SAI4EO: Symbiotic Artificial Intelligence for Earth Observation

Nicolò Taggio[1,*,†], Sergio Samarelli[1,†] and Matteo Simone[1,†]

[1]Planetek Italia, via Massaua, Bari, 70123, Italy

Abstract

Symbiotic Artificial Intelligence (SAI) refers to the symbiotic relationship between artificial intelligence systems and human users, characterized by mutual interaction and cooperation. Earth Observation (EO) involves collecting data about the Earth's surface and atmosphere using technologies like satellites and aerial sensors. Leveraging the synergies between these domains, this work explores the convergence of SAI and EO through the implementation of an EO assistant-chatbot system. In particular, using natural language and remote sensing data, two primary tasks will be investigated: image captioning for object counting and detection, and scene description for image classification. This integration promises to revolutionize automated analysis and interpretation of EO data, with significant implications in the evolution of smart cities, for environmental monitoring, land use planning, and related fields.

Keywords

Symbiotic AI, Earth Observation, Artificial Intelligence, Smart Cities

1. Introduction

Symbiotic Artificial Intelligence (SAI) explores the complex aspects of the relationship between humans and artificial intelligence, including scientific, societal, economic, legal, and ethical considerations. With the pervasive integration of AI systems into our daily routines, the imperative to address existing limitations and constraints in human-machine cooperation has gained paramount importance. While the effectiveness and precision of an autonomously operating AI agent remain central concerns, the landscape becomes more intricate within a collaborative framework where humans and intelligent AI systems collaborate towards shared objectives. The AI system must possess the capacity to comprehend not only human actions, but also their cognitive frameworks. Symbiotic AI holds the potential to revolutionize human-machine interaction, fostering symbiotic partnerships that amplify and enrich human cognitive capabilities rather than removing them.

Earth Observation (EO) encompasses a broad spectrum of technologies and methodologies aimed to collect comprehensive insights into the Earth's surface and atmosphere. Through the utilization of remote sensing technologies, including satellites, aerial sensors, and ground-based instruments, EO attemps to capture and interpret various aspects of the Earth environment. These observations span a diverse array of phenomena, ranging from natural processes such as weather patterns, land cover changes, and geological formations, to human-induced activities like urbanization, deforestation, and agricultural practices. EO plays a pivotal role in facilitating our understanding of global dynamics, enabling scientists, policymakers, and stakeholders to monitor environmental changes, assess natural hazards, and make informed decisions regarding resource management, disaster mitigation, and sustainable development efforts. Moreover, EO data serves as a valuable resource for a multitude of applications across disciplines, including climate research, biodiversity conservation, and urban planning.

Anyway, the intricacy of satellite-acquired data, coupled with the vast volume encompassing various types such as optical, SAR (Synthetic Aperture Radar), multispectral, hyperspectral sensors, among others, presents a significant challenge in swiftly translating this wealth of information into actionable insights for end-users. Given these considerations, AI, particularly SAI, emerges as an important link to bridge human requirements with the wealth of information derived from remote sensing technologies. For this purpose, we highlight the tasks of "Image Captioning" and "Scene Description" applied to the EO domain. Both tasks focus on describing the contents of the images, the former by detecting the objects that are present while the latter by assigning a class to them from a pre-defined set.

In this project, starting from a comprehensive review of existing literature (Section 2), an exploration of two distinct tasks will be investigated (Section 3), describing how SAI can facilitate the development of an EO assistant chatbot tailored for EO applications (Section 4). Lastly, the discussion will delve into the challenges encountered and outline future avenues of exploration (Section 5).

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy

*Corresponding author.

†These authors contributed equally.

taggio@planetek.it (N. Taggio); samarelli@planetek.it (S. Samarelli); simone@planetek.it (M. Simone)

0009-0003-6392-3099 (N. Taggio); 0009-0004-6822-231X (M. Simone)

We highlight the possibility of leveraging Multimodal Large Language Models (MLLMs) to solve EO tasks and underline the lack of state-of-the-art approaches to tackle the domain challenges.

2. State of the Art

In the following, a state of the art review is detailed, exploring latest developments, methodologies, and applications of algorithms and datasets in Symbiotic AI for EO.

2.1. SAI Methods

Although recent advances in deep learning applied to the EO field have demonstrated promising results in visual analysis tasks, e.g. object detection and instance segmentation, current methodologies tipically rely on task-specific architectures. These approaches, while effective for individual tasks, often struggle to handle the complexities of multi-sensor remote sensing data, accommodate multiple tasks, and generalize to open-set reasoning scenarios.

In contrast, the emergence of MLLMs has gathered attention in the domain of natural images, showcasing impressive multi-task reasoning abilities in real-world settings. Unlike domain-specific models tailored for specific tasks, MLLMs exhibit versatile performances and can generalize effectively to new situations, enabling zero-shot capabilities across open-set tasks.

Key architectures employed in trending MLLMs include the Vision Transformer (ViT)[1], Large Language Model (LLM) and Contrastive Language-Image Pre-training (CLIP) [2]. ViT stands out for its departure from traditional convolutional neural network (CNN) approaches, relying solely on self-attention mechanisms to process images as sequences of tokens. This innovative methodology enables ViT to achieve remarkable performances across various vision tasks, highlighting its adaptability and versatility. On the other hand, CLIP, developed by OpenAI, leverages contrastive learning principles to jointly understand images and text, representing them in a shared embedding space. Nowadays, Large Language Models are cutting-edge linguistic tools crafted to empower computers with the capability to understand human language. The LLaMA (Large Language Model Meta AI) [3] family stands out for its remarkable ability to catch intricate contextual connections, marking a significant leap forward in natural language processing. These foundational models enhance the transformer architecture's capacity for comprehending natural language, owing to their extensive set of trainable parameters. Some open source models, like LLaMAntino [4], have been released for Italian language using Language Adaptation strategies.

Through the integration of these recent architectures, symbiotic AI systems can process and synthesize information from heterogeneous sources such as text, images, and other sensory data. This capability allows systems to perceive and interpret the surrounding world in a more comprehensive and multidimensional manner. Furthermore, leveraging MLLMs in symbiotic AI provides opportunities for natural interaction between humans and machines. The models can understand human requests in a more contextual manner, interpreting not only the text but also associated images, enabling a more seamless and intuitive interaction.

However, applying MLLMs to EO data poses significant challenges due to the substantial differences between natural and remote-sensing images, including variations in imaging conditions, environmental scales, and acquisition angles. Consequently, there is a scarcity of studies in literature focused on the application of MLLMs to EO data, highlighting an important area for future research and innovation. Among these research works, it's noteworthy to highlight GeoChat [5] and EarthGPT [6], despite their significant constraints and lack of suitability for the Italian language.

2.2. EO Datasets

EO datasets serve as the base of intelligent systems capable of extracting valuable insights from images. Within this domain, extensive research has been conducted, considering the utility of these datasets for both scientific and commercial purposes. Indeed, the literature presents numerous datasets crucial for significant EO tasks such as object detection and semantic segmentation. For instance, the DOTA dataset [7] is curated explicitly for object detection, featuring satellite images prepared for oriented bounding box detection. Additionally, xView [8] stands out as one of the largest datasets available, boasting one million object instances across 60 diverse classes. However, it is important to highlight that these datasets are often encumbered by licenses limiting their usage to non-commercial applications. Despite these restrictions, the demand for freely accessible datasets remains imperative, particularly from a service-oriented perspective.

In this context, datasets like SODA [9] and RarePlanes [10] emerge as viable options for Task 1, which involves image captioning and object detection. Alternatively, for the Task 2, datasets such as, Coastal Zone EEA and Open Street Map are preferred due to their comprehensive coverage and unrestricted access. Specifically, the SODA dataset serves as a comprehensive benchmark for small object detection. SODA-A, a subset of SODA, comprises 2513 high-resolution images meticulously annotated with oriented rectangles across nine distinct classes, and is

distributed under the MIT license [9]. Furthermore, to recognize the significance of particular object in military applications and to be able to perform analysis, the RarePlanes dataset has been incorporated into the project. This dataset encompasses a diverse examples of both real and synthetic generated satellite imagery, categorized into various classes based on airplane characteristics such as propulsion type, length, civilian or military designation, number of tail fins, and more.

On the other hand, in the task of land cover land use classification, identify classes present in a buffer along coastal zone is a crucial task. In fact, the European Environment Agency (EEA) has destined an entire project for the first Copernicus Land-VHR Coastal Zone hotspot thematic mapping produced on the European coastal zones. A consortium of EO service providers, spearheaded by Planetek Italia, has been commissioned by the EEA to develop a novel product focusing on Coastal Zones (CZ). This initiative aims to enhance the Thematic Hotspot Mapping (THM) category of the Copernicus Land Monitoring Service (CLMS). THM attempts within the CLMS complement the broader wall-to-wall mapping efforts by furnishing specific and comprehensive land cover/land use (LC/LU) data to tackle environmental challenges effectively. The upcoming products will encompass the entire European coastal region up to an inland depth of 10 km, spanning approximately 730,000 km². These products will feature a minimum mapping unit of 0.5 hectares and delineate approximately 71 LC/LU classes. The project delivers a comprehensive and highly precise LC/LU map encompassing the entire European coastline, boasting an impressive accuracy rate exceeding 90%. This serves as an important illustration of how SAI can expedite the extraction of valuable insights from vast amounts of data, leveraging approximately 10 terabytes of remote sensing images within the realm of EO. Finally, for the Task2, other available layers like OSM will be investigated to extract useful information in terms of buildings, streets, industrial zones and more.

3. Tasks description

In this section, two crucial tasks in EO are detailed where SAI, particularly MLLMs, could bridge the gap between EO applications and human interaction, facilitating direct access to insightful information derived from remote sensing data.

3.1. Image Caption

In the context of EO, conducting a thorough analysis of a scene emerges as a valuable effort for EO experts. Various scenarios present themselves to EO experts when examining images representing Very High Resolution (VHR) data. For instance, monitoring the activity within strategic areas such as airports, including the counting of airplanes, helicopters, and similar entities, serves to detect unexpected occurrences. Conversely, understanding the multitude of objects within a scene, ranging from vehicles to bus routes and sports facilities, holds relevance within the context of smart cities. Moreover, having this information represented in terms of bounding boxes, ideally with geographical coordinates or oriented bounding boxes, can significantly expedite standard Geographic Information System (GIS) processes for domain experts. In light of these requirements, the development of an EO chatbot tailored for EO becomes both necessary and highly desirable, enabling optimal interrogation of VHR data.

Nevertheless, certain challenges warrant attention. Firstly, the scarcity of readily available ground truth data in AI-like format is noteworthy. Within the EO domain, it's crucial to emphasize the scarcity of data, particularly for deep learning methodologies. While some ground truth datasets exist, their accessibility is restricted by commercial licensing restrictions. For instance, datasets like DOTA or xView, which encompass diverse objects, cannot be utilized for commercial purposes due to licensing constraints, as previously mentioned. Secondly, obtaining commercial Very High Resolution (VHR) images poses a significant obstacle.

So that, the **Task 1**, called *image caption - object detection* aims to detect and count objects in a remote sensing images using very high resolution data (i.e. from 30 to 50 cm) and with 3 or 4 bands.

3.2. Scene Description

In the same context, the task of analyzing a scene in terms of LC/LU, potentially on a global scale, presents a significant challenge in the EO domain. This challenge stems from various factors, including the complexities involved in classifying numerous types of LC/LU due to their spectral similarities. Additionally, resolving the issue requires addressing the limitations of satellite image resolution, both spatial and spectral. While certain vegetation classes may benefit from spectral information (i.e. tree cover, shrubland, grassland, etc..), discerning others, such as urban or anthropic areas (i.e. dense or sparse urban), necessitates intricate patterns and, consequently, higher spatial resolution.

Another crucial aspect lies in the potential of utilizing open data to enhance the EO chatbot's capability in describing LC/LU. For instance, various freely available layers such as OSM, Microsoft Open Buildings, and similar resources can be leveraged without additional training for computer vision step. The EO chatbot needs to possess the capability to utilize geographical information for data extraction, crop it to the area of interest,

and conduct inquiries to retrieve additional information.

The **Task 2**, called *scene description - image semantic segmentation* aims to describe the LC/LU following the classes defined in the **coastal zone** dataset using auxiliary information to integrate available pre-computed classes (i.e. residential, and non-residential buildings) in a remote sensing images with high resolution images (i.e. from 2 to 4 m) and with 3 or 4 bands.

4. EO assistant

The objective of the project is to use the theory behind LLMs and MLLMs in assisting EO experts in the creation of operational services. As part of this work, it is developing an EO assistant, tailored to address the aforementioned tasks efficiently. To provide a practical perspective, the tasks have been delineated into specific questions that need to be addressed from an operational standpoint. This approach ensures that the EO assistant is equipped to tackle the essential aspects of EO workflows effectively.

Figure 1 shows an example of what are the expected questions and answers in the Task 1. In particular, the

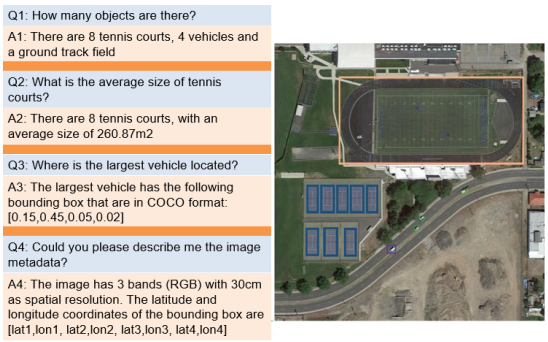

Figure 1: An example of questions and answers between an end-user (blu lines) and the EO chatbot (red lines). On the right side, an example of a VHR image [6]

EO experts need information about the precise spatial positioning of objects within EO imagery. Specifically, they require detailed information regarding the location of objects depicted in the images. This entails not only identifying the objects themselves but also extracting valuable statistical insights related to their dimensions. Additionally, experts often aim to identify the largest object within the scene, as it may hold significant relevance for various analyses and applications. Moreover, the chosen dataset plays a critical role in fulfilling these requirements. It is essential that the dataset provides not only visual representations of objects but also accompanying metadata that offer contextual information about each object's characteristics and spatial attributes. One important aspect is the capability to interpret metadata effectively. EO experts rely on metadata to understand the context of the data and extract meaningful insights. Additionally, the ability to transform bounding box annotations, such as those formatted in COCO (Common Objects in Context), into a geographical context is crucial. This transformation facilitates the integration of EO data with GIS and other spatial analysis tools, enabling a more comprehensive understanding of the landscape and its features. So that, EO experts require not only visual representations of objects in imagery but also detailed statistical information and spatial coordinates. The seamless integration of metadata interpretation and bounding box transformation into a geographical context enhances the usability and relevance of EO data for various applications, ranging from environmental monitoring to urban planning and beyond.

However, the ability to recognize LC/LU on a global scale, or at least in Europe, remains a challenge even with traditional deep learning approaches. Generative AI, specifically SAI, supported by LLMs and MLLMs, can assist EO experts in interpreting remote sensing data and aid end-users in extracting information using natural and straightforward language. Figure 2 illustrates potential use cases that could significantly expedite the replication of the European Environment Agency (EEA) Coastal Zone project for a new product next year. Given that the EEA updates CZ products every six years, there is an expectation for a new product within a relatively shorter timeframe.

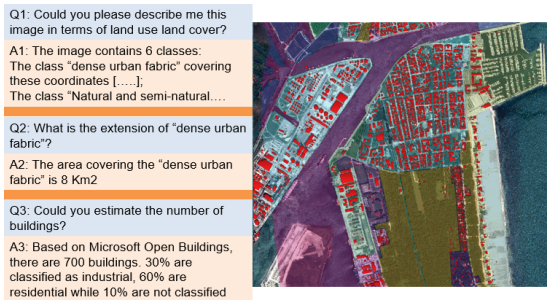

Figure 2: An illustration of a dialogue between an end-user (shown in blue) and the EO chatbot (displayed in red). On the right, there is a depiction of layers beneficial for the EO chatbot. Specifically, the base map is sourced from the Pléiades constellation, while the colored and transparent polygons originate from the OSM (red polygons) and CZ dataset.

It is clear how an EO expert can benefit from an EO chatbot that "comprehends" LC/LU. In fact it enables the monitoring of changes over time, identifying trends like urban expansion, deforestation, or agricultural intensification. It also aids in supporting urban planning by pinpointing suitable areas for development, infrastruc-

ture planning, and land zoning. Additionally, leveraging auxiliary information extracted from sources like OSM could expedite analysis. Importantly, ensuring the information is available in geographical coordinates facilitates direct use by EO experts, not just end-users.

5. Conclusion

In this study, symbiotic artificial intelligence methods have been employed to aid EO experts and end-users in extracting valuable insights related to land use land cover, and image captioning tasks, addressing tangible challenges in the advancement of smart cities, environmental monitoring, land use planning, and related domains. Specifically, an exhaustive review of the current state-of-the-art has been conducted, elucidating the most viable algorithms for two practical scenarios within the EO domain: image captioning and scene description.

The ongoing work has involved the selection of MLLMs based on existing LLMs , such as LLaMantino, alongside diverse EO datasets with open licensing. This amalgamation aims to develop an EO chatbot tailored for both EO experts and end-users. The delineated tasks underscore the necessity for seamless interaction between end-users through natural language and the system's proficiency in retrieving information from EO data, encapsulating the essence of Symbiotic AI for EO (SAI4EO).

Subsequently, the projects will embark on creating ground truth data comprising "*questions-answers-EO data*", followed by a feasibility assessment prior to establishing an end-to-end service catered to both end-users and EO experts.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[3] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[4] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, Llamantino: Llama 2 models for effective text generation in italian language, arXiv preprint arXiv:2312.09993 (2023).

[5] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, F. S. Khan, Geochat: Grounded large vision-language model for remote sensing, 2023. arXiv:2311.15826.

[6] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, X. Mao, Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain, arXiv preprint arXiv:2401.16822 (2024).

[7] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, Dota: A large-scale dataset for object detection in aerial images, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3974–3983.

[8] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, B. McCord, xview: Objects in context in overhead imagery, arXiv preprint arXiv:1802.07856 (2018).

[9] R. Duan, H. Deng, M. Tian, Y. Deng, J. Lin, Soda: A large-scale open site object detection dataset for deep learning in construction, Automation in Construction 142 (2022) 104499.

[10] J. Shermeyer, T. Hossler, A. Van Etten, D. Hogan, R. Lewis, D. Kim, Rareplanes: Synthetic data takes flight, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 207–217.