



HOKKAIDO UNIVERSITY

Title	A Study on Visual Recognition with Large Multimodal Models and Its Applications
Author(s)	甘, 耀宗
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第16993号
Issue Date	2026-03-25
DOI	https://doi.org/10.14943/doctoral.k16993
Doc URL	https://hdl.handle.net/2115/99837
Type	doctoral thesis
File Information	Yaozong_Gan_review.pdf, 審査の要旨 https://creativecommons.org/licenses/by/4.0/



学位論文審査の要旨

博士の専攻分野の名称 博士 (情報科学) 氏名 甘 耀宗

審査担当者 主 査 教 授 小川 貴弘
副 査 教 授 土橋 宜典
副 査 教 授 長谷山 美紀

学位論文題名

A Study on Visual Recognition with Large Multimodal Models and Its Applications

(大規模マルチモーダルモデルを用いた視覚認識およびその応用に関する研究)

本論文は、大規模マルチモーダルモデル (LMM: Large Multimodal Models) を実世界の視覚認識タスクへの適応・拡張可能性を探究した研究成果をまとめたものである。

深層学習は人工知能 (AI) の中核技術となり、さまざまなタスクにおいて著しい進展を牽引してきた。大規模画像認識における畳み込みニューラルネットワーク (CNN) の成功以降、ResNet や EfficientNet に代表される一連の CNN ベースのアーキテクチャが、階層的かつ表現力の高い視覚特徴を学習することにより、コンピュータビジョン分野を大きく発展させた。一方、自然言語処理 (NLP) においては、BERT や GPT-2 などのモデルが、大規模事前学習によりデータから構造化された知識を抽出する有効性を示した。これらの発展の融合により、AI は単一モダリティからの認識を越えて、視覚・言語など複数モダリティを統合的に活用するマルチモーダル理解へと進化し、より包括的な推論および意思決定が可能になった。

これらを背景として登場した大規模マルチモーダルモデル (LMM) は、AI 研究における重要なパラダイム転換を示している。Qwen、Gemini、GPT などの LMM は、異種モダリティを統合した大規模データセットで学習され、テキストと画像を含む入力間で共有された意味表現を習得する。これらのモデルは、視覚と言語の整合、クロスモーダル検索、マルチモーダル推論、生成タスクにおいて顕著な性能を示している。

一方、LMM を実環境の視覚認識タスクに適用するにあたり、以下に示すいくつかの未解決課題が残存している。第一に、既存の LMM は大量のペアデータおよび膨大な計算資源に依存しており、データが限られた領域や特定ドメインへの展開に制約がある。第二に、異なるモダリティ間の整合は必ずしも完全ではなく、ドメインバイアス、曖昧な推論経路、複雑な実世界画像に対する解釈性の低下などを招く。第三に、事前学習済み LMM は高い視覚認識能力を有するものの、分布の変化や未知カテゴリに直面した際のクロスドメイン一般化能力および細粒度識別には限界がある。このように、LMM は視覚認識分野を大きく前進させたものの、データ依存性や実世界の視覚認識シーンにおけるドメイン特化性といった重要課題が依然として残存している。これらを解決することは、知的交通システムやリモートセンシングを含む複雑な視覚環境において信頼性の高いマルチモーダル認識システムを構築するために不可欠である。

本論文の目的は、LMM を実世界の視覚認識タスクへどのように効果的に適応・拡張できるかを探究することである。特に、交通標識認識やリモートセンシングといった応用に注目し、クロスドメイン環境における汎化性、解釈性、データ効率の改善を重視する。本論文では前述の課題に対処

するため、段階的に発展する一連のマルチモーダル視覚理解手法を提案し、各手法はこれらの問題の一部を緩和するよう設計されており、総合的に LMM の適応力と推論能力を高めることを目指している。

本論文の構成は以下のとおりである。第 1 章では研究背景と目的を述べる。第 2 章では深層学習、LMM、および交通標識認識・リモートセンシングにおける関連研究をレビューする。第 3 章ではゼロショット条件下の視覚認識を扱い、カテゴリ間の構造関係を分析する。第 4 章では少数ショット学習への拡張を扱う。第 5 章ではマルチモーダル推論による細粒度・クロスドメイン視覚認識を検討する。第 6 章ではリモートセンシング領域における大規模マルチモーダル応用への拡張を示す。第 7 章では本論文の結論と今後の展望を述べる。

以上を要約すると、本論文は、LMM を実世界の視覚認識タスクへの適応・拡張可能性を探究した研究成果を創出したものであり、多様なタスクを対象とした広範な評価により、提案技術は既存手法を大きく上回る性能改善を示している。この貢献は、情報科学分野の発展に寄与するものと認められる。したがって、本論文における著者は、北海道大学博士 (情報科学) の学位を授与される資格を有するものと認める。