TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

利用課題名　GeoVLM：マルチビュー協調型視覚言語モデルによる包括的な都市理解に向けて
英文：GeoVLM: Towards comprehensive urban understanding via multi-view coordinated vision-language model

利用課題責任者　横矢直人
First name Surname　Naoto Yokoya

所属　東京大学大学院新領域創成科学研究科
Affiliation　Graduate School of Frontier Sciences, The University of Tokyo

邦文抄録（300 字程度）

地球観測システムの急速な発展により、リモートセンシング画像、地点情報、地域気候区分、土壌タイプなど、膨大な地理空間データが日々取得されている。これらの異種データから有用なセマンティック情報を効果的に抽出・統合するためには、国連が掲げる持続可能な開発目標およびデジタルシティの理念に沿った、環境モニタリングと都市管理を協調的に行うための大規模なマルチモーダルモデルの開発が求められている。本研究では、日本全国を対象地域とし、スーパーコンピューティング資源を活用して、サブメートル精度の土地被覆マッピング、都市計画支援、農業管理の自動化を実現する大規模マルチモーダルモデルを開発する。

英文抄録（100 words 程度）

With the rapid development of earth observation systems, massive amounts of geographic data (remote sensing imagery, points of interest, local climate zones, soil types, etc.) are being acquired daily. To effectively extract and comprise useful semantic knowledge from the heterogeneous data, we need to develop a large multi-modal model for collaborative environmental monitoring and urban management, meeting the United Nations' sustainable development and digital city goals. Regarding the whole of Japan as the starting point, we hereby apply for super-computing resources to develop a large multi-modal model, automatically achieving sub-meter land-cover mapping, city planning guidance, and agriculture management.

Keywords: sustainable urban development, vision-language model, remote sensing

背景と目的

都市計画者は、ENVI、ArcGIS、SPSS といった地理空間および統計分析ツールに多くを依存し、マルチモーダルな地球科学データを分析して報告書や都市計画を作成している。しかしビッグデータ時代において、これらの作業には高度な専門知識が求められる上、膨大な時間も要するようになり、ますます負担が増している。こうした状況を背景に、膨大な地球科学データを効率的に処理し、有意義な情報を抽出して実用的な地球科学知識へと導く、インテリジェントな地球科学システムの整備が急務となっている。とはいえ、このような多次元かつ大規模な地質データのインテリジェントな処理・解析には、いくつかの重要な課題が存在する。

①リモートセンシング画像、ストリートビュー、テキスト情報といった異なる種類のデータから、どのように統一された特徴表現を得ることができるのか？

②地球科学的な問いに対する回答や、都市計画における意思決定を、どのようにして推論できるのか？

本プロジェクトでは、問題①に対して、マルチモーダルな地球科学データを統一的に表現し、各種機能を一貫したセマンティクスに変換するためのコントラスト学習に基づく自己教師あり学習戦略を設計した。問題②に対しては、大規模言語モデルの高い推論能力を活用し、オブジェクト間の空間的関係や意味的分析を可能にするマルチモーダル視覚言語モデルを構築した。これにより、都市の持続可能な発展に関するレポートを自動生成する仕組みを実現している。

概要

本研究では、リモートセンシング（RS）画像と街路景観（SV）画像に由来するマクロおよびミクロレベルのセマンティクスを統合することで、都市の持続可能性に関する課題に対処するビジョン言語モデル（VLM）の機能拡張を目指した。まず、20,589 枚の RS 画像、110 万枚の SV 画像、80 万件の質問応答ペアから構成される国家規模の地理空間ビジョン言語データセット GeoVLSet を構築した。このデータセットは、地理空間オブジェクトの推論、社会的要素の分析、都市経済の評価、持続可能な開発に関するレポートの自動生成を促進するものである。さらに、RS 画像と SV 画像が持つ補完的なセマンティクスを効果的に統合し、多層的な都市理解を実現するマルチイメージ対応の視覚言語モデル GeoVLM を開発した。提案手法により、エンドツーエンドの指示調整を通じて、直接的かつ堅牢な都市分析が可能であることを示した。本研究の成果は、リモートセンシングにおける VLM 研究に新たな視点を提供するとともに、将来の都市分析システムの高度化・効率化に貢献するものである。

結果および考察

本研究で提案する視覚モデルは、日本全国の複数の都市において、既存のあらゆる手法を上回る品質の評価結果を示した。すべてのシーンは、商業、工業、住宅、未開発または半開発の 4 つのクラスに分類している。既存の基本情報に基づき、すべてのリモートセンシング画像に対して適切な都市レポートを作成できるよう、アノテーターには十分なトレーニングを行った。図 1 には、異なるシーンの特徴を反映した 3 都市の持続可能な開発レポートの例を示している。

まとめ、今後の課題

都市分析の要件に基づき、20,589 枚のリモートセンシング（RS）画像、110 万枚のストリートビュー（SV）画像、80 万件の質問応答ペアを含む、全国規模の地理空間視覚言語データセット GeoVLSet を提案した。本データセットは、地理空間オブジェクトの推論、社会的キ―オブジェクトの分析、都市経済の評価、持続可能な開発レポートの生成といったマルチレベルのタスクに対応し、大規模視覚言語モデル（VLM）の高次推論能力を拡張するものである。

図 1. 多様なリモートセンシングシーンにおける都市持続可能性レポートの代表例

RS 画像と SV 画像の統合により、補完的な意味情報の取得が可能となる一方で、特に RS-SV 間の関係性や長文コンテキストのモデリングにおいては、依然として重要な課題が存在する。これらの課題に対処するため、本研究では、地理空間的な位置埋め込みと長文コンテキスト処理に対応した大規模言語モデル（LLM）を組み込んだ、堅牢な視覚言語ベースライン GeoVLM を提案した。

広範な実験を通じて、GeoVLM の優れた性能と高い堅牢性が確認された。GeoVLSet および GeoVLM は、Earth Vision 分野におけるドメイン特化型 VLM の開発に資するベンチマークとなることが期待される。今後は、興味地点（POI）などのソーシャルメディアデータの統合や、都市災害といったさらなる応用シナリオの拡充を予定している。