLMs）

在大型语言模型的发展历程中，早期模型如 GPT-2 [4] 和 BERT [3] 标志着自然语言处理领域的重大突破。这些基础模型在规模庞大的文本数据集上进行训练，为 NLP 领域带来了创新。随后，为了提高容量和扩大训练数据，研究者们开发了一系列先进的 LLMs，包括 GPT-3、Megatron-turing NLG、PaLM、Gopher、Chinchilla、OPT 和 BLOOM 等。近期的工作重点在于优化 LLMs，使其更有效地与人类指导和反馈进行交互。代表性的研究包括 InstructGPT 和 ChatGPT，它们展示了强大的语言理解和执行复杂任务的能力，如写作改进和编码助手。

与 LLMs 的进展同时涌现的是 LLaMA [6] 语言模型的崛起。为了实现与 ChatGPT 类似的人类指导能力，一些研究试图使用高质量的指导性数据对 LLaMA 模型进行微调。其中包括 Alpaca、Vicuna 和 MPT 等模型，以及从人类反馈数据中学习的开源语言模型，如 Falcon 和 LLaMA-2，这些模型表现出令人印象深刻的性能。

2.2 大语言模型与视觉对齐（Visual Aligning）

LLMs 具有出色的泛化能力，因此研究者们将其扩展到多模态领域，通过将视觉输入与 LLMs 进行对齐。早期的研究如 VisualGPT 和 Frozen 使用预训练的语言模型来改善图像字幕和视觉问题回答任务的性能。这为后续的视觉语言研究奠定了基础，例如 Flamingo 和 BLIP-2。GPT-4 展示了许多先进的多模态能力，例如根据手写文本生成网站代码。这些能力启发了其他视觉语言 LLMs，包括 LLaVA [5] 和 MiniGPT-4 [9]，它们通过适当的指导性微调与大型语言模型 Vicuna 对齐，展示了对齐后的先进多模态能力。

一些工作如 Vision-LLM、Kosmos-2、Shikra [2]，也证明了多模态 LLMs 能够通过生成文本格式的边界框来执行视觉定位（Visual Grounding），通过语言模型生成边界框的文本格式，这进一步展示了在理解和融合不同模态信息方面的进展，尤其是在指令理解、多模态理解与生成能力上展现了巨大的潜力。

3 本文方法

3.1 本文方法概述

本文主要从模型架构和数据集两方面探讨了图像-语言模型的优化方法。首先，对于模型架构，本方案采取了进行适当的改进。对于线性投影层部分，采用了将视觉特征投影到语言模型空间的策略，并应用较高效的策略如 4 合 1 策略，减少了视觉输入标记的数量并大幅提高了处理效率。在大型语言模型部分，选用了开源的 LLaMA2-chat(7B) 作为主干网络且直接产生包含边界框的文本表示以便标注空间位置，统一了输入接口且提高了视觉定位任务的效能；对于数据集的应用，本文选择了几个在视觉问答任务中常用的数据集，如 OK-VQA，GQA，VSR，IconVQA，VizWiz 和 HM 等。这些数据集的利用大大支持了图像理解和自然语言处理任务的实现，并为模型的训练和评估提供了丰富多样的资源，是提升模型性能的关键环节。

3.2 模型架构

本模型包含三个核心部件：视觉主干网络、线性投影层和大型语言模型，具体配置如下：

视觉主干网络：本模型采用 EVA 模型作为视觉主干网络，以支持对输入图像的特征提取。在整个训练过程中，视觉主干网络的参数保持冻结。针对高分辨率的图像输入，模型采用的图像分辨率为 448×448，并且对位置编码做插值处理，以适配较高图像分辨率。

线性投影层：本模型旨在将冻结的视觉主干网络所提取的视觉特征投影到语言模型空间。为了有效处理如 448×448 这种高分辨率的图像输入，采用了将相邻的 4 个视觉特征在嵌入空间进行合并，再投影到大型语言模型的同一特征空间中的策略。这样不仅减少了视觉输入标记的数量，而且在训练和推理阶段大幅提高了处理效率。

大型语言模型：本模型采用了开源的 LLaMA2-chat(7B) 作为语言模型的主干网络。语言模型作为处理各类视觉-语言输入的统一接口。对于那些需要生成空间位置的视觉定位任务，选择让语言模型直接产生包含边界框的文本表示形式，以便标注空间位置。架构如图 1所示：

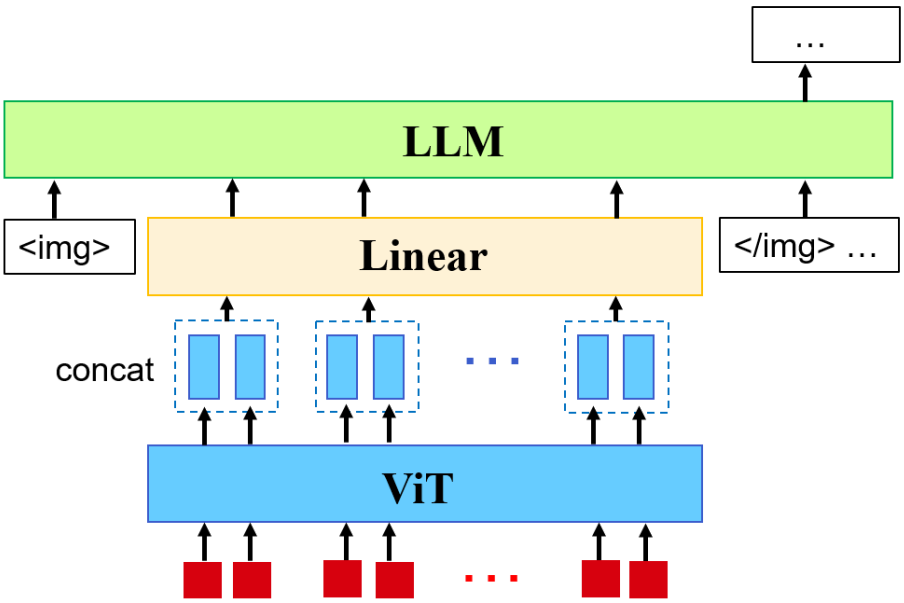

图 1. 模型架构

3.3 数据集介绍

视觉问答 (Visual Question Answering，VQA) 是一项结合了图像理解和自然语言处理的挑战性任务，旨在使机器能够根据给定的图像内容回答相关的问题。随着深度学习技术的快速发展，研究人员需要大规模、多样化的数据集来训练并评估他们的模型。下面详细介绍几个在 VQA 研究中常用的数据集：OK-VQA，GQA，VSR，IconVQA，VizWiz 以及 HM。

3.3.1 OK-VQA (Outside Knowledge Visual Question Answering)

OK-VQA 是一个专注于需要外部知识的视觉问答任务的数据集。该数据集涵盖了不同的知识领域，如科学、地理、历史等，旨在促进对于需要普遍常识的问题的研究。它包含了超

过 14,000 张图像和 14,000 个问题-答案对，每个问题都需要对图像内容和世界知识的理解来正确回答。

3.3.2 GQA (eneral Question Answering)

GQA 数据集旨在提升 VQA 系统中的推理能力。GQA 包含了超过 113k 张图像，和 22M 的问题-答案对，这些问题是基于图像内容自动生成的，并且设计成涵盖不同类型的推理，如对象识别、属性识别、计数、比较、分类和联合等。GQA 的目标是促使模型更好地理解视觉场景，并在复杂的推理任务上展示其能力。

3.3.3 VSR (Visual Semantic Reasoning)

VSR 是一个用于衡量深度学习模型在复杂视觉-语义推理任务上性能的数据集。它由图像以及需要对图像内容进行深入推理的问题-答案对组成。VSR 的挑战在于，它要求模型并非仅仅识别图像中的物体，而是需要进行抽象思维和推理，比如理解事件和情感，从而提供正确的答案。

3.3.4 IconVQA (Icon Visual Question Answering)

IconVQA 是专门为图标类图像设计的 VQA 数据集。它包括大量图标和与之相关的问题，强调对抽象图形的理解和问答。由于图标通常是简化和符号化的，因此 IconVQA 为研究抽象视觉信息提供了良好的试验场。

3.3.5 VizWiz

VizWiz 是一个真实世界问题下的 VQA 数据集，收集自视觉受损者的日常生活中。数据集中的每一张图片都配有视障者提出的问题，涉及物体识别、文字阅读和日常任务等方面。VizWiz 的特别之处在于其问题来源和图像质量，这些图像可能模糊、不完整或光照不良，这些挑战让 VQA 模型对现实世界有更好的适应性。

3.3.6 HM (HowMany-QA)

HM 数据集专注于计数任务，它含有大量的视觉内容和对应的计数问题。这个数据集的目的是解决 VQA 中的一个特定子问题——对象计数。计数任务在现实应用中相当重要，如人群分析、库存管理等。HM 数据集要求模型具有较强图像分割和识别的能力，从而准确计算图像中的特定类别物体数量。

4 复现细节

4.1 与已有开源代码对比

改进点 1 在模型架构中。采用了更有效的策略，将视觉 Vit 特征投影到线性层，该策略是通过使用 4 合 1 策略来减少了视觉输入标记的数量，大大提高了处理效率。这种策略是通过把视觉特征信息和语言模型空间进行融合，使两者在同一空间经行操作，从而提高模型的理

解和预测能力；改进点 2 在模型的多模态理解能力上有所提升。模型引入了语言表达的相对坐标边界框，以增强模型在视觉问题回答（VQA）任务的表现。这种优化是通过引入一个新的任务类型——“对象解析与定位”，以进一步提升模型对图像中多物体的识别和描述能力。

4.2 语言边界框视觉定位

本模型在进入实验阶段时，根据不同的数据集和任务类型进行了细致的调整和优化。为了提高模型对图像中多物体的识别和描述能力，特别利用了 Flickr30k 数据集对其进行了针对性的微调。此外，模型引入了一个新的任务类型——“对象解析与定位”，以进一步增强模型的多模态理解能力。在多轮对话环境中，针对模型可能出现的多任务处理挑战，创造了混合多任务对话数据集，并将其纳入模型训练。此举旨在在多轮复杂上下文中，提升模型的任务处理性能。

4.3 训练过程

本模型训练时采用了带有任务标识符的多任务指令模板。在输入指令中嵌入具体任务的标识符，让模型在训练过程中增强对多任务理解的倾向。通过三个阶段对模型进行训练，以期实现更佳的视觉对齐效果：

4.3.1 阶段一

预训练。使用了包括弱标注和精细标注数据集在内的大量视觉-语言数据，以确保模型具备较为广泛的视觉-语言知识。弱标注数据集赋予了较高的采样比率，目的是让模型获取更丰富的知识。

4.3.2 阶段二

：多任务训练。本阶段重点关注的是提升模型在各视觉任务上的表现，因此仅使用精细数据集进行训练，并且根据任务的频率调整数据采样比率，确保模型优先学习高质量的图像-文本对齐数据。

4.3.3 阶段三

多模态指令微调。接下来，这一阶段关注于使用更多的多模态指令数据集对模型进行微调，并增强其作为聊天机器人的对话能力。继续使用第二阶段的数据集，并增加新的指令数据集，包括 LLaVA、Flickr30k 以及之前构建的混合多任务数据集和语言数据集 Unnatural Instruction。对此赋予第二阶段细粒度数据集更低的采样比率，同时赋予新指令数据集更高的采样比率。

5 实验结果分析

5.1 VQA 指标评估

本节将展示在不同数据集中，新提出的方法与业界领先的 VQA 模型 LLaVA 之间的性能对比分析。这一对比涵盖了 OKVQA、GQA、VSR、IconVQA、VizWiz 和 HM 六个不同的视觉问题回答数据集。通过评估结果，可以清楚地看出在多个数据集中新模型实现了显著的性能提升。在 OKVQA 数据集中，模型的准确率从 54.4% 提高到了 56.9%，增加了 2.5 个百分点。在 GQA 数据集上的提升则更为显著，从 41.3% 跃升至 60.3%，增幅近 19 个百分点。对于 VSR 数据集，准确率也有所增长，从 51.2% 提升至 60.6%，增加了 9.4 个百分点。在 IconVQA 数据集上，准确率从 43.0% 增加至 47.7%，提升了 4.7 个百分点。在 VizWiz 数据集上，新模型得到了 32.9% 的准确率，而与 LLaVA 模型无法比较，因为 LLaVA 没有在该数据集上报告结果。同样，在 HM 数据集上的表现也优于 LLaVA，达到了 58.2% 的准确率。这些定量指标有效证实了新模型在视觉问题回答任务上的出色表现。尤其在 GQA 数据集上，新模型的性能大幅度超越 LLaVA 模型。根据提供的材料，GQA 数据集的 top-1 准确率从 41.3% 提升至 60.3%，这表明新模型在理解视觉问题、整合信息、抽象概念理解以及生成准确答案等方面所作的改进是成功的。

表 1. VQA 各数据集评估结果

Method	OK-VQA	GQA	VSR	IconVQA	VizWiz	HM
LLaVA	54.4	41.3	51.2	43.0	-	-
VQABBoxpro	56.9	60.3	60.6	47.7	32.9	58.2

5.2 实验分析

本节将通过具体案例来分析新模型的性能。在处理需要多步推理和对图像细节深入理解的复杂问题时，新模型展示了其明显的优势。例如，当问题涉及到识别特定物体属性时，新模型能够精准地提取关键信息，给出更准确的答案。在一些需要深度图像解析和理解的抽象问题上，新模型的表现也优于 LLaVA 模型。举例而言，当被询问“图片中的人物从事什么职业”时，新模型能够结合服装特征和背景信息进行有效判断，而 LLaVA 模型在这类问题的回答中往往表现欠缺。

6 总结与展望

综合以上分析，新模型在不同的视觉问答任务中均展现出优越的性能。通过性能评估和案例分析得出的结论，均突显了新模型的效能和其在实际应用中的潜力。展望未来，将在更广泛的数据集中验证模型性能，并继续进行优化以适用于更多样化的视觉问答环境和挑战。

参考文献

[1] Y Chang, X Wang, J Wang, et al. A survey on evaluation of large language models. *ArXiv preprint*, abs/2307.03109, 2023.

[2] K Chen, Z Zhang, W Zeng, et al. Shikra: Unleashing multimodal llm's referential dialogue magic. *ArXiv preprint*, abs/2306.15195, 2023.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[4] Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China, 2019. Association for Computational Linguistics.

[5] H Liu, C Li, Q Wu, et al. Visual instruction tuning. *ArXiv preprint*, abs/2304.08485, 2023.

[6] H Touvron, T Lavril, G Izacard, et al. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971, 2023.

[7] W Wang, Z Chen, X Chen, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *ArXiv preprint*, abs/2305.11175, 2023.

[8] Y Zhou, A I Muresanu, Z Han, et al. Large language models are human-level prompt engineers. *ArXiv preprint*, abs/2211.01910, 2022.

[9] D Zhu, J Chen, X Shen, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv preprint*, abs/2304.10592, 2023.