위성영상 기반 Remote Sensing을 위한 Vision-Language 데이터셋 활용 방안

문경렬, 금형철, 백승호

LIG넥스원

kyeongryeol.moon@lignex1.com, hyungcheol.geum@lignex1.com, seungho.baek@lignex1.com

Applications of Vision-Language Datasets for Remote Sensing Based on Satellite Images

Moon Kyeongryeol, Geum Hyeongcheol, Baek Seungho

LIG Nex1

요 약

최근 인공지능 기술은 이종 데이터간 연관성을 기반으로 입력 데이터의 복잡한 추론을 시도하고 있다. 특히 Vision-Language Models(이하 VLMs)의 급격한 발전은 주어진 영상 속 오브젝트의 관계 또는 장면 상황을 이해하는 방법론으로 주목받고 있다. 이러한 기법으로 위성영상을 분석하여 상황 인지를 시도하는 도전적인 문제에도 많은 적용이 이루어지고 있으나, 위성영상 - 자연어 쌍으로 이루어진 데이터셋은 그 특성상 쉽게 수집하기 어렵다. 따라서 본 논문에서는 위성영상을 활용한 상황인지 연구에 활용할 수 있는 다양한 Remote Sensing을 위한 Vision-Language 데이터셋을 소개하며 미래 국방 분야에서 이들의 활용 방안을 소개한다.

Ⅰ. 서 론

최근 AI 기술은 대형 모델을 중심으로 급속도로 발전하고 있다. 그 중에서도 Vision-Language Models(이하 VLMs)의 발전은 영상, 자연어 간 연관성에 기반한 복잡한 추론이 가능해졌다. 구체적으로 GPT-4V, Gemini, Ferret 등 오픈소스로 공개되거나 상용화된 모델들이 있다.

하지만 Remote Sensing(이하 RS)분야는 현재 널리 사용하는 Vision-Language 데이터셋을 그대로 적용하기 어려운 특수한 영역이다. 또한 도메인 용어가 많이 존재하여 모델이 이를 고려한 적절한 추론을 하기 쉽지 않다. 특히 미래 국방 분야에서 JADC2와 같이 전 영역에 기반한 지휘 통제 체계에서 요구하는 수준은 위성 자산을 활용한 감시 정찰 정보도 포함한다. 이에 기반한 상황 분석은 pre-trained 모델을 기반으로 하여 의미론적으로 특징을 공유하는 인공지능 기술이 필요하다.

무엇보다 현재까지 Remonte Sensing 문제는 Vision-oreiented된 태스크에 한정되어 있다. 이는 위성 영상의 객체 관계 및 장면을 이해하는 상황 인지에 적절한 구조가 아니다. 따라서 위성 영상과 자연어로 구축된 데이터셋으로 RS-specific한 대형 모델 구축을 고려할 수 있다. 최근 VLMs가 발전하면서, 이러한 방향에 맞추어 RS-specific한 Vision-Language 데이터셋의 수가 증가하는 추세이다.

본 논문에서는 기존에 공개된 방대한 데이터셋으로 Vision-Language 표현을 잘 학습한 모델에 기반하여 군 도메인에 특화된 위성 영상-자연어 데이터셋을 파인튜닝하는 방법과 최근 발전하는 RS 문제의 Vision-Language 데이터셋 동향을 소개하며 그 특징에 따른 활용 방안을 제시한다.

Ⅱ. 본론

Sydney-Captions(Image Captioning)[1] 데이터셋은 Sydney 데이터셋을 기반으로 하여 Google Earth 데이터로 구축되었다. Sydney 데이터셋은 호주 시드니의 18000 x 14000 크기와 0.5m 해상도의 픽셀 이미지를 주거지, 공항, 초원, 강, 바다, 산업단지, 활주로 등 총 7개의 클래스로 분류하였으며 총 613장의 이미지로 구성된다. 해당 데이터셋은 Convolutional Neural Network(CNN)을 통해 이미지의 특징을 추출한 후 Recurrent Neural Network(RNN)또는 Long Short-Term Memory(LSTM)으로 매칭되는 이미지 캡션을 결합하여 고해상도 이미지-캡션 데이터셋을 구성한 것 특징이다.

상세한 구성은 각 이미지에 대하여 5개 캡션이 포함되고, 총 613장의 이미지와 3,065개의 캡션이 쌍을 이룬다. Sydney 데이터셋은 80%를 학습, 10%를 검증, 나머지 10%를 테스트셋으로 구분하고, CNN 모델 4가지(AlexNet, VGG-16. VGG-19, GoogLeNet)을 통해 인식한 이미지를 RNN과 LSTM 모델을 각각 조합하여 평가 지표를 산출하였다. BLEU, METEOR, CIDEr 지표를 계산하였으며 VGG-19와 LSTM의 조합을 통한 지표가 가장 높게 계산되었다.(BLEU 1=54.8, METEOR=20.8, CIDEr=37.9) Sydney-Captions에 존재하는 7개의 클래스는 국방 도메인에서 고려할 수 있는 클래스이다. 따라서 이러한 데이터로 pre-trained된 모델의 표현 가용성이 높을 것으로 보인다.

RSICD[2]는 Remonte Sensing 이미지에 대하여 정확한 크기, 수량 등의 표현으로 명확한 정보를 사용하고, 촬영 특성을 고려하여 방향 용어를 제외하여 상대적인 위치 관계를 나타낸 캡션을 포함하는 데이터셋이다. Google Earth, Baidu Map, MapABC, Tianditu 등에서 수집된 10,921장 이상의 다양한 해상도의 원격탐사 데이터를 수집하였고, 224 x 224 픽셀 크기의 이미지로 구성된다. 캡션은 각 이미지 당 원격탐사 분야 전문가에 의해 생성된 5개의 설명이 포함되어 총 24,333개의 문장으로 구성되며 단어 수는 332,333개이다. 30개의 클래스마다 각각 수백 개의 이미지가 포함되어 있으며 동일한 이미지에 5개의 문장이 없는 경우 기존 문장을 무작위로 복제하여 총 54,605개의 문장으로 구성되었다. 데이터셋의 일반성 평가를 위해 UCM, Sydney, RSICD에서 사용한 모델에 학습하여 METEOR: 0.20459, CIDEr: 1.18011의 객관적 지표값을 얻었다. 해당 데이터셋은 전문가에 의해 생성된 캡션들과 클래스의 다양성을 통해 고수준 추론 성능을 보여줄 수 있다.

SkyScript(Image Captioning)[3]는 남극을 제외한 모든 대륙에 해당하는 범위를 가지며 그 중 북미와 유럽에 Visual Grounding에 스위스(GSD: 0.1m, RGB), 스페인(GSD: 0.1m, RGB), 미국 농업부 제공 이미지(GSD: 0.6~1m, RGB) 등 1m 미만의 고해상도 영상이 집중되어 있어 많은 객체를 포함한 이미지 컬렉션으로 구성된다. 특히 인구 밀집 지역일수록 더 많은 객체를 포함하는 특징이 있으며 OpenStreetMap DB에서 추출한 객체로 캡션을 구성하여 총 520만개의 Vision-Language 데이터로 구성된다. 또한 OpenAI의 ViTL/14 CLIP 모델로 이미지와 캡션 임베딩 간 유사도를 계산하여 상위 랭크의 데이터들로만 구성하여 사용할 수 있다. (20, 30, 50 등) 해당 데이터셋으로 SkyClip 모델을 학습하였고, 학습에 사용되지 않은 벤치마크 데이터셋을 통해 검증한 결과 평균 최고 정확도는 59.93%이며 CLIP-laion-RS, Remote- CLIP 등 비교 모델들보다 2~6%P가량 높은 성능을 보였다. 해당 데이터셋은 다양한 객체와 고해상도 영상의 강점을 가지며 넓은 지역을 포함하는 다양성으로 이에 기반한 모델 표현력이 지리적 도메인에 강건한 특징을 가질 수 있다.

RSVQA[4]는 기존 데이터셋 부족을 해결하기 위해 질문과 답변이 포함된 Remonte Sensing 이미지 데이터셋으로 구성되었다. 미리 정의된 템플릿을 사용하여 자연어 기반 질문을 생성한 뒤, 공개 데이터인 OpenStreet- Map을 통해 답변을 생성하여 구축하였다. 구체적으로 저해상도(LR: Low Resolution) 데이터셋은 네덜란드 상공에서 촬영된 Sentinel-2 이미지를 기반으로 가시광선 대역에서 10m 해상도의 이미지를 제공하며, 256 x 256 크기의 이미지 772개와 770,232개의 질의응답으로 구성된다. 고해상도(HR: High Resolution) 데이터셋은 USGS의 고해상도 정사영상(HRO)에서 15cm 해상도로 미국 북동부 해안의 161개 타일을 512 x 512 크기의 이미지 100,65개로 분할하여 10,066,031개의 질문과 답변으로 구성된다. VQA 모델을 통해 LR 데이터셋과 HR 데이터셋을 각각 3회 학습하여 모델 성능을 평가하였다. LR 데이터셋에서는 전체 정확도 79%, HR 데이터셋 테스트셋에서는 83%의 정확도를 보여주었다. 해당 데이터셋은 해상도 측면의 다양성이 높아, 저렴한 위성 촬영 장비로 얻은 데이터를 활용할 수 있다.

RSVG(Visual Grounding) 데이터셋(RSVGD)[5]은 탐지 데이터셋인 DIOR 에서 샘플링하여 구축한 벤치마크 데이터셋이다. DIOR 데이터셋은 클래스 20개, 객체 인스턴스 192,472개, 이미지 수 23,463장(해상도: 800 x 800픽셀), 클래스 다양성 등이 상당히 큰 대규모 데이터셋이다. 이 데이터셋은 DIOR에서 캡션 오류가 있는 데이터를 제거하고, 카테고리 객체를 최대 5개까지 제한하여 박스 샘플링, 속성 추출, 질의 표현 생성, rapid judgement의 4단계에 걸친 작업자 검증을 거쳐 한단계 업그레이드된 데이터셋이다. 최종적으로 RSVGD는 17,402개의 원격탐사 이미지를 활용하여 38,320개의 원격탐사 이미지-쿼리 쌍으로 구성된다. 평균 텍스트 길이는 7.47 단어이며, 사용한 단어 종류는 100개이다. 전체 데이터셋의 40%를 학습, 10%를 검증, 50%를 테스트로 랜덤 분할하여 트랜스포머 기반의 모듈, Multi-Level Cross-modal Featuring Learning Module(MLCM)을 평가하였다. Visual Encoder는 ResNet-50, Language Encoder는 BERT 모델을 활용하여 SOTA method에서 Pr@(0.5, 0.6, 0.7, 0.8, 0.9)=(76.78, 72.68, 66.74, 56.42, 35.07), meanIoU(68.04), cumIoU(78.41)의 결과를 보여주었다. 이는 TransVG, VLTVG 등 다른 트랜스포머 기반 모델들에 비하여 높은 수치이다. MLCM 모듈의 활용 유무에 따른 비교로는 RSVGD 테스트셋을 MLCM 없이 학습하였을 때 Pr@0.5 = 72.41%의 결과가 나왔으며, MLCM을 사용했을 때 Pr@0.5 = 76.78%의 결과로 4.37%P의 성능 향상을 보였다. RSVGD를 MLCM과 함께 활용한다면 트랜스포머 기반 모델의 장점에 따라 장거리 의존성 학습에 효과적이며 병렬처리가 가능하여 학습 및 추론 속도를 높여 시간적, 성능적 측면에서 큰 강점을 보일 것으로 생각된다.

위에서 나온 데이터셋들의 주요 특징을 아래 표에서 확인할 수 있다.

	클래스 수	해상도	이미지 수	캡션 수	검증 결과
Sydney-Captions[1]	7	0.5m	613	3,065 (각5)	BLEU 1=54.8, METEOR=20.8, CIDEr=37.9
RSICD[2]	30	224 x 224	10,921	54,605	METEOR: 0.20459, CIDEr: 1.18011
SkyScript[3]	-	1m 미만	520만	520만	정확도:59.93%
RSVQA-LR[4]	-	10m	772	770,232	정확도:79%
RSVQA-HR[4]	-	15cm	100,659	10,066,031	정확도:83%
RSVGD[5]	20	800 x 800	17,402	38,320	Pr@0.5=76.78%

표 1 Vision-Language 데이터셋 개요

Ⅲ. 결론

본 논문은 위성 영상 기반 RS 기술에서 상황 인지와 객체 간 관계 추론 능력을 위한 Visual-Language 데이터셋 동향을 조사하며 이들의 활용 방안을 제시하였다. 최근의 많은 연구는 위성 영상의 시각적 정보뿐만 아니라 자연어 등 여러 데이터의 의미론적 관계를 이해하기 위해, 위성 영상 도메인에서의 Visual-Language 데이터를 구축하고, 사전 학습한 VLMs의 가능성을 보여주었다.

그러나 RS 분야의 특성상, 수집할 수 있는 데이터셋 규모는 일반적인 Visual-Language 데이터셋 규모에 비해 현저히 작다. 이는 앞으로 고수준, 대량의 데이터셋 확보가 RS 기술을 활용한 상황 인지 성능을 결정하는 중요한 과제임을 암시한다.

또한 국방 도메인에서의 VLMs 또한 처음부터 학습한 모델이 아닌, 일반적 데이터셋으로 사전 학습한 모델을 호출하여 활용하는 것이 시간, 비용, 유지보수 측면에서 효율성을 담보할 수 있다. 이의 구현은 오픈소스 기반 모델에서부터 시작하여 매우 적은 양의 레이블링 데이터로 학습하는 프로세스가 필요하다. 하지만 소량의 데이터로 추론 성능을 유의미하게 향상하는 방법은 향후 이론적 수준에서 고민하여야 하는 중요한 문제이다.

참 고 문 헌

[1] B. Qu, X. Li, D. Tao, and X. Lu. Deep semantic understanding of high resolution remote sensing image. In 2016 International conference on computer, information and telecommunication systems(Cits)

[2] X. Lu, B. Wang, X. Zheng, and X. Li. Exploring models and data for remote sensing image caption generation, IEEE Transactions on Geoscience and Remote Sensing(TGRS)

[3] Zhecheng Wang et al: SkyScript: A Large and Semantically Diverse Vision-Language Dataset for Remote Sensing. AAAI 2024

[4] S. Lobry, D. Marcos, J. Murray, and D. Tuia. RSVQA: Visual question answering for remote sensing data. IEEE Transactions on Geoscience and Remote Sensing

[5] Y. Zhan, Z. Xiong, and Y. Yuan. RSVG: Exploring data and models for visual grounding on remote sensing data. IEEE Transactions on Geoscience and Remote Sensing