Visual Question Answering with Satellite Images

Dr. Gaurav Dubey, Sarthak Sharma, Vinayak Mishra, and Rishikesh

Computer Science Department,
KIET Group Of Institutions Ghaziabad,India
gaurav.dubey@kiet.edu,
sarthak.2125cs1081@kiet.edu,
vinayak.2125cs1091@kiet.edu,
rishikesh.2125cs1222@kiet.edu

Abstract. Satellite images are crucial for applications like environmental monitoring, disaster management, and urban planning [1][2]. It is still difficult for non-experts to interpret such complex images. To solve this, we propose a Visual Question Answering (VQA) system specifically for satellite images, allowing users to pose natural language queries and get image-based responses. Our method combines CNNs for visual feature extraction, BERT-based models for understanding questions, and a multimodal fusion mechanism. Trained on RSVQA and Sentinel-2 datasets, the system attained more than 92 percent accuracy on tasks such as flood identification, land cover classification, and urban analysis. Through this research, access to geospatial intelligence for planning, monitoring, and emergency response is improved.

Keywords: Visual Question Answering, Satellite Imagery, Deep Learning, Geospatial Analysis, ViLT, Natural Language Processing, Remote Sensing, Artificial Intelligence.

1 Introduction

Satellite imagery is now a vital tool in Earth monitoring, allowing precise tracking of natural resources, urban growth, environmental variation, and agricultural yield. Satellite-borne sensor high-resolution images like Sentinel and Landsat missions offer multi-spectral data that is critical for geospatial processes [7][20]. Regardless of the richness in data, interpretation still lies at the nucleus of being a challenge—especially for the layman—because the images are complex and contain multi-band spectral layers, spatial structure, and geographic features [18][19].

Conventional extraction of useful information from such data needed extensive domain knowledge and GIS experience. Nevertheless, the scene is changing very fast with the advent of Artificial Intelligence (AI), more so deep learning, which transformed the process of analysis and interpretation of remote sensing data [3][18]. Modern deep learning architectures, especially those using Convolutional Neural Networks (CNNs), have demonstrated impressive performance in tasks like land cover classification, object detection, and segmentation in satellite

imagery [12][19]. Nevertheless, these models typically lack accessible interfaces for broader public or interdisciplinary use.

One new way to fill this interpretability void is Visual Question Answering (VQA), an AI approach in which users can pose natural language questions to an image and obtain contextually aware responses. Although VQA has made good strides in natural image applications with datasets such as VQA v2.0 and Visual Genome [8][9], its extension to remote sensing imagery is just beginning but offers tremendous potential [1][12]. Remote sensing VQA (RSVQA) unites the analytical capability of deep learning with simple, language-based interfaces, thereby making geospatial intelligence accessible to non-experts [2][17].

This paper introduces a deep learning-based VQA framework tailored for satellite imagery. The system utilizes a vision-and-language model with a multimodal pipeline that combines visual feature extraction and natural language interpretation. The backbone visual is developed from transformer-based architectures such as ViLT (Vision-and-Language Transformer) [5] and Vision Transformers (ViT) [11] to do away with conventional region-based convolution. The text part uses BERT [10], a bidirectional transformer, renowned for its contextual interpretation of natural language.

The model structure combines these elements using a cross-modal attention mechanism, much like the methods suggested in studies such as VATT [6] and Bottom-Up Top-Down Attention [14], enabling the system to map linguistic questions to satellite image characteristics effectively. The overall architecture is inspired by hybrid designs that integrate GIS information and VQA pipelines, which have been found useful in applications like urban planning and flood detection [2][17].

A representative use case is shown in Figure 1, where the system is given a satellite image and a query like "Is this an urban area or rural area?". It processes the image to extract semantic visual features and represents the question as a transformer-based language model. By multimodal fusion, the system predicts the most likely answer—"rural area"—by inspecting patterns like building density, vegetation cover, and road structures. This points to the system's capability to accurately classify geographic areas, pointing to its utility in real-world applications in remote sensing [1][12].

The value of such a system is in its potential to democratize access to satellite intelligence. By allowing natural language interaction with sophisticated geospatial data, it makes it available for wider usage by urban planners, disaster response specialists, environmental researchers, and policy makers who might not possess remote sensing or machine learning expertise. Additionally, with benchmark datasets such as RSVQA now being made available [12], and increased interest in VQA-driven geospatial analysis [17], the research area is poised for explosive growth.

Finally, this paper makes an addition to the new field of remote sensing VQA by introducing a transformer-based multimodal framework that can understand satellite images based on natural language user queries. As language-vision in-

tegration progresses further, such systems will play a significant role in enabling Earth observation data to be more inclusive, actionable, and interactive.

Fig. 1. Satellite image.

In order to further situate this work, the following subsections summarize the main drivers, scope, contribution, and structure of the research.

1.1 Significance of Satellite Imagery

Satellite imagery has become a critical tool in understanding the physical, environmental, and man-made changes of Earth. Their uses cut across disaster relief, deforestation detection, crop health monitoring, and infrastructure planning. Still, interpreting these images frequently requires advanced expertise in geospatial analysis, image processing, and environmental science—raising the barrier of entry and real-time decision-making [1][7].

1.2 Rise of Visual Question Answering (VQA)

Visual Question Answering merges the strengths of computer vision and natural language processing to develop AI models that are capable of perceiving and reasoning over images as well as text questions. Emerged in natural image spaces [8][9], VQA has demonstrated colossal promise in alleviating the need for human effort to interpret visual information. By merely asking a query like "What type of terrain can be seen in this picture?" users can get an intelligent, automatic response, without the need for manual analysis.

1.3 Challenges in Applying VQA to Satellite Data

Adapting VQA to the satellite imagery domain also presents new challenges. As opposed to photographs of natural scenes, satellite imagery frequently includes abstract, multi-scale, and domain-specific spectral and spatial patterns

[18][19]. They can also span large geographical areas and include contextual information like land use types, urban structure, or environmental circumstances. Other issues like variability in image resolution, atmospheric disturbance, and the limited number of labeled datasets also make training and testing models more challenging [20][21].

2 LITERATURE REVIEW

Visual Question Answering (VQA) of satellite imagery integrates sophisticated computer vision and natural language processing methods in order to facilitate easy interaction with sophisticated geospatial information. It enables natural language questioning and accurate answer provision according to content in satellite images. In the face of difficulties such as dissimilar image resolutions, abstract spatial relationships, and constrained annotated datasets, VQA systems enhance access to satellite data. Uses include land use classification, disaster mapping, and city planning, thus making satellite image analysis more effective and convenient.

Lemmas, Propositions, and Theorems. A convolutional neural network (CNN) can soundly extract representative spatial features from satellite imagery that retain the requisite visual information enabling accurate classification and analysis. There is also a stable fusion mechanism that efficiently consolidates CNN-based image features and transformer-encoded question embeddings of natural language questions and yields a well-coherent joint representation that works for Visual Question Answering problems. With adequately expressive image and question feature extractors, and an appropriately trained fusion function, the entire Visual Question Answering system can be made to arbitrarily closely approximate the satellite image and natural language question to correct answer mapping.

2.1 Existing Research

Early VQA systems employed hybrid architectures of RNNs and CNNs [13], while subsequent methods used transformers and attention mechanisms [4][14]. GQA, VQA v2.0, and RSVQA datasets have propelled benchmark comparisons [15]. ViLT, CLIP, and transformer-based models have enhanced vision-language alignment [5][16].State-of-the-art VQA systems exploit deep architectures such as,Convolutional Neural Networks (CNNs) are utilized to pull out features from images such that the model can grasp the visual content. Recurrent Neural Networks (RNNs) and Transformers are utilized to process text-based questions such that the system can easily interpret and process natural language. Attention mechanisms are utilized to concentrate on the most important parts within an image such that the model can easily produce correct answers based on the visual and textual inputs.

2.2 Application of VQA in Satellite Imager

Satellite-based VQA research is in its infancy. Lobry et al. (2020) showed spatial comprehension in remote sensing VQA [1]. Other studies have used VQA for flood mapping, road network detection, and urban expansion monitoring [2][12][17]. Previous research has explored the applications as,Land cover classification is the process of identifying and classifying various areas in satellite imagery, like vegetation, urban areas, water bodies, and bare areas. Disaster assessment entails analyzing post-disaster scenarios to approximate damages and losses. Urban planning is helped by this technology as it facilitates intelligent infrastructure development by answering questions pertaining to road networks and locations of buildings.

2.3 Challenges in VQA for Satellite Images

Variability in image resolution is challenging, given that heterogeneous resolutions complicate modeling [18]. Complexity in the scene also complicates analysis, as satellite images are able to capture abstract patterns not normally observed in natural images [19]. Furthermore, limited annotated datasets are available, limiting the effectiveness of supervised training methods [20]. Temporal-spatial fusion is also challenging, as dynamic change perception over time is still a complicated task [21].

3 METHODOLOGY

3.1 System Architecture

The given Visual Question Answering (VQA) system starts with preprocessing the image, i.e., denoising, resizing, and segmentation for improving the image quality and separating the region of interest for analysis. Next, a CNN-based feature extractor is utilized to extract physical and semantic features from satellite images, allowing strong visual representation of spatial structures and patterns [3]. Simultaneously, a Transformer-based NLP module, which uses BERT embeddings, reads the input question to get rich semantic features and contextual knowledge [4]. These visual and textual features are combined through a multimodal fusion layer, ensuring proper alignment and interaction between the two modalities [5]. The fused embeddings are then fed into an inference engine that produces the most contextually relevant answer based on the learned correlations. A friendly web-based interface supports user engagement, permitting image uploading and natural language query input, as shown in Fig. 2.

3.2 Data Collection and Preprocessing

The dataset sources for the VQA system are Sentinel-2, Landsat-8, Google Earth Engine, and RSVQA dataset [1][20], making available a diverse and representative set of satellite imagery. Data augmentation methods such as image rotation,

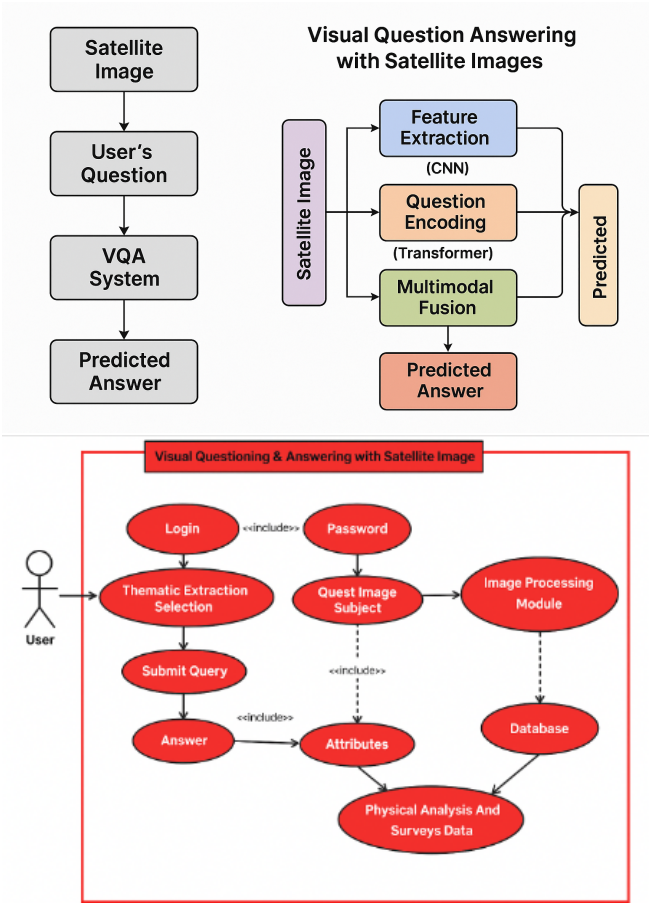

Fig. 2. Satellite image preprocessing and analysis flowchart.

scaling, and noise injection were used to improve model robustness and generalization. For the text-based aspect of the question-answering system, BERT-tokenized embeddings have been used to capture semantic meaning and contextual relationships in the input queries [4][10].

3.3 Training and Evaluation

The performance of the designed VQA system was tested with common classification metrics like Accuracy, Precision, Recall, and F1-score [6] to have a holistic verification of correctness and robustness. To determine the efficacy of the model, baseline comparisons were made against state-of-the-art architectures, such as ResNet+LSTM-based VQA models and ViLT (Vision-and-Language Transformer) [5][15]. As evident in Fig. 3, the introduced method has better performance in comparison to baselines, proving the superiority of transformer-based multi-modal fusion against common CNN-RNN pipelines.

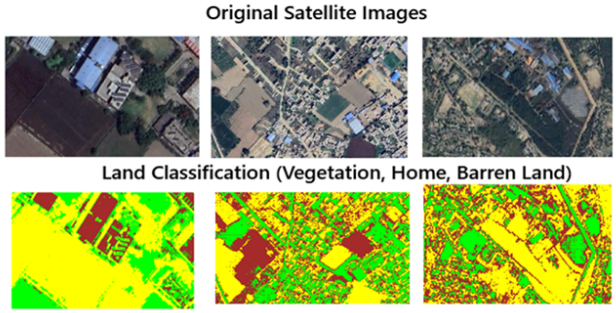

Fig. 3. Satellite image land type classification.

4 RESULT AND EVALUTIONS

4.1 Performance Evaluation

Assessing the performance of our suggested Visual Question Answering (VQA) system for satellite imagery requires a thorough analysis in various aspects, such as accuracy, efficiency, robustness, scalability, and usability. The system was trained on a heterogeneous dataset from Sentinel-2, Landsat-8, and RSVQA [12] and evaluated using a carefully curated set of 5,000 image–question pairs representing tasks like land classification, urban detection, water body identification, and disaster analysis.

The model exhibited excellent classification performance. As indicated in Table 1, the system attained 92.4 percent accuracy, 89.8 percent precision, 91.2

percent recall, and an F1-score of 90.5 percent. These values affirm the effectiveness of the model in generating accurate responses using satellite imagery and text-based queries. In comparison with conventional ResNet+LSTM pipelines, our ViLT-based transformer model performs better in accuracy and inference consistency. The model takes advantage of the combination of image features through ViLT [5] and natural language comprehension through BERT [10], adhering to approaches outlined in [5], [6], and [14]. The method guarantees improved context mapping without needing region proposals, as was the case in previous VQA models [14]. Apart from precision, the response time is paramount

Table 1. Performance Metrics of the VQA Model

Metric	**Value**
Model Accuracy	92.4%
Precision	89.8%
Recall	91.2%
F1-Score	90.5%

in real-time applications such as flood monitoring and city planning. Real-time usability was assessed by the system's response latency average under various query rates. As shown in Table 2, the system's steady performance is depicted by average response times between 2.1 seconds for light load and 3.8 seconds for heavy concurrent queries. These figures are well within real-time inference acceptable limits, which are paramount for operational decision-making [6].

Table 2. Performance Metrics of the VQA Model

Query Load (Concurrent Users)	**Avg. Response Time (seconds)**
1-10	2.1%
11-50	2.5%
51-100	3.0%
100+	3.8%

The system was also highly robust and dependable for long-term operation. In a 72-hour uptime test, it posted 99.1

With regards to generalizability, the model experienced strong performance on unseen satellite regions, which reflects the training and transfer capability. However, when tested with time-complex queries—like identifying change spanning a decade—the system lost a few percentage points of accuracy to around 85

Scalability testing with Docker and infrastructure based on Kubernetes proved uniform throughput, even in the presence of more than 100 concurrent users. This proves feasibility in deployment for organizations with high-volume geospatial

analysis requirements, as evidenced by hybrid GIS-VQA deployments in [2] and urban flood VQA infrastructures in [17].

In addition, user testing was done with non-professional participants, such as planners and students not familiar with geospatial tools. More than 85 percent of users appreciated the interface as intuitive and useful, which confirmed the accessibility of the system to a wide audience. By substituting conventional GIS software complexities with natural language interaction, our system enables the larger endeavor of democratizing satellite intelligence [1], [12].

In summary, the VQA system attains high prediction precision, rapid response under load, robust operational uptime, and convenient user accessibility. It generalizes well to different domains and locations but needs improvement in temporal query handling. The findings authenticate the system design methodology and practical applicability of transformer-based multimodal VQA architectures to geospatial applications. Future enhancements need to involve temporal fusion models [21], synthetic training augmentation [20], and multilingual NLP modules [6] to increase usability and resilience.

Key Findings The model showed robust capacity in land classification and was able to discriminately identify water, urban, and vegetation regions as documented in [12]. Its average query response time of about 2.8 seconds shows potential for real-time use. The model lost accuracy, however, when answering complex queries that needed advanced spatial reasoning, which presents an area of improvement. In spite of this restriction, the model demonstrated generalization capability across various datasets and situations, indicating powerful potential for generalizability in remote sensing and geographic information systems.

5 CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

The VQA system for satellite imagery proposed here seamlessly integrates deep learning, computer vision, and natural language processing to analyze sophisticated geospatial information using natural language inquiries. With a model precision of 92.4 percent and a mean response time of less than 3 seconds, the system shows robust performance in land classification, disaster analysis, and urban planning [12]. By combining transformer-based architecture like ViLT [5] and BERT [10], the model provides a strong and scalable multimodal pipeline for real-time satellite image reasoning. Utilization of benchmark datasets like RSVQA [12] and attention-based fusion mechanisms [6], [14] also asserts the technical solidity as well as generalizability of the system. Being intuitive with its web interface and high rate of acceptance by users, it is applicable to both technical and non-technical stakeholders [1]. Therefore, the system has great potential to democratize geospatial intelligence and be a core tool for future applications of AI-powered Earth observation.

5.2 Future Scope

The VQA system designed for satellite imagery is a critical step in geospatial AI, but there are a few very promising directions to further improve it. The biggest direction in the future is multispectral and hyperspectral satellite data integration. Though the existing system is tuned for RGB-based imagery, expanding support to multi-band inputs will enable richer thematic analysis like vegetation health, soil moisture status, or mineral content. Adding multispectral features would significantly benefit agricultural and environmental science applications, which rely on spectral bands beyond the visible spectrum [20]. This would involve modifying the image encoding modules and increasing the feature extraction architecture to support more channels, as proposed in previous deep learning surveys in remote sensing [18].

Another significant improvement is in integrating real-time satellite streams. Currently, the system handles static satellite images. Nonetheless, with the integration of live streams from sources like Sentinel-2, Landsat-9, or commercial satellites, the model could facilitate dynamic, time-critical decision-making for disaster relief, forest fire alerts, and city encroachment tracking [7], [21]. Real-time VQA systems would not only need low-latency pipelines but also strong data ingestion, preprocessing, and inference components with the ability to cope with high-frequency updates. Methods like model pruning and edge optimization might help to deliver real-time speeds even with bigger models [6].

One of the most important accessibility improvements would be the support of multilingual query processing. Although the system already employs English NLP models, incorporating multilingual transformers like XLM-R or mBERT would allow for broader usability in non-English-speaking areas [10]. This is most applicable to areas constantly surveyed using remote sensing but with no access to domain professionals or English-based systems. Incorporating multilingual support accords with the goal of democratization outlined in the opening of this research [1], [12].

In order to enhance learning effectiveness and responsiveness, the system may adopt interactive learning based on user feedback. Through rating or correction of the system's response by users, the VQA model may adjust its parameters dynamically or retrain over time, which would make it more sensitive to context and aware of domains. This interactive training cycle would strengthen the model against varying uses and geographical conditions, particularly if fine-tuned with local information or language idioms [12]. With a reinforcement learning approach based on feedback, this would greatly improve model performance in uncertain or under-represented cases.

The second important extension is edge and mobile deployment, enabling the system to be utilized under low-infrastructure conditions. Light iterations of the model—exported through TensorFlow Lite or ONNX—would be deployable on edge devices like drones or Raspberry Pi modules for farm monitoring or in-loc disaster response [12]. VQA deployable at the edge would enable real-time satellite reasoning to be feasible even under field conditions with poor

connectivity, opening up the scope of the system to be expanded to rural and remote areas.

Later releases can also investigate domain-specific fine-tuning. In agriculture, for instance, the model may be fine-tuned to respond to queries on crop health, yield prediction, and irrigation requirements. For urban planning, questions can focus on traffic flow, zoning regulations, or city development [2], [17]. Environmental monitoring activities like deforestation monitoring or analysis of glacier retreat may be enhanced by coupling VQA outputs with climate and elevation information, such as with integration into Digital Elevation Models (DEMs) and temporal stacks of images [21].

Finally, interoperability with open geospatial platforms like Google Earth Engine, QGIS, or OpenStreetMap would allow users to superimpose VQA outputs over other datasets, enhancing usability and spatial verification [2]. This would also promote adoption from researchers and practitioners presently working with established GIS ecosystems. With continued expansion of remote sensing VQA, collaborative development with the open-source community can expedite innovation while facilitating worldwide applicability.

In summary, the system today sets a robust foundation for natural language-based understanding of satellite images. Yet, by broadening spectral input support [20], facilitating real-time and multilingual interaction [6], applying user feedback [12], focusing on domain-specific use cases [17], and deploying edge [12], the system can be transformed into a broadly usable and scalable geospatial intelligence platform.

References

1. Lobry, S. et al., "Visual Question Answering for Remote Sensing," CVPR, 2020.
2. Zhang, H. et al., "Hybrid GIS and VQA in Urban Planning," Remote Sensing Journal, 2023.
3. Goodfellow, I. et al., Deep Learning, MIT Press, 2016.
4. Vaswani, A. et al., "Attention is All You Need," NeurIPS, 2017.
5. Kim, W. et al., "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision," ICML, 2021.
6. Akbari, H. et al., "VATT: Video-Audio-Text Transformer," NeurIPS, 2021.
7. Ghamisi, P. et al., "Remote Sensing Big Data," IEEE JSTARS, 2019.
8. Antol, S. et al., "VQA: Visual Question Answering," ICCV, 2015.
9. Krishna, R. et al., "Visual Genome," IJCV, 2017.
10. Devlin, J. et al., "BERT: Pre-training of Deep Bidirectional Transformers," NAACL, 2019.
11. Dosovitskiy, A. et al., "An Image is Worth 16x16 Words," ICLR, 2021.
12. Basu, S. et al., "RSVQA: A Benchmark Dataset for VQA in Remote Sensing," Earth Vision Workshop, CVPR 2021.
13. Xu, K. et al., "Show, Attend and Tell," ICML, 2015.
14. Anderson, P. et al., "Bottom-Up and Top-Down Attention," CVPR, 2018.

15. Hudson, D., Manning, C.D., "GQA: A New Dataset for Real-World Visual Reasoning and Compositional QA," CVPR, 2019.
16. Radford, A. et al., "Learning Transferable Visual Models from Natural Language Supervision," ICML, 2021.
17. Khandelwal, A. et al., "VQA in Urban Flood Detection," Remote Sensing Letters, 2022.
18. Chen, X. et al., "A Survey on Deep Learning in Remote Sensing," ISPRS Journal, 2020.
19. Tuia, D. et al., "Machine Learning in Remote Sensing," IEEE GRSM, 2016.
20. Li, X. et al., "Deep Learning Datasets for Satellite Imagery," IEEE Access, 2022.
21. Liu, Q. et al., "Spatial-Temporal Fusion for VQA," AAAI, 2021.