摘要

近期，多模态大模型（Large Multimodal Model，LMM）已成为人工智能领域的研究焦点，其通过结合强大的大语言模型（Large Language Model，LLM）与视觉编码器（vision encoder）来处理包含多种模态的任务。这些模型展现了诸多新颖能力，例如基于图像编写故事以及在无需光学字符识别（Optical Character Recognition，OCR）技术支持下进行数学推理，这不仅在传统方法中难以实现，也指明了向通用人工智能（Artificial General Intelligence, AGI）发展的潜在道路。

尽管当前对多模态大模型，尤其是其组成部分——视觉编码器与大语言模型之间的交互方式已有深入研究，但关于视觉编码器本身的探讨却相对较少。一些研究表明，视觉编码器的性能局限可能成为制约多模态大模型整体性能的瓶颈，并尝试通过融合不同视觉编码器来克服这一限制。然而，现有研究主要聚焦于视觉编码器的组合策略及其数值分析，缺乏对融合后视觉编码器如何共同作用以提升模型性能的深入分析。

鉴于此，本文旨在探究视觉编码器融合与优化在多模态大模型中的应用，以及这些方法如何促进模型性能的提升。具体研究内容包括：

1. 选取不同的视觉编码器，在保持多模态大模型整体架构不变的前提下，对其在相同数据集和评测标准下的表现进行训练和评估。通过分析不同评测指标的变化，总结出视觉编码器对多模态大模型在不同领域能力影响的规律，并找到一套适合评估视觉编码器对多模态大模型影响的评测指标。
2. 利用基于注意力矩阵和梯度的可视化技术，深入理解融合不同视觉编码器如何增强多模态大模型的能力，进而提出一种基于特征拼接的视觉编码器融合方法。该方法在降低计算成本的同时，显著提升了模型的认知与感知能力。为提高方法的普适性，进一步设计了适用于任意视觉编码器组合的基于互注意力的视觉拼接模块，并通过模型和数据的优化训练，训练出了超过基线性能的多模态大模型，同时在更大规模的大语言模型上验证了该方法的扩展性。

综上所述，本文在深入探讨了多模态大模型中视觉编码器的作用及其融合的生效机制的基础上，提出了一种在比现有方法计算成本更低的情况下依旧维持高性能的视觉编码器融合方案，并设计了基于互注意力的特征拼接模块，实现了其在任意视觉编码器组合中的适用性。通过进一步的模型优化，在更大的模型上验证了该方法的有效性，为未来工作选择和优化多模态大模型中的视觉编码器提供了新的视角。

关键词：视觉编码器，多模态，大语言模型，模型融合