

A Project Report on

Visual Question Generation From Remote Sensing Images

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the
academic requirements for the award of the degree.

Bachelor of Technology

In

Computer Science and Engineering

Submitted by

Y. Laxmi Narayana
(20H51A0581)

B. Pravalika
(20H51A05B6)

Under the esteemed guidance of

Ms.M. Kamala
(Assistant Professor)



Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY
(UGC Autonomous)

*Approved by AICTE *Affiliated to JNTUH *NAAC Accredited with A⁺ Grade

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2020- 2024

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



CERTIFICATE

This is to certify that the Major Project report entitled "**VISUAL QUESTION GENERATION FROM REMOTE SENSING IMAGES** " being submitted by Y. Laxmi Narayana (20H51A0581), B. Pravalika (20H51A05B6) in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodied in this project report have not been submitted to any other University or Institute for the award of any Degree.

Ms. M. Kamala
Assistant Professor
Dept. of CSE

Dr. Siva Skandha Sanagala
Associate Professor and HOD
Dept. of CSE

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

With great pleasure, we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project a grand success.

We are grateful to **Ms. M. Kamala, Assistant Professor**, Department of Computer Science and Engineering for her valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala**, Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving force to complete my project work successfully.

We are very grateful to **Dr. Ghanta Devadasu**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Major Dr. V A Narayana**, Principal, CMR College of Engineering and Technology, for permitting us to carry out this project successfully and fruitfully.

We would like to thank the **Teaching & Non- teaching** staff of the Department of Computer Science and Engineering for their cooperation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, and **Shri. Ch. Abinav Reddy**, CEO, CMR Group of Institutions for their continuous care and support.

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in the completion of this project work.

Y. LaxmiNarayana - 20H51A0581
B. Pravalika - 20H51A05B6

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	LIST OF FIGURES	iii
	ABSTRACT	iv
	INTRODUCTION	1
1	1.1 Problem Statement	2
	1.2 Research Objective	3
	1.3 Project Scope and Limitations	4
2	BACKGROUND WORK	6
	2.1 Image Captioning And Question Answering Models	7
	2.1.1 Introduction	7
	2.1.2 Merits, Demerits & Challenges	8
	2.1.3 Implementation	9
	2.2 Transfer Learning From Vision& Language Models	10
	2.2.1 Introduction	10
	2.2.2 Merits, Demerits & Challenges	11
	2.2.3 Implementation	12
	2.3 Rule-Based Systems	13
	2.3.1 Introduction	13
	2.3.2 Merits, Demerits & Challenges	14
	2.3.3 Implementation	15
3	PROPOSED SYSTEM	18
	3.1 Objective of Proposed Model	19
	3.2 Algorithms Used for Proposed Model	20
	3.3 Designing	25
	3.3.1 UML Diagrams	25
	3.4 Stepwise Implementation and Code	26
	3.5 Model Architecture	35

	3.6 System Requirement	36
4	RESULTS AND DISCUSSION	38
	4.1 Output Screens	39
	4.2 Performance	42
5	CONCLUSION	43
	5.1 Conclusion and Future Enhancement	44
	REFERENCE	45
	GITHUB LINK	47
	Paper Publishing & Certificate	

List of Figures

FIGURE NO.	TITLE	PAGE-NO
2.1.3.1	Flow Chart of Existing Solution 1	9
2.2.3.1	Flow Chart of Existing Solution 2	12
2.2.3.2	Architecture	13
2.3.3.1	Model Diagram	17
3.2.1	The architecture of the proposed method	24
3.3.1	Class Diagram	25
3.5.1	Model Architecture	35
4.1.1	User Interface	39
4.1.2	Uploading Of Image	39
4.1.3	Extract Features & Generate Questions (eg1).	40
4.1.4	Extract Features & Generate Questions (eg2).	40
4.1.5	Extract Features & Generate Questions (eg3).	41
4.1.6	Extract Features & Generate Questions (eg4)	41

ABSTRACT

Visual question generation (VQG) is an emerging research area that aims to generate natural language questions based on visual content automatically. Here, this focuses on the application of VQG to remote sensing images, which are obtained from aerial or satellite sensors and provide valuable information for various domains, including agriculture, urban planning, and environmental monitoring. The proposed approach leverages deep learning techniques to extract meaningful features from remote sensing images and subsequently generate coherent and contextually relevant questions.

Convolution Neural Networks (CNNs) are utilized to capture spatial information from the images while Pre-trained models like Gemini API are employed to model sequential patterns in the generated questions. The combination of these architectures enables the model to comprehend and translate the visual context into human-readable questions effectively. To ensure the quality of the generated questions, a novel attention mechanism is introduced, allowing the model to focus on relevant image regions while formulating the questions.

CHAPTER 1

INTRODUCTION

CHAPTER 1

INTRODUCTION

1.1. Problem Statement

Remote sensing images provide valuable insights into various geographical, environmental, and infrastructural aspects of our planet. However, extracting meaningful information from these images often requires human interpretation, which can be time-consuming and labor-intensive. In recent years, there has been growing interest in developing automated systems capable of understanding and analyzing remote sensing images to assist in various applications such as urban planning, environmental monitoring, and disaster response.

This project aims to develop a system for automatically generating questions from remote-sensing images. This involves addressing several key challenges:

Image Understanding: Remote sensing images often contain complex spatial and spectral information. The system must effectively process and understand these images to identify salient features, objects, and patterns.

- **Question Generation:** The system must be capable of generating diverse and relevant questions based on the content of the remote sensing images. These questions should cover various aspects such as object identification, spatial relationships, and environmental conditions.
- **Natural Language Generation:** The questions generated should be grammatically correct, coherent, and semantically meaningful. Additionally, they should be formulated in a way that is easily understandable to users with diverse backgrounds.
- **Contextual Understanding:** Generating relevant questions requires understanding the context of the remote sensing images, including the geographical location, environmental conditions, and any relevant metadata associated with the images.
- **Evaluation Metrics:** Developing robust evaluation metrics to assess the quality and relevance of the generated questions is crucial for accurately evaluating the system's performance.

1.2. Research Objective

The objective of the project "Visual Question Generation From Remote Sensing Images" is to develop a novel methodology for automatically generating natural language questions based on visual content extracted from remote sensing images.

This research aims to bridge the gap between computer vision and natural language processing domains by leveraging deep learning techniques to understand and interpret remote sensing imagery in order to generate meaningful questions.

Specifically, the project aims to:

- Investigate and implement state-of-the-art deep learning architectures for image understanding and question generation, adapting them to the unique characteristics of remote sensing imagery.
- Explore techniques for extracting relevant visual features from remote sensing images, considering low-level features such as texture, color, and shape, as well as high-level semantic concepts.
- Investigate the incorporation of contextual information, such as geospatial metadata and temporal data, to enhance the relevance and specificity of generated questions.
- Evaluate the performance of the developed models using both quantitative metrics (e.g., question-answer consistency, diversity of generated questions) and qualitative assessments by human evaluators.
- Explore potential applications of the generated questions in various domains, including environmental monitoring, urban planning, disaster response, and agricultural management.
- Investigate strategies for fine-tuning and adapting pre-trained question generation models to specific remote sensing tasks and domains, considering factors such as domain transferability and generalization capabilities.

1.3 Project Scope:

- **Application Focus:** The project focuses on the application of Visual Question Generation (VQG) to remote sensing images.
- **Utilization of Deep Learning:** Deep learning techniques, particularly Convolution Neural Networks (CNNs), are employed to extract spatial information from remote sensing images.
- **Integration of Pre-trained Models:** Pre-trained models like Gemini API are utilized to capture sequential patterns in question generation, enhancing the coherence and relevance of the generated questions.
- **Cross-domain Relevance:** Remote sensing images cater to various domains including agriculture, urban planning, and environmental monitoring, thereby expanding the applicability of the proposed approach.
- **Quality Assurance:** A novel attention mechanism is introduced to ensure the quality of generated questions, enabling the model to focus on relevant image regions while formulating questions.

Limitations:

- **Data Availability:** The effectiveness of the proposed approach heavily relies on the availability and quality of annotated remote sensing image datasets, which may be limited or challenging to obtain in certain domains.
- **Generalization:** While the model may perform well on certain types of remote sensing images, its generalization across diverse scenarios and environments might be limited, requiring fine-tuning or domain-specific adaptations.
- **Shi, Z., & Zou, Z. [5]** This paper examines a charming address within the inaccessible detecting field: “Can a machine produce humanlike dialect portrayals for a further detecting image.
- **Computational Resources:** Deep learning models, especially when dealing with large-scale remote sensing images, require substantial computational resources for training and inference, which could pose constraints in practical deployment scenarios.
- **Interpretability:** Despite generating contextually relevant questions, the interpretability of the model's decision-making process might be challenging, raising concerns regarding transparency and trustworthiness, especially in critical applications.

CHAPTER 2

BACKGROUND WORK

CHAPTER 2

BACKGROUND WORK

2.1 IMAGE CAPTIONING AND QUESTION ANSWERING MODEL

2.1.1. Introduction:

- Purpose: Image captioning and question-answering models aim to bridge the gap between visual content and natural language understanding. These models generate descriptive captions for images and answer questions related to the content of the images.
- Xiong, Z., Zhang, F., Wang, Y., Shi, Y., & Zhu, X. X [1] Soil perception, pointing at observing the condition of planet Soil utilizing further detecting information, is basic for progressing our everyday lives and living environment.
- Applications: They find applications in various domains such as assistive technologies for visually impaired individuals, content recommendation systems, autonomous vehicles, medical image analysis, and more.
- Technological Landscape: These projects leverage advancements in deep learning, particularly in computer vision and natural language processing (NLP), to understand image content and generate coherent textual descriptions or answers.
- Data Requirements: Successful implementation relies heavily on large, diverse datasets with annotated images and corresponding captions or questions, such as COCO (Common Objects in Context) for image captioning and VQA (Visual Question Answering) for question answering.

2.1.2. Merits, Demerits, and ChallengesMerits:

- Enhanced Accessibility: These models provide accessibility to visual content for individuals with visual impairments by converting images into descriptive text or answering questions about them.
- Improved User Experience: They enhance user experience in applications such as content recommendation systems by providing more informative captions or answering user queries related to images.

- **Versatility:** Image captioning and question-answering models are versatile and can be applied across various domains, from entertainment to healthcare, providing valuable insights and information

Demerits:

- **Bias Amplification:** These models can inherit biases present in the training data, leading to biased captions or answers, which may reinforce societal biases.
- **Ambiguity Handling:** Ambiguities in images or questions can pose challenges for these models, leading to inaccurate captions or answers.
- **Scalability:** Developing robust models that can handle a wide range of images and questions requires significant computational resources and training data, making scalability a challenge.
- **Ethical Considerations:** Generating captions or answers may raise ethical concerns, especially in sensitive domains like healthcare or law enforcement, where inaccuracies could have serious consequences.

Challenges:

- **Semantic Understanding:** Understanding the semantic context of images and questions accurately remains a challenge, particularly in complex scenarios or with abstract concepts
- **Data Quality:** Ensuring the quality and diversity of training data is crucial for building robust models that generalize well to unseen images and questions.
- **Evaluation Metrics:** Developing effective evaluation metrics to measure the performance of image captioning and question-answering models accurately is non-trivial due to the subjective nature of tasks.
- **Real-time Processing:** Achieving real-time performance for these models is challenging, especially in applications where low latency is critical, such as autonomous vehicles or live video streaming platforms.

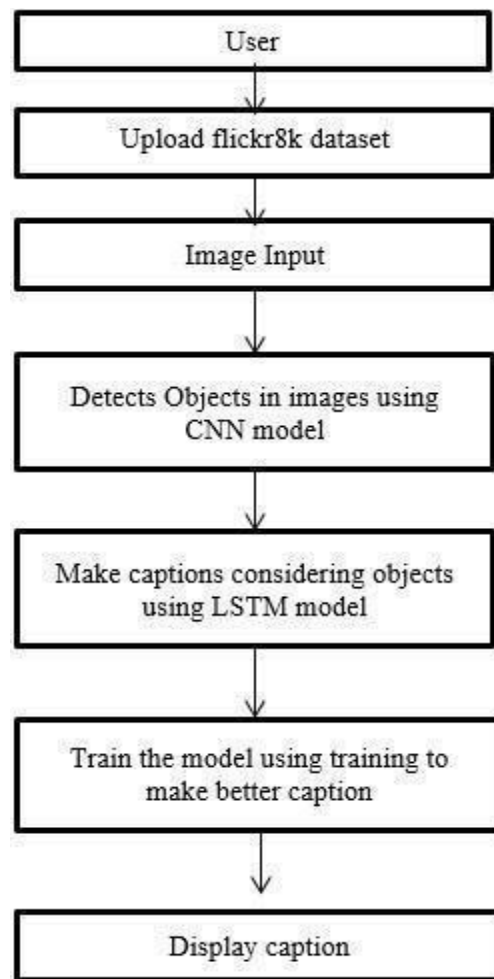
2.1.3. Implementation:

Figure: Flow Chart

Fig :2.1.3.1 Flow Chart of Existing Solution 1

- **Model Architecture:** Implementing state-of-the-art architectures such as Transformer-based models like BERT or T5 for question answering, and architectures like CNN-LSTM for image captioning.
- **Data Preprocessing:** Preprocess image and text data, including image resizing, tokenization, and data augmentation techniques to improve model robustness.
- **Training Pipeline:** Set up a training pipeline using frameworks like TensorFlow or PyTorch, leveraging powerful hardware such as GPUs or TPUs for efficient training.
- **Deployment:** Deploy trained models using platforms like TensorFlow Serving or ONNX Runtime, optimizing for performance and scalability, and integrating them into target applications or systems.

2.2.1. TRANSFER LEARNING FROM VISION AND LANGUAGE MODELS

2.2.1. Introduction:

- **Transfer Learning Concept:** Transfer learning from vision and language models refers to leveraging pre-trained models developed for tasks such as image recognition and natural language processing (NLP) to improve performance on related tasks. This approach exploits the knowledge learned from large-scale datasets and tasks to benefit smaller or related tasks, thus reducing the need for extensive data and computational resources.
- **Fu, K., Li, Y., Zhang, W., Yu, H., & Sun, X [3]** The encoder-decoder system has been within the further detecting picture captioning task.
- **Interdisciplinary Approach:** This project combines insights from computer vision and natural language processing, tapping into the synergy between these domains. By integrating vision and language models, it aims to enable machines to understand and interpret both visual and textual information simultaneously, mimicking human cognitive abilities.
- **Scope of Applications:** The applications of transfer learning from vision and language models are vast, ranging from multimodal understanding tasks such as image captioning, visual question answering, and multimodal sentiment analysis to more complex tasks like medical image analysis, autonomous driving, and virtual assistants.
- **State-of-the-Art Advancements:** Recent advancements in deep learning, especially in transformer-based architectures like BERT (Bidirectional Encoder Representations from Transformers) and vision transformers (ViTs), have significantly improved the feasibility and effectiveness of transfer learning from vision and language models. This project aims to leverage these advancements to push the boundaries of multimodal AI.

2.2.2. Merits, Demerits, and Challenges Merits:

- **Efficiency in Model Training:** Transfer learning reduces the computational resources required for training new models from scratch by leveraging pre-trained weights. This efficiency makes it feasible for smaller organizations or research teams with limited resources to develop advanced AI systems

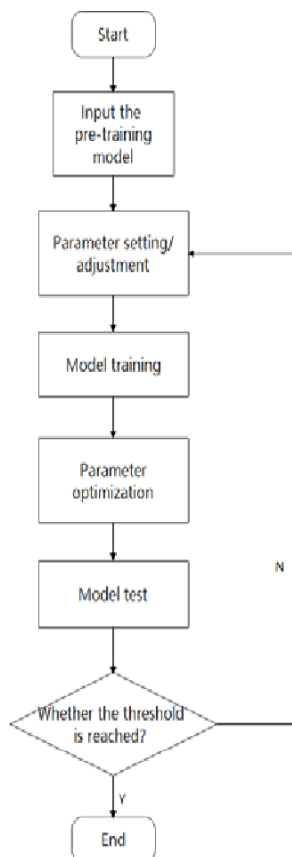
- **Domain Adaptation:** Transfer learning facilitates domain adaptation, enabling models trained on one domain to be effectively applied to related but different domains. This capability is particularly valuable in scenarios where collecting labeled data in the target domain is challenging or expensive.
- **Faster Deployment:** With transfer learning, models can be deployed more quickly since a significant portion of the learning process has already been completed during pre-training. This rapid deployment is crucial in dynamic environments where timely solutions are essential.

Demerits:

- **Limited Task Specificity:** While pre-trained models capture generic features, they may not always capture domain-specific nuances crucial for certain tasks. Fine-tuning these models on specific tasks can mitigate this limitation to some extent, but there's still a risk of insufficient adaptation to the target task.
- **Overfitting Risks:** Transfer learning runs the risk of overfitting to the target task's training data, especially when the target dataset is small or significantly different from the pre-training data. Balancing the fine-tuning process to avoid overfitting while retaining the benefits of transfer learning requires careful regularization techniques.
- **Dependency on Pre-training Data:** The effectiveness of transfer learning heavily depends on the quality and representativeness of the pre-training data. Biases and limitations present in the pre-training data can propagate to the fine-tuned models, potentially leading to biased or unreliable predictions, especially in sensitive applications like healthcare or criminal justice.
- **Computational Resources for Fine-tuning:** While transfer learning reduces the computational burden compared to training from scratch, fine-tuning still requires significant computational resources, particularly for large-scale models and datasets. This can pose challenges for organizations with limited access to high-performance computing infrastructure.

Challenges:

- **Data Heterogeneity:** Combining vision and language data introduces heterogeneity, as these modalities may have different distributions, scales, and semantics. Aligning these diverse data sources effectively while preserving their respective features is a non-trivial challenge.
- **Semantic Alignment:** Ensuring that the learned representations from both vision and language modalities are semantically aligned is crucial for effective multimodal understanding. Developing techniques to bridge the semantic gap between visual and textual information remains a challenging research problem.
- **Evaluation Metrics:** Assessing the performance of multimodal models poses challenges in defining appropriate evaluation metrics that capture the nuances of both vision and language tasks. Developing comprehensive evaluation protocols that account for the interplay between different modalities is essential for meaningful performance assessment.

2.2.3. Implementation:**Fig :2.2.3.1 Flow Chart**

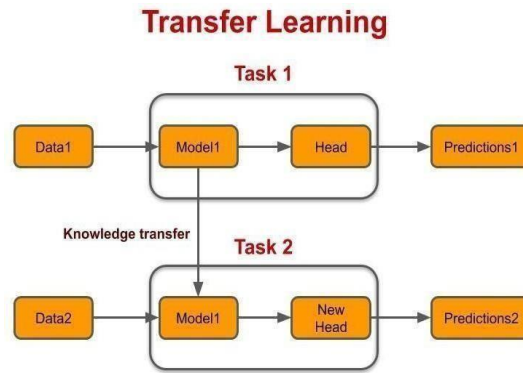


Fig :2.2.3.2 Architecture

- **Data Preprocessing:** Prepare the multimodal dataset by preprocessing both vision and language data, including tasks such as image resizing, tokenization, and text normalization
- **Model Selection:** Choose appropriate pre-trained vision and language models based on the specific task requirements and available computational resources. Popular choices include pre-trained vision transformers (ViTs), BERT, and their variants
- **Fine-tuning Strategy:** Define a fine-tuning strategy that balances between leveraging the knowledge from pre-trained models and adapting to the target task. This involves selecting hyperparameters, regularization techniques, and optimization algorithms
- **Evaluation and Iteration:** Evaluate the fine-tuned model on validation data using appropriate evaluation metrics. Iterate on the fine-tuning process by adjusting hyperparameters and model architectures based on validation performance until satisfactory results are achieved.
- **Deployment and Monitoring:** Deploy the trained model in the target environment and continuously monitor its performance to ensure robustness and reliability. Fine-tune the model further if necessary based on real-world feedback and evolving task requirements.

2.3 RULE-BASED SYSTEMS

2.3.1. Introduction

- **Definition:** Rule-based systems (RBS) are a type of artificial intelligence (AI) system that operates on a set of predefined rules and logical operations to make decisions or perform tasks. These rules are typically represented in the form of "if-then" statements.

- Purpose: RBS is commonly used in various domains such as expert systems, decision support systems, and business rule engines to automate decision-making processes and streamline operations.
- Functionality: The core functionality of RBS involves evaluating conditions based on input data and executing corresponding actions according to the predefined rules.
- Zhang, Z., Diao, W., Zhang, W., Yan, M., Gao, X., & Sun, X [4] Critical advances have been made in further detecting picture captioning by encoder-decoder systems.
- Examples: Examples of RBS include medical diagnosis systems, fraud detection systems, and recommendation engines, where specific rules are applied to analyze data and generate appropriate responses.

2.3.1. Merits, Demerits and Challenges Merits:

- Transparency: Rule-based systems are inherently transparent since the rules governing their behavior are explicitly defined. This transparency is crucial in domains where understanding the decision-making process is essential, such as legal or medical contexts.
- Ease of Maintenance: Modifying or updating the rules in RBS is relatively straightforward, allowing for easy adaptation to changing requirements or environments. This flexibility reduces the cost and effort associated with system maintenance.
- Interpretability: Due to their explicit rule representation, RBS provides clear explanations for their decisions, enhancing user trust and facilitating easier debugging in case of errors or unexpected outcomes.
- Scalability: RBS can handle large amounts of data and complex decision-making scenarios efficiently, making them suitable for applications ranging from simple business rules to intricate expert systems.

Demerits:

- **Limited Complexity:** RBS is best suited for problems with well-defined rules and clear decision boundaries. They may struggle with handling complex, ambiguous, or uncertain situations where rules cannot be easily formulated
- **Brittleness:** Since RBS relies heavily on predefined rules, they may exhibit brittleness .
- **Knowledge Acquisition Bottleneck:** Building effective rule sets often requires significant domain expertise and effort to accurately capture the decision-making logic. Acquiring, validating, and updating these rules can be time-consuming and resource-intensive.
- **Maintenance Overhead:** While RBS offers ease of maintenance compared to some other AI approaches, managing a large number of rules and ensuring their consistency and relevance over time can still pose challenges, especially in dynamic environments.

Challenges:

- **Rule Conflict Resolution:** In complex systems with numerous rules, conflicts may arise where multiple rules apply to the same situation, leading to ambiguity or inconsistency. Resolving such conflicts effectively is a non-trivial challenge.
- **Handling Uncertainty:** Real-world data often contains uncertainties and inaccuracies that cannot be fully captured by deterministic rules. Integrating techniques for handling uncertainty, such as probabilistic reasoning, into RBS is a challenging task.
- **Adaptability to Dynamic Environments:** RBS may struggle to adapt to dynamic environments where rules need to be continuously updated or refined in response to changing conditions or new information. Ensuring timely and accurate rule adaptation is a significant challenge
- **Integration with Other Systems:** Integrating RBS with existing IT infrastructure or other AI technologies can be challenging, especially when dealing with heterogeneous data sources or complex architectures. Ensuring seamless interoperability and data flow is crucial for successful integration

2.3.2. Implementation

- **Requirements Analysis:** Begin by thoroughly understanding the problem domain and identifying the specific requirements and objectives that the RBS should address.
- **Rule Elicitation and Formalization:** Collaborate with domain experts to elicit and formalize the rules governing the decision-making process. Document these rules in a clear and structured manner.
- **Rule Engine Selection:** Choose an appropriate rule engine or development framework based on factors such as scalability, performance, and compatibility with existing infrastructure.
- **Prototype Development and Testing:** Develop a prototype RBS implementation incorporating the defined rules and test it rigorously using representative datasets and scenarios validate its accuracy and performance.
- **Deployment and Integration:** Deploy the RBS into the production environment, ensuring seamless integration with other systems and providing necessary interfaces for data input and output. Monitor the system closely post-deployment to identify and address any issues that may arise.
- **Maintenance and Iteration:** Establish a process for ongoing maintenance and iteration, including regular updates to the rule set based on feedback and changing requirements. Continuously monitor the system's performance and effectiveness, making adjustments as necessary to ensure optimal functionality.
- **Deploy the Rule-Based System (RBS)** into the production environment, ensuring smooth integration with existing systems and establishing interfaces for data input and output. Continuously monitor the system post-deployment to promptly detect and resolve any issues.
- **Work closely with subject matter experts** to extract and formalize the rules guiding decision-making. Clearly document these rules in a structured format.
- **Rule Engine Selection:** Choose an appropriate rule engine or development framework based on factors such as scalability, performance, and compatibility with existing infrastructure.

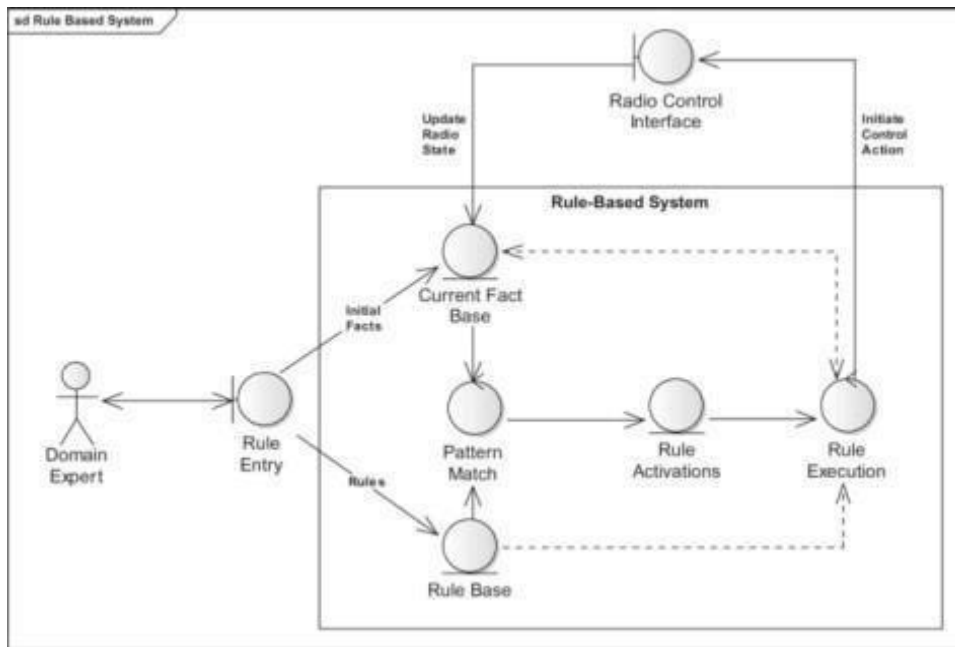


Fig 2.3.3.1: Model Diagram

CHAPTER3

PROPOSED SYSTEM

CHAPTER 3

PROPOSED SYSTEM

3.1. The Objective of the Proposed Model

- **Integration of Gemini API with BERT:**

The integration of the Gemini API with BERT in visual question generation represents a powerful fusion of natural language processing (NLP) and computer vision techniques. GeminiAPI provides access to advanced image analysis capabilities, while BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art NLP model known for its contextual understanding of language.

- **Enhanced Image Understanding:**

Enhanced Image Understanding in Visual Question Generation refers to the integration of advanced image processing techniques with natural language processing (NLP) to generate questions about images. This approach aims to improve the quality and relevance of questions generated by considering not only the visual content of the image but also its context and semantics.

- **Context-Aware Question Generation:**

Context-Aware Question Generation in visual question generation refers to the process of creating questions based on the content and context of visual information, such as images or videos. This approach aims to generate questions that are relevant and meaningful with respect to the visual content, enhancing the overall understanding and engagement of users.

- **Evaluation Metrics and Benchmarking:**

Evaluation metrics and benchmarking play a crucial role in the development and advancement of visual question generation (VQG) systems. These metrics are essential for assessing the quality and performance of VQG models objectively. Common evaluation metrics include measures of question relevance, diversity, and grammaticality. Additionally, benchmark datasets such as VizWiz and COCO-QA provide standardized benchmarks against which VQG models can be evaluated and compared..

- **Real-World Application and Impact Assessment:**

Real-world applications of visual question generation (VQG) systems span diverse domains such as education, accessibility, and automation. In education, VQG facilitates interactive learning experiences by generating questions based on visual content, fostering deeper understanding and engagement among learners. Additionally, via inaccessibility, VQG empowers visually impaired individuals by enabling them to interact with visual content through textual questions, thus enhancing their access to information and digital resources. Furthermore, in automation, VQG streamlines content creation processes by automatically generating questions for training datasets or assessments, saving time and effort for educators and content creators. The impact assessment of VQG lies in its ability to democratize access to education and information, improve learning outcomes, and enhance efficiency in content-generation workflows across various sectors. Through continuous evaluation and refinement, VQG systems hold the potential to revolutionize how we engage with visual content and foster knowledge acquisition in both traditional and digital learning environments.

3.2. Algorithms Used for Proposed

Model Gemini API

BERT

CNN

- **Gemini API :**

Gemini API is a powerful tool used in visual question generation, which is the process of automatically creating questions about images or visual content. The Gemini API leverages state-of-the-art natural language processing (NLP) and computer vision techniques to analyze images and generate relevant and meaningful questions. It allows developers and researchers to integrate advanced question-generation capabilities into their applications or systems with ease.

Key features of the Gemini API include:

Image Understanding: The API utilizes advanced computer vision algorithms to understand the content and context of images, enabling it to generate questions that are pertinent to the visual information present.

- **Natural Language Processing:** Leveraging sophisticated NLP models, the Gemini API generates questions that are grammatically correct, semantically meaningful, and contextually relevant to the image.
- **Customization:** Users can customize the question generation process according to specific requirements or preferences, such as adjusting the complexity of questions, selecting question types, or specifying target audiences.
- **Scalability and Efficiency:** Gemini API is designed to be scalable and efficient, allowing it to process large volumes of images and generate questions quickly and accurately.
- **Integration:** The API provides easy integration with existing applications or systems through well-documented APIs and SDKs, enabling seamless incorporation of visual question generation capabilities into various platforms.

In summary, Gemini API plays a crucial role in automating the process of generating questions from images, offering advanced image understanding and natural language processing capabilities to developers and researchers in the field of visual question generation.

- **BERT:**

- BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art natural language processing model developed by Google, known for its ability to understand context and generate high-quality representations of text. In the context of visual question generation (VQG), BERT has been applied to enhance the generation of questions based on images
- Using BERT in VQG involves leveraging its pre-trained language understanding capabilities to process both the textual information related to an image and the visual features extracted from the image itself. By feeding both modalities into the model, BERT can effectively understand the context of the image and generate relevant questions about it.
- **Bidirectional Context:** Unlike previous models that processed text in one direction (e.g., left-to-right), BERT can capture the context from both directions, which leads to a better understanding of the semantics of a sentence.

State-of-the-Art Performance: BERT has achieved state-of-the-art results on a wide range of NLP benchmarks, surpassing previous models in tasks such as text classification, named entity recognition, question answering, and more.

- **Contextual Understanding:** BERT excels at capturing the contextual nuances of language, enabling it to generate questions that are more relevant and contextually appropriate based on the visual content of an image.
- **Fine-tuning:** BERT can be fine-tuned on VQG datasets to adapt its pre-trained representations specifically for the task of generating questions about images. Fine-tuning allows BERT to learn the correlations between visual features and corresponding questions more effectively.
- **Multimodal Fusion:** BERT can integrate information from both text and image modalities through techniques such as attention mechanisms or fusion layers. This multimodal fusion enables BERT to generate questions that are not only based on the visual content but also take into account relevant textual context.
- **Improved Question Quality:** By leveraging BERT's powerful language understanding capabilities, VQG systems incorporating BERT tend to produce questions that are grammatically correct, semantically meaningful, and contextually coherent, leading to higher-quality question generation overall.
- **Challenges:** Despite its effectiveness, integrating BERT into VQG systems poses challenges such as computational complexity, especially when processing large-scale datasets or real-time inference scenarios. Additionally, ensuring that BERT understands the intricacies of both textual and visual information accurately requires careful training and optimization.

CNN:

Convolutional Neural Networks (CNN) play a pivotal role in visual question generation, a task that involves generating natural language questions based on visual content.

CNN stands for Convolutional Neural Network, which is a type of deep learning algorithm mainly used for image recognition and classification tasks. It's inspired by the organization of the animal visual cortex and designed to automatically and adaptively learn spatial hierarchies of features from raw data.

CNNs consist of multiple layers, including convolutional layers, pooling layers, and fully connected layers. Convolutional layers apply convolution operations to the input data, extracting features through filters or kernels. Pooling layers downsample the feature maps, reducing their dimensionality and extracting the most important information. Fully connected layers combine the features extracted by the previous.

- **Feature Extraction:** CNNs are utilized to extract meaningful visual features from images. These networks are adept at capturing hierarchical representations of visual information, enabling them to identify objects, shapes, textures, and other relevant visual elements.
- **Integration with Language Models:** Visual features extracted by CNNs are combined with textual information using techniques like attention mechanisms or fusion methods. This integration facilitates the generation of coherent and contextually relevant questions based on the visual input.
- **End-to-end Models:** Some approaches employ end-to-end models that jointly learn to extract visual features and generate questions, leveraging the capabilities of CNNs for feature extraction and recurrent neural networks (RNNs) or transformer models for sequence generation.
- **Data Efficiency:** CNNs are known for their ability to learn hierarchical features from large-scale visual datasets, which is crucial for training robust visual question generation models. The availability of pre-trained CNNs further enhances data efficiency by enabling transfer learning.
- **Semantic Understanding:** CNNs contribute to the semantic understanding of visual content, allowing question generation models to produce questions that are not only syntactically correct but also semantically meaningful concerning the depicted scene.
- **Challenges:** Despite their effectiveness, challenges such as handling diverse visual contexts, maintaining spatial relationships, and addressing biases in visual datasets remain areas of ongoing research in the integration of CNNs with visual question generation models.
- **Convolutional Layers:** These layers apply convolution operations to the input data. They consist of a set of filters that are convolved with the input data, resulting in feature maps that highlight specific features in the input.
- **Pooling Layers:** Pooling layers are used to reduce the spatial dimensions of the feature maps produced by the convolutional layers, thus reducing the computational complexity of the network while preserving important information.
- **Activation Functions:** Typically, activation functions like ReLU (Rectified Linear Unit) are applied after convolutional and pooling layers to introduce non-linearity into the network, enabling it to learn complex patterns.

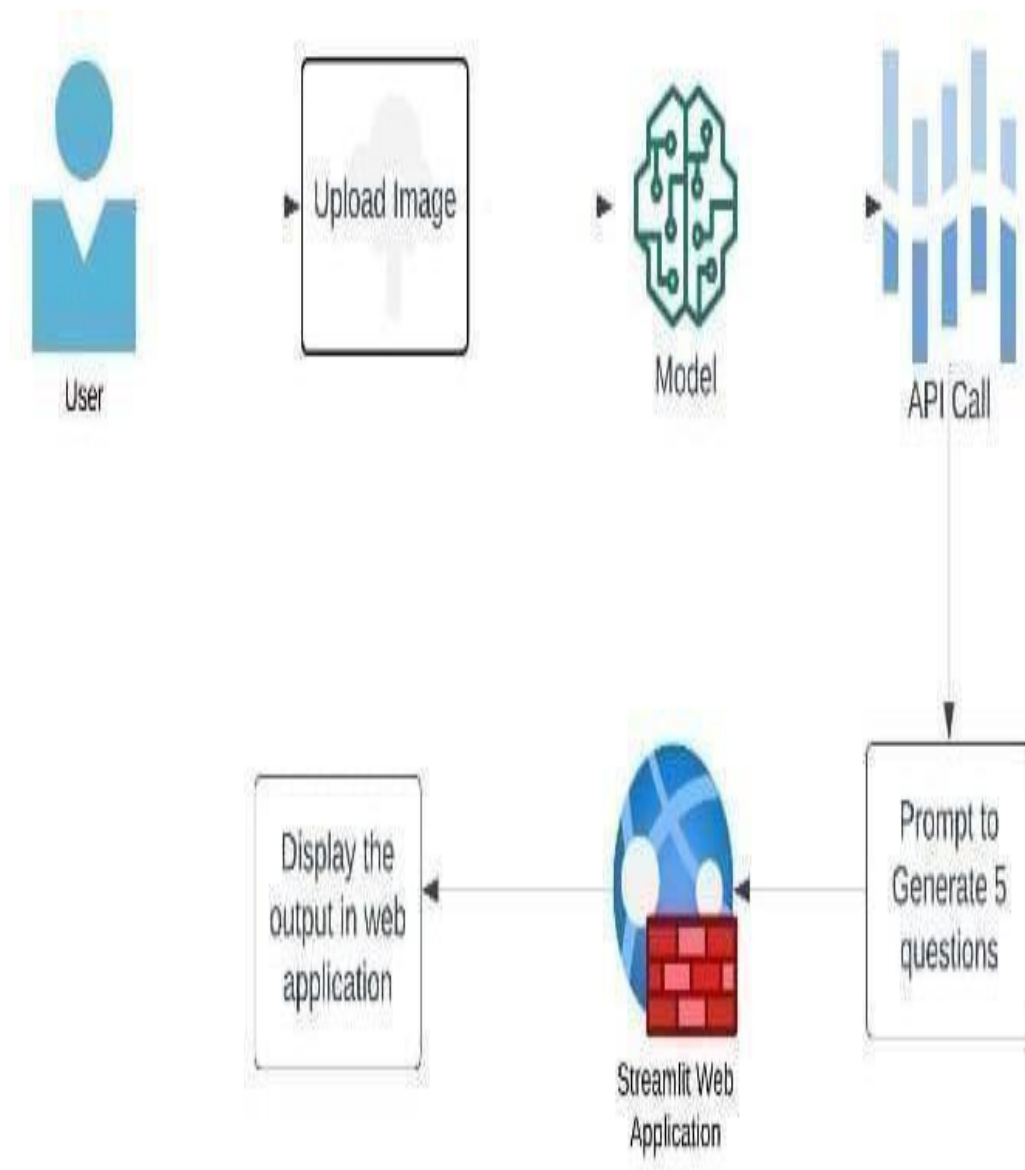


Fig:3.2.1 Architecture

Gemini API and BERT are two key components in the usage of visual question generation from remote sensing images:

Gemini API: Gemini is a platform that provides access to a variety of remote sensing data, including satellite imagery. It offers an API (Application Programming Interface) that allows developers to programmatically access and manipulate this data. In the context of visual question generation from remote sensing images, Gemini API can be utilized to retrieve the relevant images for analysis and question generation.

BERT (Bidirectional Encoder Representations from Transformers): BERT is a state-of-the-art natural language processing model developed by Google. It's pre-trained on large amounts of textual data and is capable of understanding context and nuances in language. In the context of visual question generation, BERT can be used to process textual inputs (questions) and generate appropriate responses based on the content of the images retrieved using the Gemini API.

Together, these technologies enable the automatic generation of questions about remote-sensing images. The Gemini API retrieves the images, which are then processed by BERT to generate questions based on the content of the images. This approach can be useful in various applications, such as environmental monitoring, urban planning, agriculture, and disaster management, where analyzing remote sensing imagery is crucial for decision-making.

3.3. Designing

3.3.1. UML Diagram

A. Class diagram:-

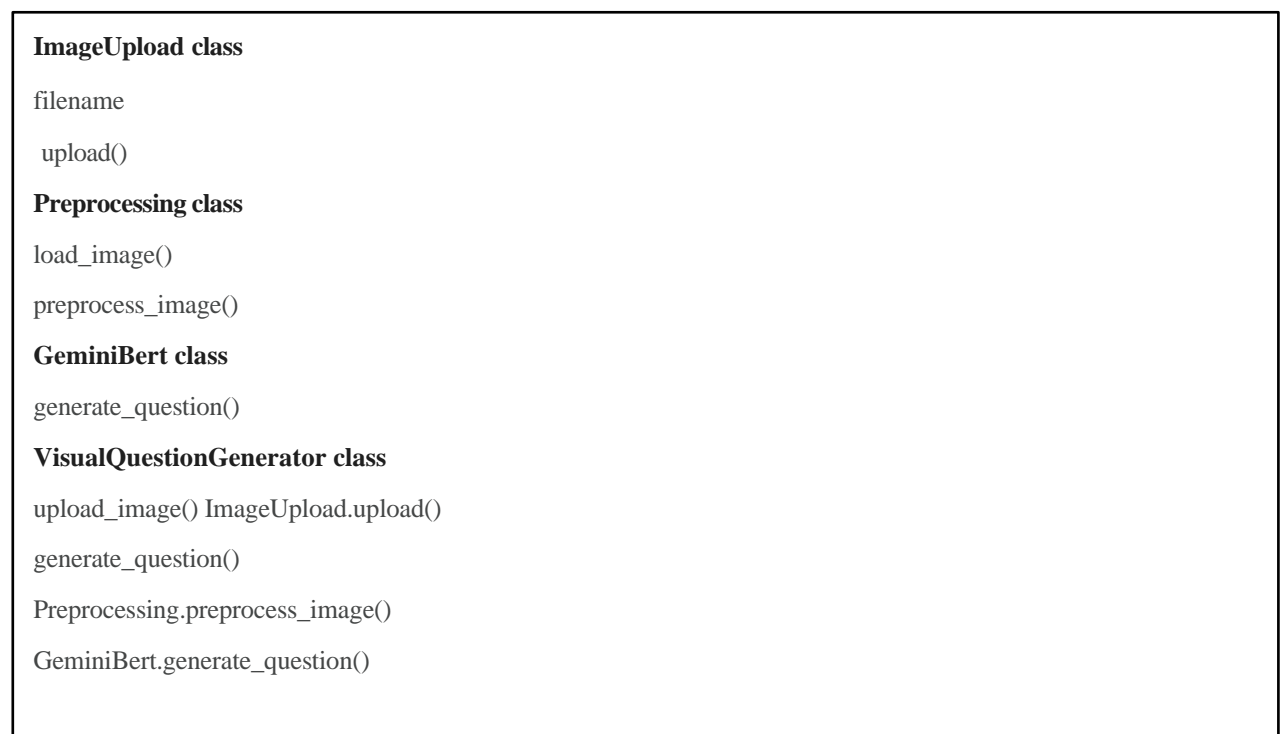


Fig:3.3.1.Class Diagram

3.4. Stepwise Implementation and Code:

- User Upload: Remote-sensing image uploaded to web app.
- API Call: Image sent to Gemini API Bert-3 server for analysis. Question Generation: Pre-trained model generates questions based on image analysis.
- Display: Questions are displayed on the web app interface for user interaction.

BERT analyses visual details, generating questions for deeper exploration. CNNs process images, extracting features to aid question generation. Bert's utilization of visual question generation from remote sensing images is vital as it can interpret intricate visual elements and convert them into coherent questions.

CODE:

App.py:-source code

```
import streamlit as st

import text wrap

import os

import PIL.Image

import google.generativeai as genai

# Used to securely store your API key

os.environ['GOOGLE_API_KEY'] = 'AIzaSyAc7Ii4wHf_whau2q--rgjfdht8-I5xhSY'

GOOGLE_API_KEY = os.getenv('GOOGLE_API_KEY')

genai.configure(api_key=GOOGLE_API_KEY)

for m in genai.list_models():

    if 'generateContent' in m.supported_generation_methods:

        generative_model_name = m.name

        break
```

```
def generate_questions_from_image(image):  
  
    # Display the uploaded image  
  
    st.image(image, caption='Uploaded Image', use_column_width=True)  
  
    # Initialize generative model  
  
    model = genai.GenerativeModel('gemini-pro-vision')  
  
    # Generate questions based on the image  
  
    response = model.generate_content(["Prepare 5 questions for the given image", image],  
stream=True)  
  
    response.resolve()  
  
    # Display the generated questions  
  
    st.write(response.text)  
  
    print(response.text)  
  
def about_project():  
  
    st.title("About Project")  
  
    st.write(  
  
        """"  
  
        ## Project Overview  
  
        The Image to Questions Generator is a web application designed to generate questions  
        based on an uploaded image. Leveraging state-of-the-art generative AI models, the application  
        processes the content of the image and formulates questions to prompt further understanding  
        or analysis.  
  
        ## Motivation  
  
        Visual content is abundant on the internet, and extracting meaningful information from  
        images is crucial for various applications such as educational content creation, image analysis,  
        and content moderation. However, manually generating questions from images can be time-  
        consuming and resource-intensive. Hence, the aim of this project is to automate this process  
        using machine learning techniques.
```

Features

- **Image Upload**: Users can upload an image in common formats like JPG, JPEG, or PNG.
- **Question Generation**: Upon image upload, the application processes the image content using a generative AI model to formulate relevant questions.
- **Streamlit Interface**: The application is built using Streamlit, a user-friendly framework for creating data-focused web applications in Python.
- **Markdown Display**: The generated questions are displayed using Markdown formatting for clarity and readability.

Technologies Used

- **Streamlit**: Used to build the web interface of the application.
- **PIL (Python Imaging Library)**: Utilized to handle image uploads and processing.
- **GenerativeAI by Google**: Integrated to access state-of-the-art generative AI models for question generation.

```
"""  
  
)  
  
st.markdown("<h1 style='text-align: left; color: white; font-size: 20px;'>Architecture of the  
project:</h1>", unsafe_allow_html=True)  
  
st.image('Gemini.jpeg', use_column_width=True)  
  
st.write(  
  
"""  
  
## How It Works
```

1. **Image Upload**: Users upload an image using the provided file uploader.
2. **Question Generation**: The uploaded image is processed by a pre-trained generative AI model capable of understanding visual content.
3. **Output Display**: The generated questions are displayed on the web interface, providing users with prompts based on the content of the uploaded image.

Future Enhancements

- **Improved Question Quality**: Refinement of the generative model or integration of feedback mechanisms to enhance the quality of generated questions.
- **Support for Different Languages**: Extension of the application to support question generation in multiple languages.
- **Customization Options**: Addition of features allowing users to customize the types or number of questions generated.

Conclusion

The Image to Questions Generator is a practical application of machine learning technology, offering a convenient solution for generating questions from visual content. Whether used for educational purposes, content analysis, or creative exploration, the application demonstrates the potential of AI-driven tools to simplify complex tasks.

```
""""  
  
)  
  
def image_to_questions():  
  
    st.title("Image to Questions Generator")  
  
    st.write("Upload an image and get questions generated based on it.")  
  
    uploaded_image = st.file_uploader("Choose an image...", type=["jpg", "jpeg", "png"])  
  
    if uploaded_image is not None:  
  
        img = PIL.Image.open(uploaded_image)  
  
        generate_questions_from_image(img)  
  
def main():  
  
    page = st.sidebar.radio("Go to", ["About Project", "Image to Questions"])  
  
    if page == "About Project":  
  
        about_project()  
  
    elif page == "Image to Questions":
```

```
image_to_questions()

if __name__ == "__main__":

    main()
```

Gemini_Image_Analysis.ipynb:-Source Code

```
!pip install -q -U google-generativeai

import pathlib

import textwrap

import google.generativeai as genai

# Used to securely store your API key

from google.colab import userdata

from IPython.display import display

from IPython.display import Markdown

def to_markdown(text):

    text = text.replace('•', ' *')

    return Markdown(textwrap.indent(text, '> ', predicate=lambda _: True))

import os

os.environ['GOOGLE_API_KEY'] = 'AIzaSyAc7Ii4wHf_whau2q--rgjfdht8-I5xhSY'

# Or use `os.getenv('GOOGLE_API_KEY')` to fetch an environment variable.

GOOGLE_API_KEY=os.getenv('GOOGLE_API_KEY')

genai.configure(api_key=GOOGLE_API_KEY)

for m in genai.list_models():

    if 'generateContent' in m.supported_generation_methods:

        print(m.name)

models/gemini-1.0-pro
```

```
models/gemini-1.0-pro-001  
models/gemini-1.0-pro-latest  
models/gemini-1.0-pro-vision-latest  
models/gemini-pro  
import PIL.Image  
img = PIL.Image.open('1.jpg')  
img  
model = genai.GenerativeModel('gemini-pro-vision')  
models/gemini-pro-vision  
model = genai.GenerativeModel('gemini-pro-vision')  
response = model.generate_content(img)  
to_markdown(response.text)
```

No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

Saving GCP ARCHETECTURE (1).jpeg to GCP ARCHETECTURE (1).jpeg

User uploaded file "GCP ARCHETECTURE (1).jpeg" with length 415262 bytes

<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=5141x3112>

What is the purpose of the Health Data Cloud Storage?

How is the data processed in the Processing Cloud Function?

What is the purpose of the Cloud SQL?

```
models/gemini-1.0-pro  
models/gemini-1.0-pro-001  
models/gemini-1.0-pro-latest  
models/gemini-1.0-pro-vision-latest  
models/gemini-pro
```

models/gemini-pro-vision

<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=474x296>

Peach leaf curl is a fungal disease that affects peach trees. The fungus overwinters in the buds of infected trees and produces spores in the spring that infect new leaves. Symptoms of peach leaf curl include puckering, curling, and yellowing of leaves. The leaves may also become distorted and stunted. Peach leaf curl can cause defoliation of trees and reduce fruit production. The disease is most severe in warm, humid climates.

There are a number of things that can be done to prevent and control peach leaf curl. These include:

Using resistant varieties of peach trees

Pruning infected trees in the fall or winter to remove diseased leaves and twigs

Applying a fungicide to the trees in the spring before buds break

Keeping the trees well-watered and fertilized

Peach leaf curl is a common disease, but it can be prevented and controlled with proper care. By following these tips, you can help keep your peach trees healthy and productive.

```
response = model.generate_content(["Prepare 3 questions for the given image", img],  
stream=True)
```

```
response.resolve(  
to_markdown(response.text))
```

What is the cause of this disease?

How does the disease affect the plant?

How can the disease be prevented or treated?

```
import textwrap
```

```
import os
```

```
import PIL.Image
```

```
from IPython.display import Markdown
```

```
import google.generativeai as genai

from google.colab import files

from google.colab import drive

# Used to securely store your API key

os.environ['GOOGLE_API_KEY'] = 'AIzaSyAc7Ii4wHf_whau2q--rgjfdht8-I5xhSY'

GOOGLE_API_KEY = os.getenv('GOOGLE_API_KEY')

genai.configure(api_key=GOOGLE_API_KEY)

for m in genai.list_models():

    if 'generateContent' in m.supported_generation_methods:

        #print(m.name)

        def to_markdown(text):

            text = text.replace('•', ' *')

            return Markdown(textwrap.indent(text, '> ', predicate=lambda _: True))

def generate_questions_from_image():

    # Upload image

    uploaded = files.upload()

    for fn in uploaded.keys():

        print("User uploaded file \"{name}\" with length {length} bytes".format(

            name=fn, length=len(uploaded[fn])))

    # Load the uploaded image

    img = PIL.Image.open(fn)

    # Display the image

    display(img)

# Initialize generative model
```



```
model = genai.GenerativeModel('gemini-pro-vision')

# Generate questions based on the image

response = model.generate_content(["Prepare 3 questions for the given image", img],
stream=True)

response.resolve()

# Display the generated questions

display(to_markdown(response.text))

generate_questions_from_image()
```

3.5. Model Architecture:

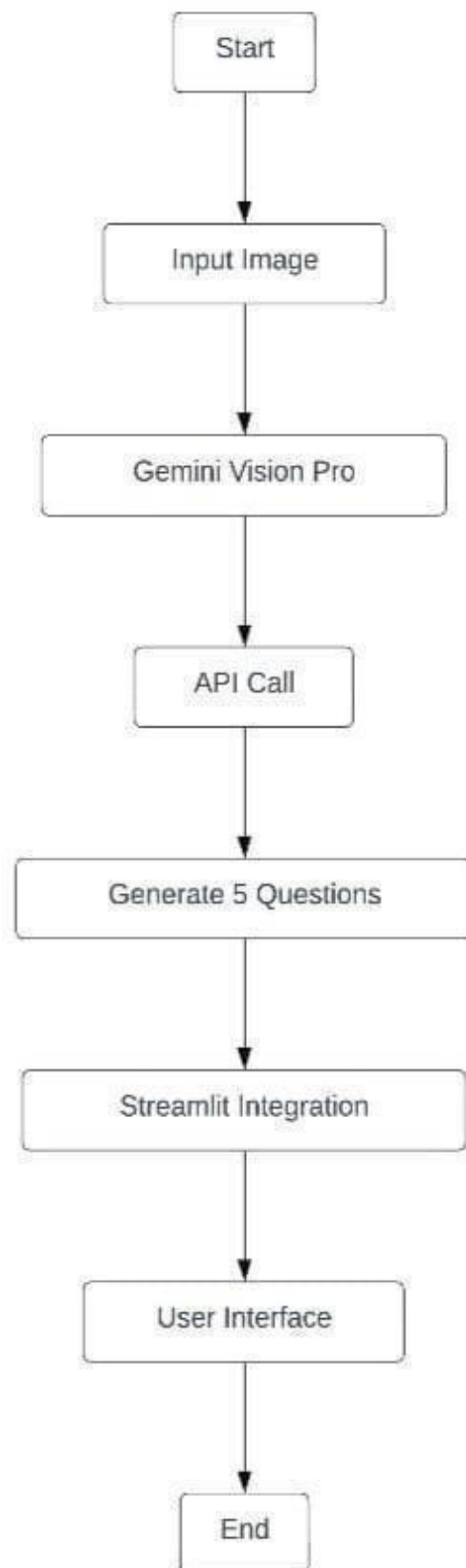


Fig 3.5.1: Model Architecture:

3.6 SYSTEM REQUIREMENT:

Software Requirements:

Python Environment:

Installation of Python (preferably version 3. x) to run the codebase and dependencies.

Deep Learning Framework: TensorFlow or PyTorch: Frameworks for implementing and training deep learning models. Ensure compatibility with the chosen pre-trained model (BERT) and Gemini API.

Gemini API:

Installation and integration of the Gemini API for accessing pre-trained language models like BERT. Follow the documentation for setup instructions and API usage.

Image Processing Libraries:

Libraries such as OpenCV or PIL (Python Imaging Library) for image preprocessing tasks like resizing, augmentation, and feature extraction from remote sensing images.

Text Processing Libraries:

Tokenization libraries like Hugging Face's Transformers or TensorFlow Text for processing textual data, including tokenization and embedding generation.

Deployment Framework (Optional):

TensorFlow Serving or ONNX Runtime: Frameworks for deploying and serving trained models in production environments. This may be necessary if deploying the model for real-world applications.

Hardware Requirements:

GPU or TPU:

High-performance hardware accelerators like Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs) are recommended for training deep learning models efficiently, especially for large-scale datasets and complex architectures like BERT.

Memory (RAM):

Sufficient RAM capacity to accommodate the dataset and model parameters during training and inference processes. The memory requirement may vary based on the dataset size and model complexity.

Storage Space:

Adequate storage space to store the dataset, pre-trained model weights, and intermediate outputs generated during training and inference stages.

Processor (CPU):

A multi-core CPU is required for general-purpose computation tasks, including data preprocessing, model setup, and post-processing operations.

Network Connectivity:

Stable internet connectivity may be required for accessing pre-trained models via APIs (such as Gemini API) and for downloading additional dependencies or updates during setup.

Cooling System (Optional):

If using hardware accelerators like GPUs for intensive training tasks, ensure proper cooling mechanisms to prevent overheating and hardware damage.

By fulfilling these software and hardware requirements, the system can effectively implement visual question generation from remote sensing images using the Gemini API with BERT-based models, enabling efficient training and deployment processes.

CHAPTER 4

RESULTS AND DISCUSSION

CHAPTER 4

RESULTS AND DISCUSSION

4.1 OUTPUT SCREENS:

Initially, when we run the required commands the user interface is opened.

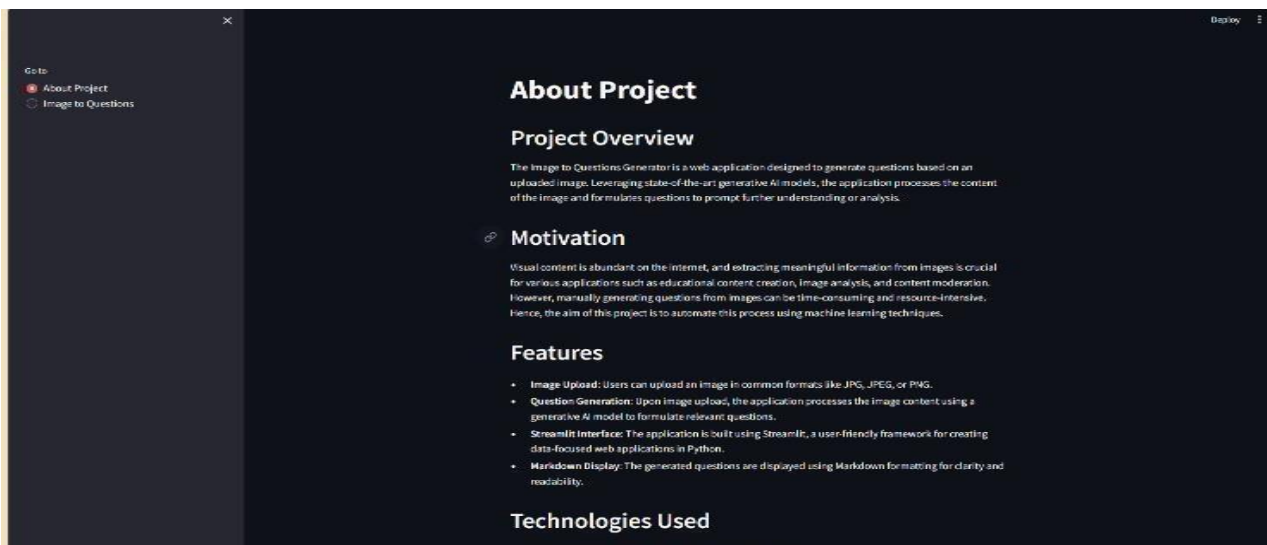


Fig 4.1.1: User Interface

Once the interface is opened then we need to upload an image in the provided userinterface.

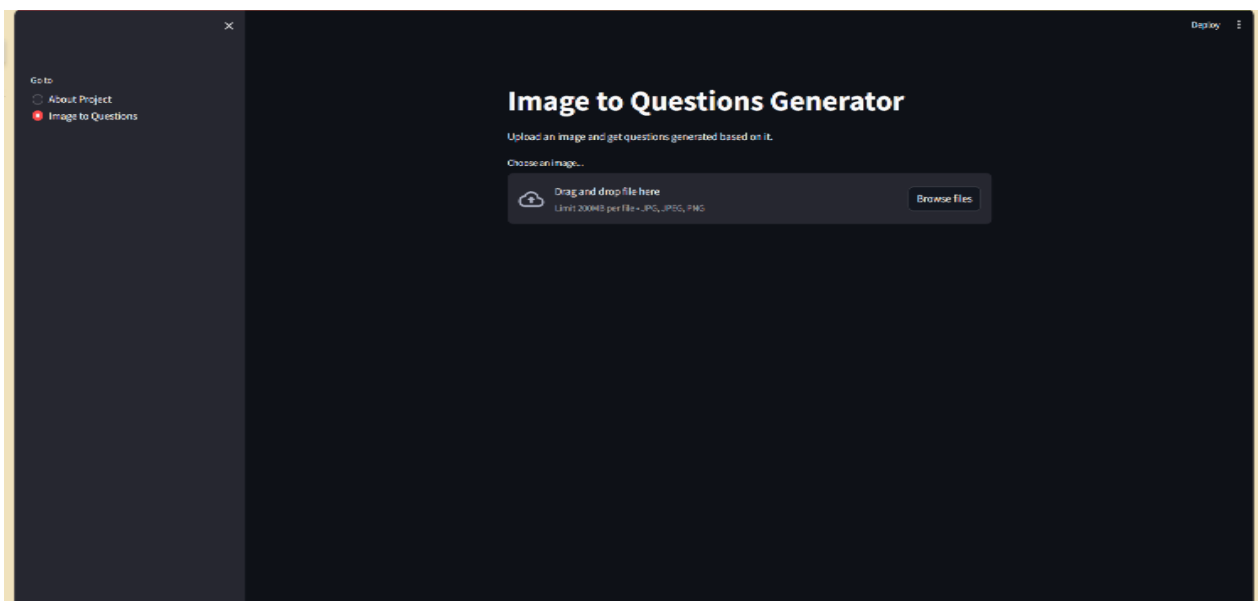


Fig 4.1.2: Uploading Of Image

Later on, using pre-trained models the features required are extracted from the image uploaded and then the required output is displayed on the screen.

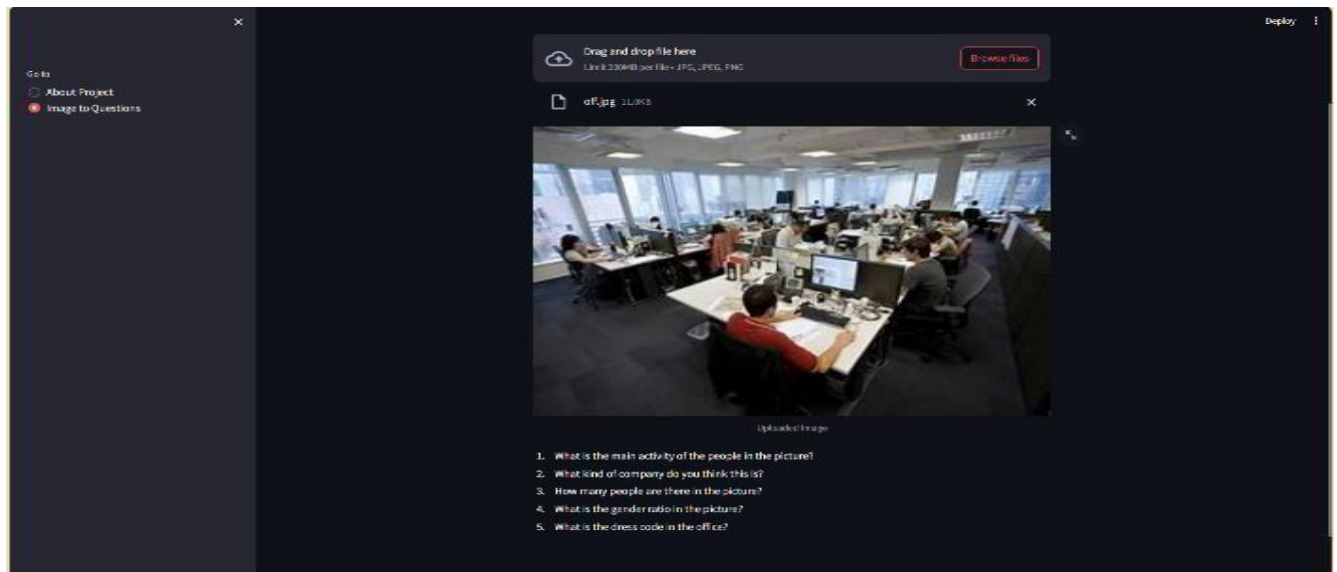


Fig 4.1.3: Extract Features and Generate Questions (eg1).

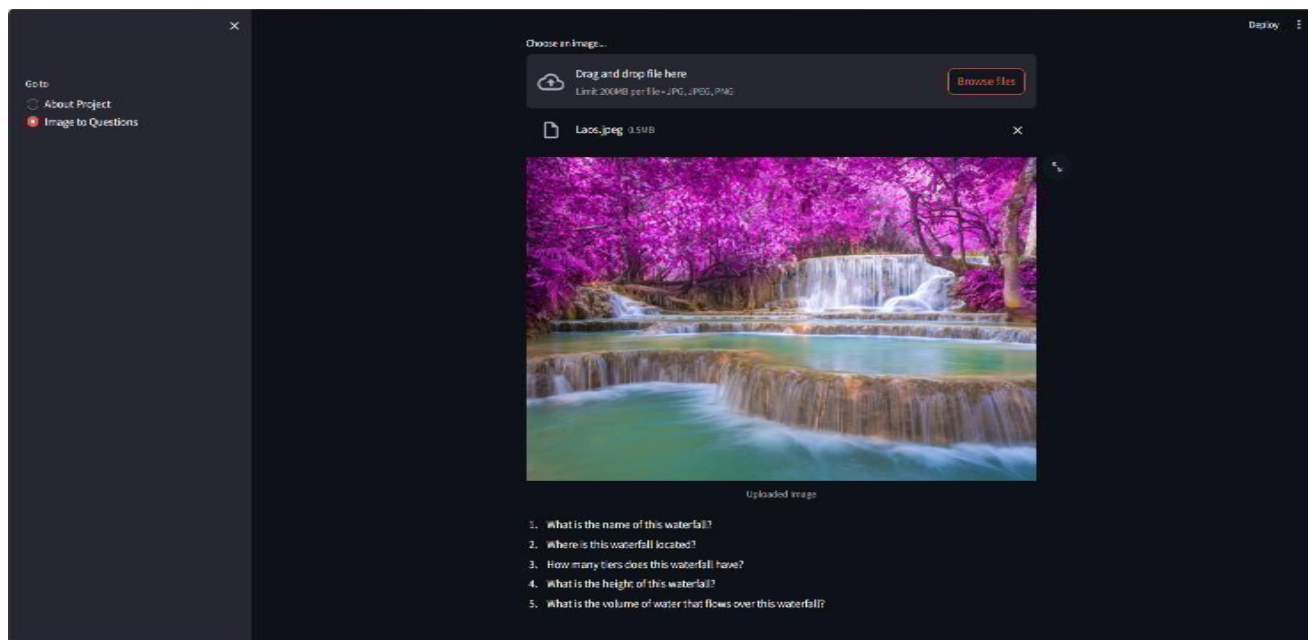


Fig 4.1.4: Extract Features and Generate Questions(eg2)

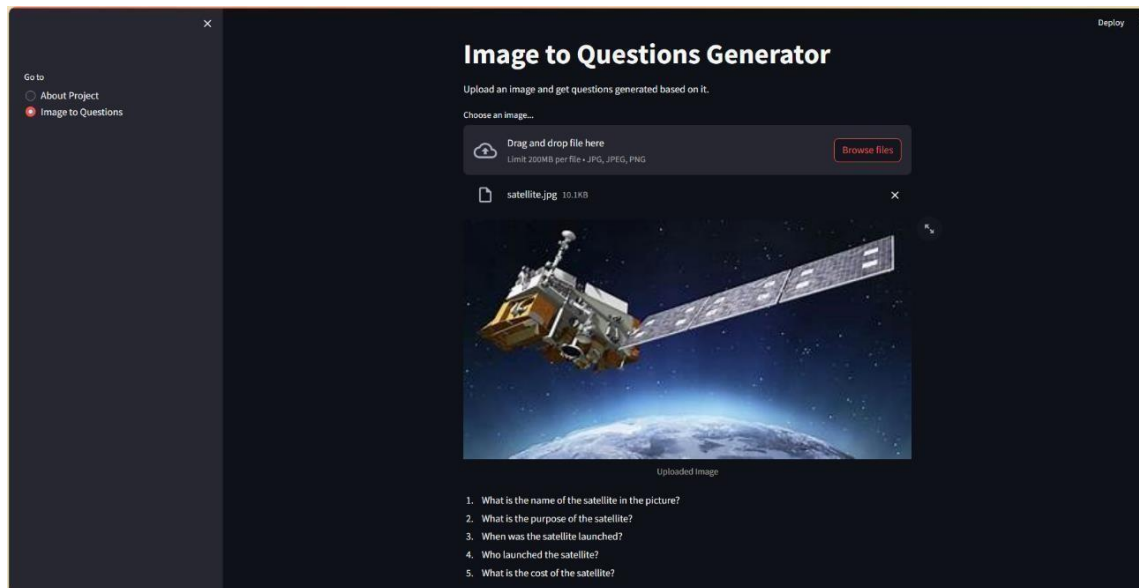


Fig 4.1.5: Extract Features and Generate Questions(eg3)

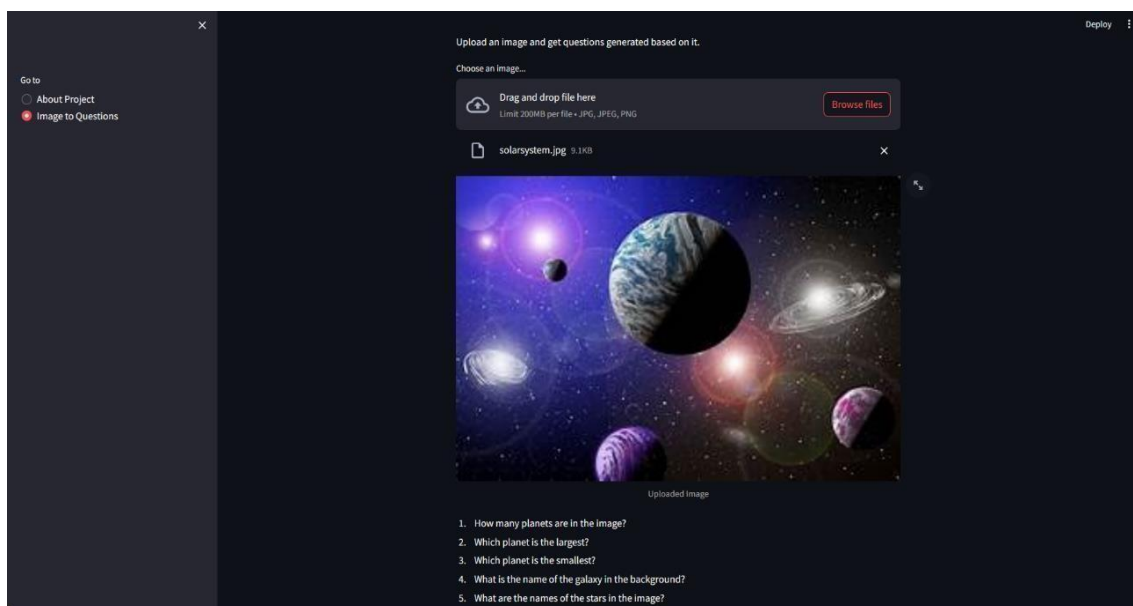


Fig 4.1.6: Extract Features and Generate Questions(eg4)

4.2 PERFORMANCE METRICS:

Question Relevance: Evaluate the relevance of generated questions by comparing them with ground truth annotations or human judgments.

Semantic Coherence: Assess the coherence and logical consistency of generated questions using semantic similarity metrics or linguistic analysis techniques.

Diversity: Measure the diversity of generated questions to ensure a wide coverage of semantic concepts and topics present in the remote sensing images.

CHAPTER 5

CONCLUSION

CHAPTER 5

CONCLUSION

5.1 CONCLUSION AND FUTURE ENHANCEMENT:

- The Image to Questions Generator showcases the practicality of machine learning, simplifying question generation from visual content for various purposes. By leveraging the Gemini API, it enhances accessibility and analysis efficiency for remote sensing images. Ongoing research hints at novel applications in environmental fields, underlining the significance of interdisciplinary collaboration in tackling environmental challenges.
- Improved Question Quality: Refinement of the generative model or integration of feedback mechanisms to enhance the quality of generated questions.
- Support for Different Languages: Extension of the application to support question generation in multiple languages.
- Customization Options: Addition of features allowing users to customize the types or number of questions generated.

REFERENCES

REFERENCES

- [1] Xiong, Z., Zhang, F., Wang, Y., Shi, Y., & Zhu, X. X. (2022). Earth nets: Empowering AI in earth observation. arXiv preprint arXiv:2210.04936.
- [2] Huang, W., Wang, Q., & Li, X. (2020). Denoising-based multiscale feature fusion for remote sensing image captioning. *Geoscience and Remote Sensing Letters*, 18(3), 436-440.
- [3] Fu, K., Li, Y., Zhang, W., Yu, H., & Sun, X. (2020). Boosting memory with a persistent memory mechanism for remote sensing image captioning. *Remote Sensing*, 12(11), 1874.
- [4] Zhang, Z., Diao, W., Zhang, W., Yan, M., Gao, X., & Sun, X. (2019). LAM: Remote sensing image captioning with the label-attention mechanism. *Remote Sensing*, 11(20), 2349.
- [5] Shi, Z., & Zou, Z. (2017). Can a machine generate humanlike language descriptions for a remote-sensing image? *Transactions on Geoscience and Remote Sensing*, 55(6), 3623-3634.
- [6] Zhang, Z., Diao, W., Zhang, W., Yan, M., Gao, X., & Sun, X. (2019). LAM: Remote sensing image captioning with the label-attention mechanism. *Remote Sensing*, 11(20), 2349.
- [7] Uppal, S., Madan, A., Bhagat, S., Yu, Y., & Shah, R. R. (2021, March). C3VQG: Category consistent cyclic visual question generation. In *Proceedings of the 2nd ACM International Conference on Multimedia in Asia* (pp. 1-7).
- [8] Fan, Z., Wei, Z., Li, P., Lan, Y., & Huang, X. (2018, July).
- [9] A Question Type Driven Framework to Diversify Visual Question Generation. In *IJCAI* (pp. 4048-4054).
- [10] Abdullah, T., Bazi, Y., Al Rahhal, M. M., Mohali, M. L., Rangarajan, L., & Zuhair, M. (2020). Texters: Deep bidirectional triplet network for matching text to remote sensing images *Remote Sensing*, 12(3), 405.

GitHub Link

<https://github.com/pravalikabommagani/Majorproject>



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12

Issue: III

Month of publication: March 2024

DOI:

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

Visual Question Generation from Remote Sensing Images Using Gemini API

M. Kamala¹, B. Pravalika², Y. Laxmi Narayana³, P. Arya Patel⁴

UG Student, Department of Computer Science & Engineering, CMR College of Engineering & Technology, Hyderabad, India

Abstract: Visual Question Generation Extracting Information from Remote Sensing Images Remote Sensing Images plays a vital role in understanding and extracting information from aerial and satellite images. Utilizing Bidirectional Encoder Representation from Transformers (BERT) for extracting valuable insights from remote sensing images. Gemini Application Programming Interface(API), and Convolution Neural Networks (CNNs) are used. First, The proposed methodology employs CNN to extract high-level features from remote sensing images, capturing spatial data and generating questions. Similarly, the Gemini Application Programming Interface(API) integrates contextual understanding into the question-generation process by providing relevant environmental data. Lastly, BERT functions as a language model in which employees enhance and refine the generated questions by taking into account both the syntax and semantics. Hence, by combining all these techniques we are capable of generating required relevant questions from remote sensing images in an enhanced and efficient way.

Index Terms: Visual Question Generation, CNN, Gemini API, Remote Sensing Images, Natural Language Processing, Deep Learning, BERT.

I. INTRODUCTION

The burgeoning field of the Visual Address Era (VQG) points to bridging the crevice between visual question substance and normal dialect by naturally producing questions almost pictures. This paper proposes a novel approach joining Convolutional Neural Systems (CNNs), the Gemini API, and Utilizing Bidirectional Encoder Representation from Transformers for creating questions from further detecting pictures. CNNs investigate visual highlights, Gemini API enhances relevant understanding, and BERT refines semantics. This strategy upgrades address pertinence and specificity, engaging clients to extricate experiences from endless visual information stores.

Xiong, Z., Zhang, F., Wang, Y., Shi, Y., & Zhu, X. X [1] Soil perception, pointing at observing the condition of planet Soil utilizing further detecting information, is basic for progressing our everyday lives and living environment. With a developing number of satellites in a circle, an expanding number of datasets with assorted sensors and inquiries about spaces are being distributed to encourage the exploration of the inaccessible detecting community Huang, W., Wang, Q., & Li, X [2] With the benefits of profound learning innovation, creating captions for inaccessible detecting pictures has gotten to be achievable, and awesome advances have been made in this field over the later long time. Be that as it may, a large-scale variety of further detecting pictures, which would lead to mistakes or exclusions in including extraction, still limits the assist advancement of caption quality. Fu, K., Li, Y., Zhang, W., Yu, H., & Sun, X [3] The encoder-decoder system has been broadly utilized within the further detecting picture captioning task. When we have to extricate inaccessible detecting pictures containing specific characteristics from the portrayed sentences for investigation, wealthy sentences can make strides the ultimate extraction comes about. Be that as it may, the Long Short-Term Memory (LSTM) organize utilized in decoders still loses a few data within the picture over time when the produced caption is long instrument is predominant in this errand but still has a few disadvantage s. The customary consideration instrument as it were employments visual data that is almost inaccessible in detecting pictures.

Zhang, Z., Diao, W., Zhang, W., Yan, M., Gao, X., & Sun, X [4] Critical advances have been made in further detecting picture captioning by encoder-decoder systems. The customary consideration without considering utilizing the name data to direct the calculation of consideration covers.

II. LITERATURE SURVEY

Shi, Z., & Zou, Z. [5] This paper examines a charming address within the inaccessible detecting field: "Can a machine produce humanlike dialect portrayals for a further detecting image The modified delineation of a farther-detecting picture (to be particular, blocked off recognizing picture captioning) may be a basic but rarely considered errand for fabricated bits of knowledge? It is more challenging as the depiction must not as it were capture the ground components of diverse scales, but moreover express their traits as well as how these components connected.

Zhang, Z., Diao, W., Zhang, W., Yan, M., Gao, X., & Sun, X [6] Critical advance has been made in inaccessible detecting picture captioning by encoder-decoder systems. The customary consideration instrument is predominant in this errand but still has a few downsides. The customary consideration component as it were employments visual data almost the further detecting pictures without considering utilizing the name data to direct the calculation of consideration covers. To this conclusion, a novel consideration component, specifically Label-Attention Instrument (LAM), is proposed in this paper. LAM moreover utilizes the name data of high-resolution inaccessible detecting pictures to create characteristic sentences to depict the given pictures.

Uppal, S., Madan, A., Bhagat, S., Yu, Y., & Shah, R. R [7] Visual Address Era (VQG) is the errand of creating common questions based on a picture. Fan, Z., Wei, Z., Li, P., Lan, Y., & Huang, X. [8] In our system, an address is developed in two steps. To begin with, an address sort is examined to decide what kind of data is asked. Moment, the substance of the address is produced conditioning on the tested address sort and the visual data of the picture. Prevalent strategies in the past have to investigate image- to-sequence designs prepared with the most extreme probability which have illustrated significant produced questions given a picture and its related ground-truth reply. VQG gets to be more challenging if the picture contains wealthy relevant data portraying its distinctive semantic categories

Abdullah, T., Bazi, Y., Al Rahhal, M. M., Mohali, M. L., Rangarajan, L., & Zuhair, M. [9] The unfaltering availability of inaccessible detecting information, especially tall determination pictures, has vivified exceptional inquiries about yields within the further sensing community. Two of the foremost dynamic points in this respect allude to picture classification and recovery [1,2,3,4,5]. Picture classification points to relegate scene pictures to a discrete set of arrival use/land cover classes depending on the picture substance [6,7,8,9,10]. As of late, with quickly extended further detecting innovations, both the amount and qual ity offurtherdetecting information have been expanded.

III. METHODOLOGY

A. Technologies used

Here, we have used various technological stuff usage of a pre-trained model Gemini API, CNN, and Bert to generate accurate questions from remote sensing images.

B. Gemini API

Gemini offers to get to a run of expansive dialect models, each with its qualities in the address era. You'll be able to select the LLM that best suits your particular needs, like producing open-ended questions, genuine requests, or imaginative prompts. The API acknowledges different input designs, counting content sections, pictures, code bits, or indeed conversational transcripts. This permits you to create questions based on diverse sorts of data, making it flexible for different applications. The API consistently coordinates with other instruments and stages, permitting you to consolidate address eras into different workflows.

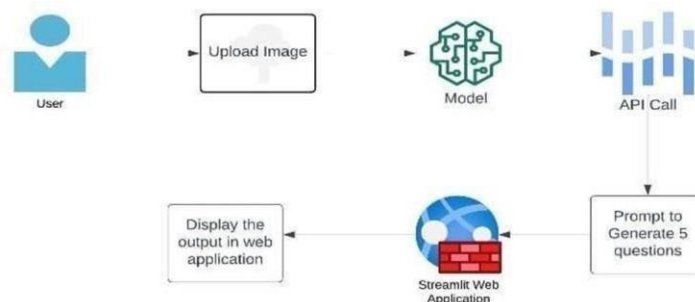


Fig 1: Proposed Methodology Architecture

C. BERT

- 1) *Gets the Picture Information:* After Gemini API bridges the hole, BERT takes the reins, accepting the complicated pointsof interest extricated from the inaccessible detecting picture.
- 2) *Interprets Visual Designs:* Like a prepared criminologist, BERT analyzes the picture, leveraging its tremendous informationon visual components and their connections. This incorporates arrival cover sorts, surfaces, shapes, and spatial courses of action.
- 3) *Makes Curious Questions:* Based on its investigation, BERT changes the visual information into characteristic dialect questions. These questions act as prompts, welcoming a more profound investigation of the image's substance and potential suggestions.

- 4) *Convolutional Neural Networks*: It Organizes, another sort of neural arrangement design. CNNs exceed expectations at recognizing designs in pictures, making them well-suited for assignments like picture classification and protest location. In this case, CNN likely analyzes the transferred picture to extricate visual highlights that advise the address era preparation.
- 5) *Input Picture*: The client transfers a picture to the framework.
- 6) *Gemini Vision Professional API Call*: The picture is sent to the Gemini Vision Professional API. The API analyzes the picture and produces five questions almost it.
- 7) *Streamlit Integration*: The created questions are shown to the client in a web application.
- 8) *Show the yield in the web application*: The point-by-point depiction of the picture is at that point shown to the client within the web application.

The Gemini API engages designers with a vigorous set of functionalities to consistently coordinate cryptocurrency exchange and account administration into their applications. Advertising both REST and WebSocket APIs, clients pick up get to real-time showcase information, arrange arrangements, and account administration highlights. With secure verification through API keys, engineers can certainly associate with the trade, guaranteeing information judgment and client security. Gemini's comprehensive documentation and back framework advance improve the improvement involvement, empowering the creation of modern exchanging calculations, portfolio administration apparatuses, and showcase examination stages.

D. Implementation of Block Diagram

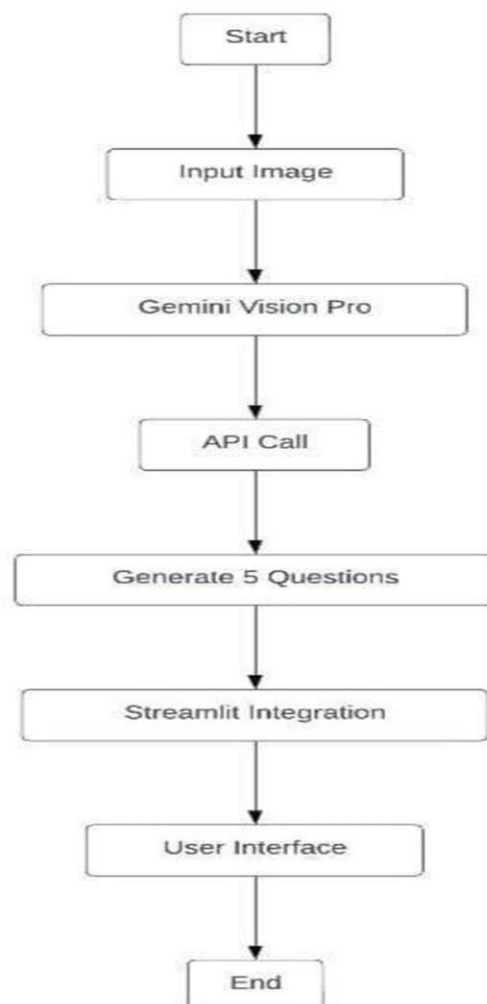


Fig 2: Execution Flow of Proposed Solution

IV. RESULTS AND DISCUSSION

A. Figures

Initially, when we run the required commands the user interface Is opened.

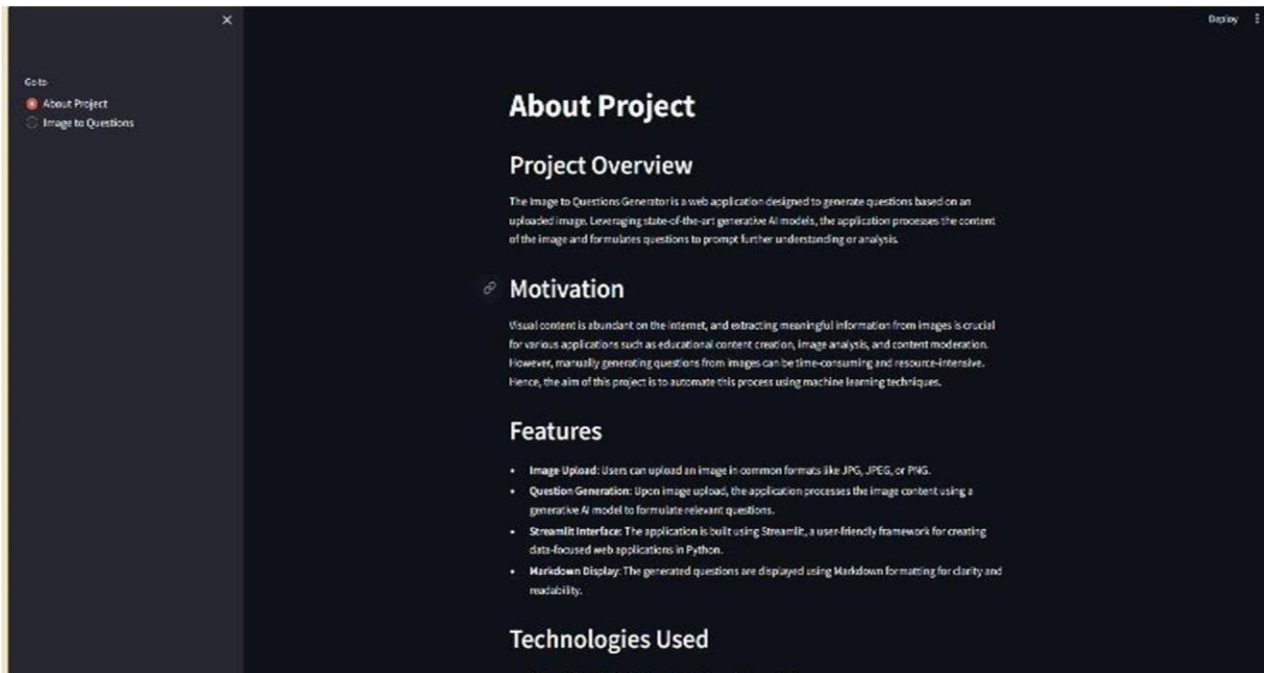


Fig 3: User Interface

Once the interface is opened then we need to upload an image in the provided user interface.

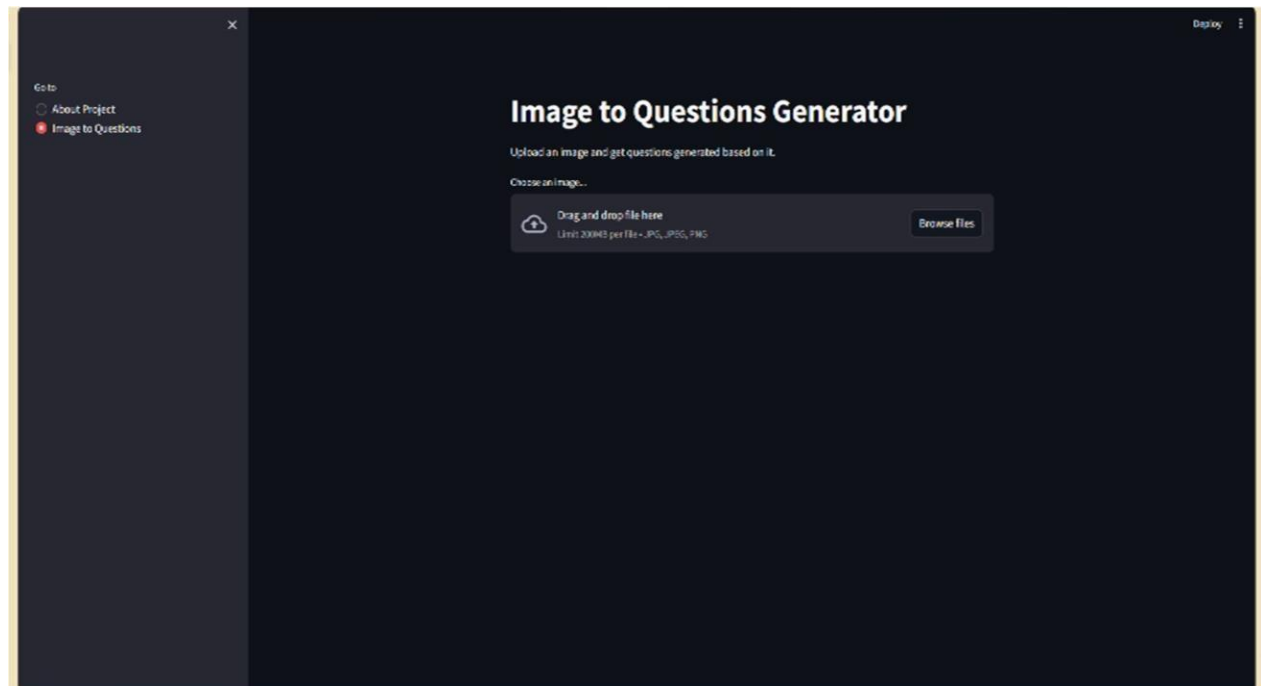


Fig 4: Uploading Of Image

Later on, using pre-trained models the features required are extracted from the image uploaded and then the required output is displayed on the screen

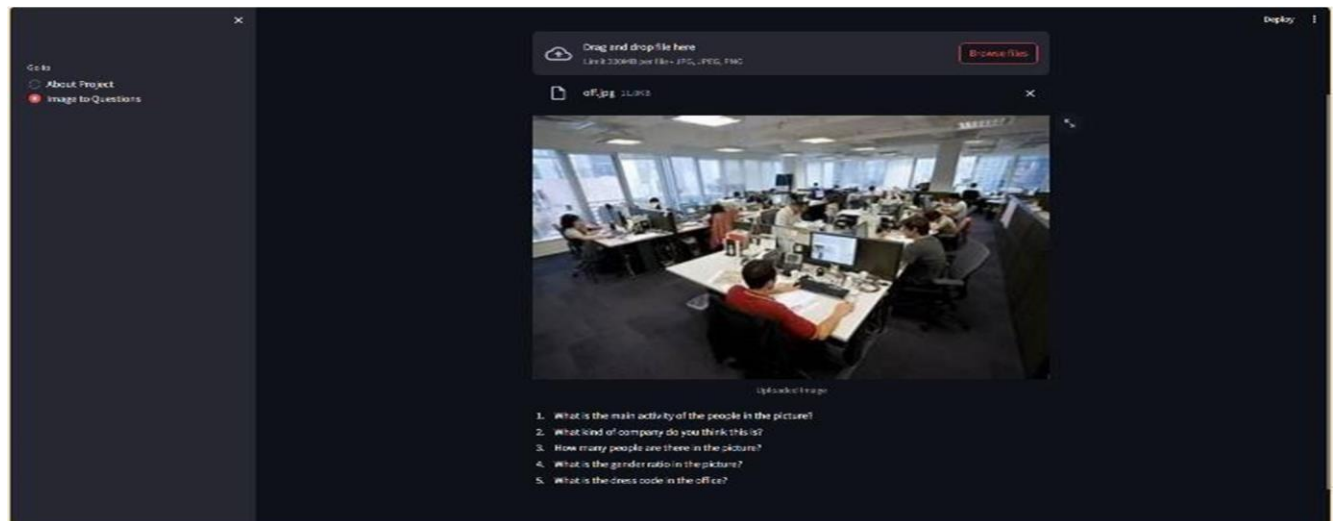


Fig 5: Extract Features and Generate Questions (eg1).

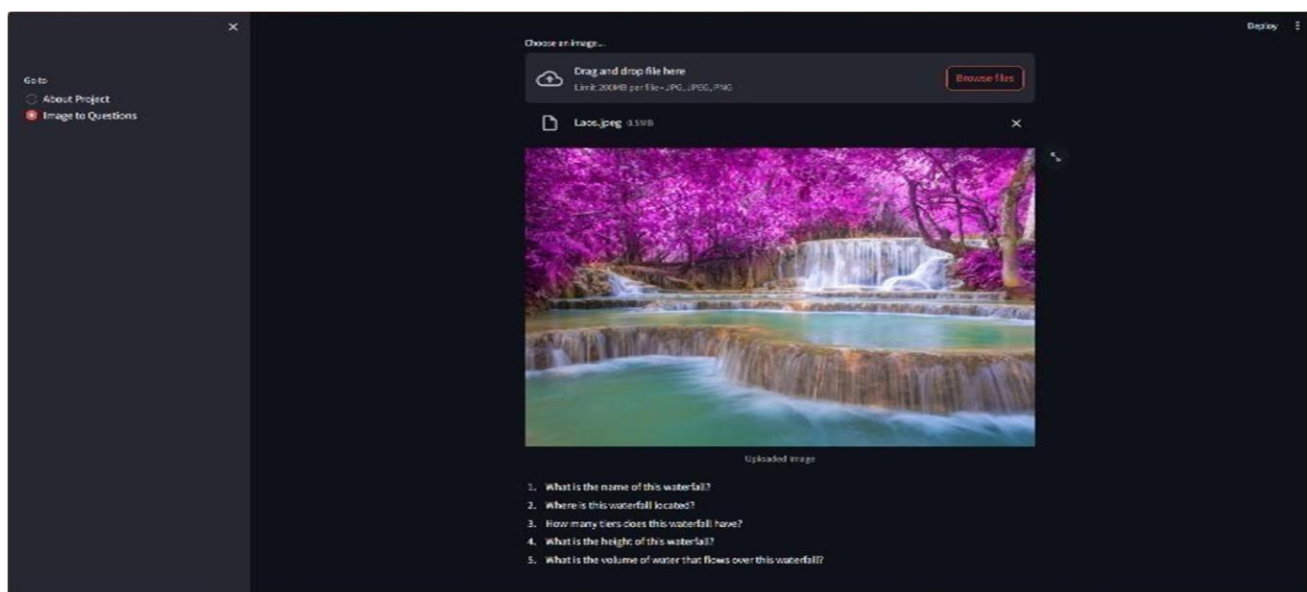


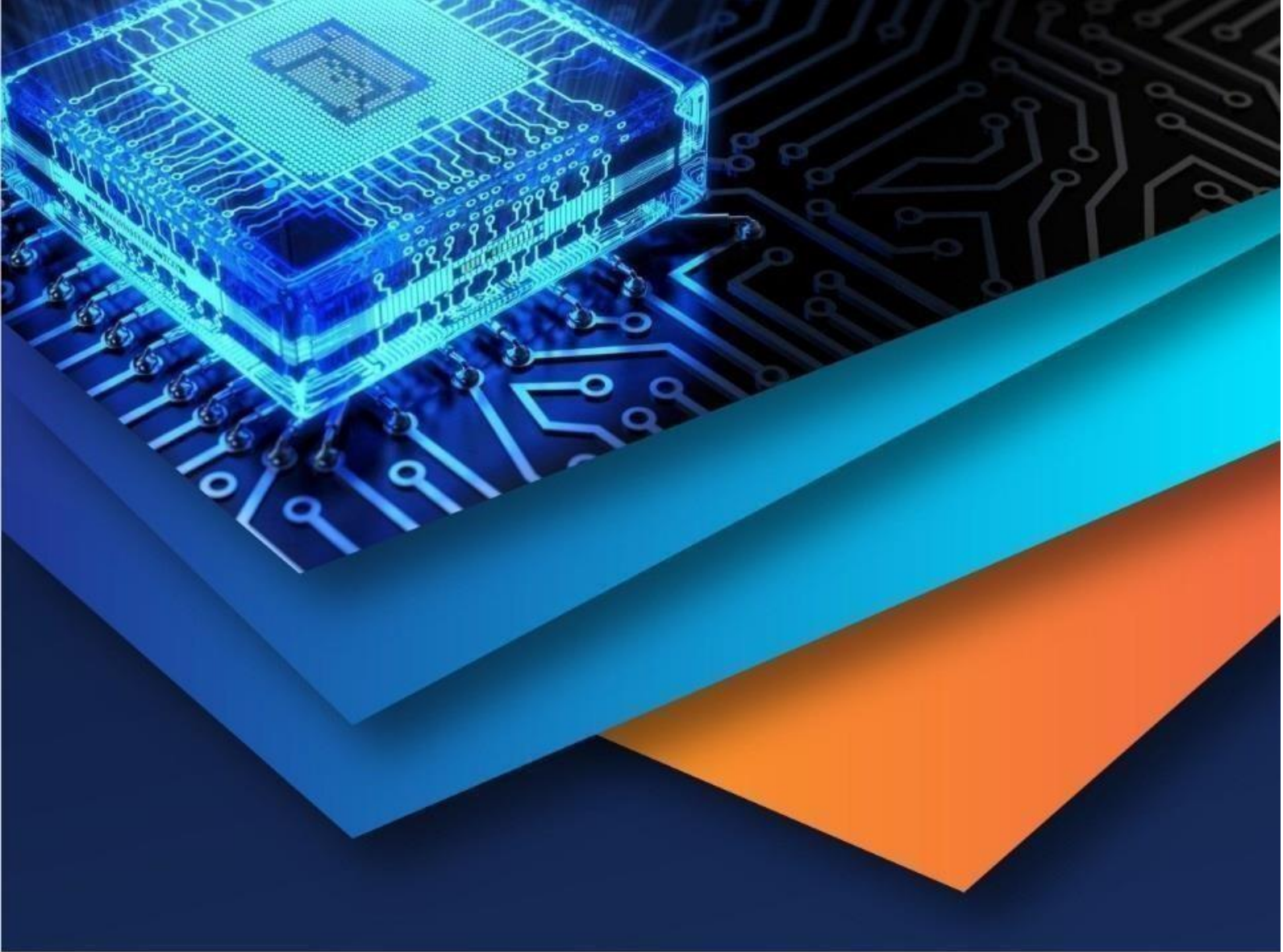
Fig 6: Extract Features and Generate Questions(eg2).

V. CONCLUSION

Visual code time is a new era of visual interaction that can leverage advances in machine learning in unprecedented ways. For example, it can be used to detect complexity in images making it possible to draw queries directly from the visual features. This approach simplifies speech creation and incorporates advanced techniques in computer vision and natural language processing to provide advanced interactivity in machine learning algorithms by extracting visual cues and by the extraction of meaningful information. Potential impacts help businesses in market research from educational seminars to image-based questionnaires or content analysis tools among others including how interesting cross-industry collaborations and partnerships (partnerships) have developed development through the integration of computer vision and natural language processing The development further highlights the potential of machine learning to modify our perception, interpretation, and communication of visual code time as presented here.

REFERENCES

- [1] Xiong, Z., Zhang, F., Wang, Y., Shi, Y., & Zhu, X. X. (2022). Earth nets: Empowering AI in earth observation. arid preprintarXiv:2210.04936.
- [2] Huang, W., Wang, Q., & Li, X. (2020). Denoising-based multiscale feature fusion for remote sensing image captioning. *Geoscience and Remote Sensing Letters*, 18(3), 436-440.
- [3] Fu, K., Li, Y., Zhang, W., Yu, H., & Sun, X. (2020). Boosting memory with a persistent memory mechanism for remote sensing image captioning. *Remote Sensing*, 12(11), 1874.
- [4] Zhang, Z., Diao, W., Zhang, W., Yan, M., Gao, X., & Sun, X. (2019). LAM: Remote sensing image captioning with the label-attention mechanism. *Remote Sensing*, 11(20), 2349.
- [5] Shi, Z., & Zou, Z. (2017). Can a machine generate humanlike language descriptions for a remote-sensing image? *IEEE Transactions on Geoscience and Remote Sensing*, -55(6), 3623-3634.
- [6] Zhang, Z., Diao, W., Zhang, W., Yan, M., Gao, X., & Sun, X. (2019). LAM: Remote sensing image captioning with the label-attention mechanism. *Remote Sensing*, 11(20), 2349.
- [7] Uppal, S., Madan, A., Bhagat, S., Yu, Y., & Shah, R. R. (2021, March). C3VQG: Category consistent cyclic visual question generation. In *Proceedings of the 2nd ACM International Conference on Multimedia in Asia* (pp. 1-7).
- [8] Fan, Z., Wei, Z., Li, P., Lan, Y., & Huang, X. (2018, July).
- [9] A Question Type Driven Framework to Diversify Visual Question Generation. In *IJCAI* (pp. 4048-4054).
- [10] Abdullah, T., Bazi, Y., Al Rahhal, M. M., Mohali, M. L., Rangarajan, L., & Zuhair, M. (2020). Texters: Deepbidirectional triplet network for matching text to remote sensing images *Remote Sensing*, 12(3), 405.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)



ISSN No. : 2321-9653

IJRASET

**International Journal for Research in Applied
Science & Engineering Technology**

IJRASET is indexed with Crossref for DOI-DOI : 10.22214

Website : www.ijraset.com, E-mail : ijraset@gmail.com

Certificate

*It is here by certified that the paper ID : IJRASET59537, entitled
Visual Question Generation From Remote Sensing Images Using Gemini API*

*by
M. Kamala*

*after review is found suitable and has been published in
Volume 12, Issue III, March 2024
in*

*International Journal for Research in Applied Science &
Engineering Technology
(International Peer Reviewed and Refereed Journal)
Good luck for your future endeavors*

By [Signature]

Editor in Chief, IJRASET

ISRA
JIF

ISRA Journal Impact
Factor: 7.429



INDEX COPERNICUS



THOMSON REUTERS
Researcher ID: 16-9681-2016



TOGETHER WE REACH THE GOAL
SJIF 7.429



ISSN No. : 2321-9653

IJRASET

**International Journal for Research in Applied
Science & Engineering Technology**

IJRASET is indexed with Crossref for DOI-DOI : 10.22214

Website : www.ijraset.com, E-mail : ijraset@gmail.com



ISRA Journal Impact
Factor: 7.429



45.98
INDEX COPERNICUS



THOMSON REUTERS
Researcher ID: N-9581-2016



TOGETHER WE REACH THE GOAL
SJIF 7.429

Certificate

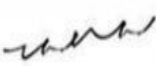
It is here by certified that the paper ID : IJRASET59537, entitled
Visual Question Generation From Remote Sensing Images Using Gemini API
by
B. Pravalika

after review is found suitable and has been published in
Volume 12, Issue III, March 2024
in

*International Journal for Research in Applied Science &
Engineering Technology*

(International Peer Reviewed and Refereed Journal)

Good luck for your future endeavors

By 

Editor in Chief, IJRASET



ISSN No. : 2321-9653

IJRASET

**International Journal for Research in Applied
Science & Engineering Technology**

IJRASET is indexed with Crossref for DOI-DOI : 10.22214

Website : www.ijraset.com, E-mail : ijraset@gmail.com



ISRA Journal Impact
Factor: 7.429



45.98
INDEX COPERNICUS



THOMSON REUTERS
Researcher ID: N-9681-2016



TOGETHER WE REACH THE GOAL
SJIF 7.429

Certificate

*It is here by certified that the paper ID : IJRASET59537, entitled
Visual Question Generation From Remote Sensing Images Using Gemini API*

*by
Y. Laxmi Narayana*

*after review is found suitable and has been published in
Volume 12, Issue III, March 2024
in*

*International Journal for Research in Applied Science &
Engineering Technology
(International Peer Reviewed and Refereed Journal)*

Good luck for your future endeavors

By [Signature]

Editor in Chief, IJRASET

