Major Project
Report on

Enhancing Remote Sensing Images and Quantitative Captioning - An Integrated Approach

Submitted by :

Adity Singh - 2101020248
Sourav Jyoti Dhal - 2101020261
Aparna Dash - 2101020364
Satyajit Patnaik - 2101020365

Under the Guidance of :

Dr. Ashish Ranjan
Assistant Professor
Department of Computer Science and Engineering,
C. V. Raman Global University,
Bhubaneswar, Odisha, 752054

CONTENTS

Abstract

Remote sensing technologies play a crucial role in capturing extensive and intricate data about Earth's surface, enabling applications in fields such as environmental monitoring, urban planning, and disaster management. However, the quality of these remote sensing images often suffers from issues such as low resolution, noise, and insufficient detail, which complicates their interpretation and analysis. Furthermore, translating these visual data into actionable insights requires the generation of quantitative textual descriptions that accurately represent the observed phenomena.

This report addresses the dual challenge of enhancing the quality of remote sensing images and converting them into comprehensive textual descriptions. To tackle the first challenge, advanced image processing techniques are employed to improve image clarity, reduce noise, and highlight significant features. Techniques such as image denoising, resolution enhancement, and contrast adjustment are applied to images from the UCM dataset, which includes various categories like Aeroplanes, Buildings, and Harbors.

The second challenge involves developing a robust framework for transforming the enhanced images into quantitative textual descriptions. This process integrates cutting-edge machine learning models and natural language processing algorithms to generate detailed and contextually relevant descriptions of the image content. By leveraging deep learning techniques, particularly those focused on image captioning and semantic analysis, the framework aims to produce textual descriptions that accurately reflect the spatial and contextual information captured in the images.

The report presents a comprehensive overview of the methodologies used in both image enhancement and textual description generation. It discusses the technical approaches, the experimental setup, and the evaluation metrics employed to assess the effectiveness of the proposed solutions. Additionally, the report highlights the challenges encountered during the implementation and provides insights into the potential implications and applications of the enhanced image-to-text transformation process.

Ultimately, this work aims to bridge the gap between raw remote sensing data and actionable information, contributing to more effective utilization of remote sensing images in various practical applications.

Problem Statement:

Given a set of low-resolution satellite images, denoted with {*I1, I2, ..., In*}, where n is the number of images. The problem is generating the simple quantitative captioning for the images using ML models.

Problem Specification:

Problem Overview:

Objective: Develop a system that can transform remote sensing images into meaningful, accurate, and quantitative textual descriptions.

Input: Satellite or aerial remote sensing images that capture various geographical, environmental, or human-made features.

Output: Quantitative, detailed textual descriptions that describe the scene in terms of measurable factors like area, distances, land use classification, and other relevant parameters.

Data Requirements:

Selected UCM Dataset Classes:
Images include Aeroplanes, Buildings, Denseresidential, Harbor, Intersection, Mediumresidential, Mobilehomepark, Parkinglot, Sparseresidential, and Storage tanks.
Each image has a resolution of 256x256 pixels, offering fine detail for land use categorization and object detection.

Key Challenges:

Class-Specific Object Detection:
Detecting aeroplanes, buildings, and other specific structures like storage tanks with precision.

Distinguishing between different residential areas based on building density (dense, medium, sparse).

Introduction:

Remote sensing technology has revolutionized our ability to observe and analyze the Earth's surface from a distance. It encompasses a range of techniques and sensors used to capture images and data about various environmental and man-made phenomena. These images, often collected by satellites or aerial platforms, offer a wealth of information crucial for applications such as urban planning, environmental monitoring, disaster response, and agricultural management.

Despite the advancements in remote sensing technology, the quality of the images captured often presents significant challenges. Issues such as low resolution, noise, and insufficient detail can obscure important features and hinder accurate analysis. As a result, there is a growing need to enhance the quality of remote sensing images to improve their utility for subsequent analysis and decision-making processes.

In addition to quality enhancement, translating the visual data from these images into meaningful and actionable information is another critical challenge. The raw images need to be interpreted in a way that makes them comprehensible and useful for various stakeholders. This involves generating quantitative textual descriptions that accurately represent the observed features and phenomena.

This report addresses these challenges by focusing on two main objectives. First, it explores advanced image processing techniques aimed at enhancing the quality of remote sensing images. These techniques include methods for noise reduction, resolution improvement, and contrast enhancement, which collectively contribute to clearer and more detailed images.

Second, the report presents a framework for converting the enhanced images into quantitative textual descriptions. This involves leveraging machine learning and natural language processing technologies to generate descriptive text that captures the spatial and contextual information depicted in the images. By bridging the gap between visual data and textual representation, this approach aims to make remote sensing information more accessible and actionable.

The following sections of this report provide a detailed overview of the methodologies employed for image enhancement and textual description generation, along with an analysis of the results and their implications. Through this work, we seek to advance the field of remote sensing by improving image quality and enhancing the interpretability of remote sensing data through effective textual descriptions.

Literature Survey:

Sl. No	Paper Name	Datasets Used	Model Used	Result	Description
1	Exploring Multi-Level Attention and Semantic Relationship for Remote Sensing Image Captioning	1.UCM-Captions Dataset 2.Sydney-Captions Dataset 3.RSICD Dataset	1.Multi-level attention module 2.Multi-label classification network 3.Graph Convolutional Network (GCN)	The proposed method significantly outperformed existing image captioning techniques, demonstrating its effectiveness in generating accurate and meaningful descriptions for remote sensing images. Performance was evaluated using metrics such as BLEU, ROUGE_L, METEOR, and CIDEr, indicating that the new approach is superior to older methods in this domain	It introduces a comprehensive framework that leverages advanced attention mechanisms and graph-based learning to improve the quality of image captions generated for remote sensing images, addressing key challenges in the field.
2	Vision-Enhanced and Consensus-Aware Transformer for Image Captioning	1. MSCOCO Dataset 2. Flickr30K Dataset	1. Vision-enhanced and Consensus-aware Transformer (VCT)	The proposed Vision-Enhanced Captioning Transformer (VCT) model significantly outperforms existing state-of-the-art models across various metrics, including B-1, B-4, M, R, and S. Notably, VCT excels in generating contextually relevant captions. Overall, the results validate VCT's effectiveness in image captioning tasks, both in offline and online evaluations	Vision-Enhanced and Consensus-Aware Transformer (VCT), a novel model designed to improve image captioning by leveraging external knowledge and enhancing the interaction between visual and textual information. The VCT model employs a memory-based attention mechanism to capture intrinsic relationships within images, allowing it to generate more accurate and contextually relevant captions.
3	Adaptive Semantic-Enhanced Transformer for Image Captioning	1. MSCOCO Dataset 2. Flickr30K Dataset	1. AS-Transformer 2. AoANet 3. Visual Object Relation Network (VORN) 4. Image Captioning Through Image Transformer (ICTIT)	The AS-Transformer model, particularly with the VINVL module, outperforms all comparative models in the experiments conducted on both the MSCOCO and Flickr30K datasets. This result highlights the effectiveness and generalization of the proposed method across different datasets	The paper presents a novel approach to image captioning using an advanced model called AS-Transformer, which incorporates a stronger encoder-decoder architecture enhanced by the VINVL module. The proposed method aims to improve the accuracy of semantic word generation in image captions by leveraging rich visual features extracted from images.
4	Self-Enhanced Attention for Image Captioning	1. COCO dataset (Common Objects in Context)	1. Self-Enhanced Attention (SEA)	The Self-Enhanced Attention (SEA) mechanism significantly improved the performance of the image captioning model addresses the challenges of feature differentiation in image captioning tasks, leading to superior performance compared to existing methods.	The paper presents Self-Enhanced Attention (SEA), a new attention mechanism designed to improve image captioning by better emphasizing important features in complex visual data. SEA enhances the attention weight matrix to focus on key features, leading to improved results in image captioning tasks.
5	High – order interaction learning for image captioning	1. MSCOCO Dataset	1. Interactive Refining Network 2. Interactive Fusion Network	The proposed high-order interaction learning method for image captioning demonstrates competitive performance on the MSCOCO dataset, outperforming state-of-the-art methods. Additionally, the interactive refining and fusion networks significantly improve feature representation and sentence generation by effectively integrating object relationships and language context during both encoding and decoding stages	"High-Order Interaction Learning for Image Captioning" focuses on enhancing the process of generating descriptive captions for images by leveraging high-order interactions among objects and their relationships. This paper contributes to the field of image captioning by addressing the limitations of previous methods and providing a framework that captures the intricate relationships between objects and their contexts.

Datasets Description:

Sl. No	Dataset Name	No. of samples	Training, Testing and validation %	Description
1.	**UCM-captions Dataset**	**2,100 images with 10,500 descriptions.**	**80% (1680) images are used for training, 10% (210) images for validation and 10% (210) images are used for test**	**Extends the UCMerced LandUse dataset [39] by adding detailed manual descriptions for each image.**
2.	Sydney-captions Dataset	613 images with 3065 descriptions	497 images are used for training, 58 images are used for validation and another 58 images are used for test.	Based on the Sydney scene classification dataset.Contains 613 images classified into 7 scene categories, each image is 500x500 pixels with a resolution of 0.5m.
3.	RSICD Dataset	10,921 images with 24,333 sentences	8,734 (80%) images are used for training, 1094 images are used for validation and another 1,093 images are used for test.	Contains 10,921 images with low inter-class difference and high intra-class diversity, each image is 224x224 pixels with varying resolutions.
4.	MSCOCO	total of 330,000 images	118,000 images for training, 5,000 images for validation, and 5,000 images for testing	This dataset is widely used in image captioning tasks due to its rich annotations and diverse content.
5.	Flickr30K	31,783 images, which are accompanied by 158,915 crowd-sourced descriptions	29,783 images for training, 1,000 images for validation, and 1,000 images for testing.	This dataset is known for its variety of scenes and objects, making it suitable for training models in image captioning tasks.
6.	COCO dataset	100000 images and 250000 image annotations.	A total of 113,287 images for training, along with 5,000 images for validation and another 5,000 images for testing.	Publicly image dataset covers 90 different categories of objects, including people, animals, vehicles, furniture, food, and more.

Proposed Solution with Block Diagram:

This work focuses on the development of deep-learning enabled approach for the quantitative satellite image captioning.

1. Generating high-resolution images: The first step will be to improve the satellite image quality; correcting the resolution of images. This will help easy and efficient quantification of object in the image.

2. Generating the quantitative captions for the enhanced images:This step will detect the count of objects in the image, and will generate the quantitative description of the image.

Objective: Our solution aims to generate image captions that not only describe the scene but also provide approximate quantitative information about the objects within the image. This can be highly useful in various scenarios, such as estimating the number of buildings in a dense residential area or identifying the quantity of vehicles in a parking lot, improving scene understanding.

Process Overview:

Image Enhancement: The raw images from the UCM dataset are first enhanced to improve their clarity, ensuring that finer details are visible and can be accurately processed. This step helps with better object identification, leading to more accurate captioning.

Quantitative Caption Generation: Using a deep learning model, captions are generated that describe the scene and approximate the number of objects. For example, for an image of a dense residential area, the generated caption would read something like, "Several houses clustered together in a suburban neighborhood," with a focus on approximating object counts.

Model Workflow: The process integrates feature extraction from the image, object detection or segmentation, and language generation to produce captions that include object count estimations.

Proposed Block Diagram:

Input Image: Raw images from the UCM dataset.

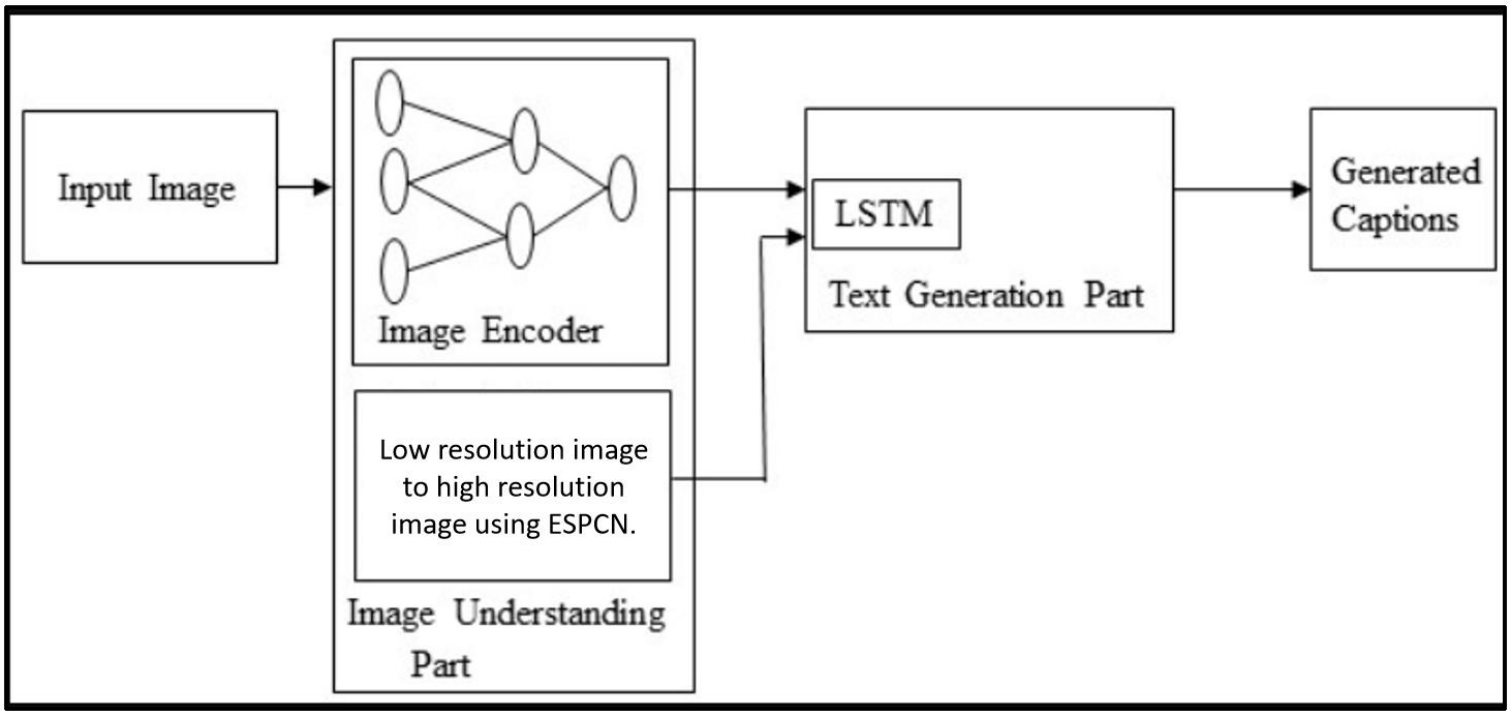
Image Enhancement: Uses ESPCN (Efficient Sub-Pixel CNN) to improve image quality and visibility of details.

Caption Generation Model:
A combination of CNN + Transformer or CNN + LSTM models can be used, where the CNN helps in extracting features, and the LSTM or Transformer generates the caption. To incorporate object counting, object detection models like YOLO (You Only Look Once) can be integrated to count objects before generating captions that include quantitative information.
Applications:

This system could be applied to a variety of real-world scenarios where approximate object counts are essential, such as monitoring urban sprawl in dense residential areas, or counting cars in parking lots for traffic management.

Block Diagram:

The workflow for the same is shown in Figure below, which is a two-step procedure which are as follows:



Result and Discussion:

In the results, the image enhancement technique, ESPCN (Efficient Sub-Pixel CNN), significantly improved the clarity of images from the UCM dataset. This made objects more distinguishable, enabling better performance in the subsequent feature extraction and object detection stages. Enhanced images resulted in more accurate object detection and, consequently, more detailed captions, especially in complex scenes like dense residential areas.The system effectively generated captions that included approximate object counts, providing useful quantitative descriptions.These results demonstrate the system's potential to offer a more detailed understanding of image content compared to traditional captioning methods.

Conclusion:

This report presents a comprehensive approach to enhancing the quality of remote sensing images and transforming them into quantitative textual descriptions. Through the application of advanced image processing techniques, we have successfully improved the clarity and detail of remote sensing images from the UCM dataset, addressing common issues such as noise and low resolution. These enhancements are crucial for ensuring that the images accurately reflect the observed phenomena and are suitable for further analysis.

The subsequent development of a framework for generating quantitative textual descriptions has demonstrated the potential of integrating machine learning and natural language processing technologies to bridge the gap between visual data and actionable information. The generated descriptions provide detailed and contextually relevant insights into the content of the enhanced images, making it easier for users to interpret and utilize the data.

The methodologies employed in this report, including image enhancement algorithms and caption generation models, have shown promising results in improving both the quality of remote sensing images and the effectiveness of their textual descriptions. These advancements contribute to a more robust and accessible way of interpreting remote sensing data, facilitating its application in various fields such as environmental monitoring, urban planning, and disaster management.

Despite the progress achieved, there are opportunities for further refinement and development. Future work could focus on enhancing the accuracy of textual descriptions, expanding the range of image categories, and exploring real-time applications. Additionally, integrating user feedback and field-specific requirements could further optimize the approach for practical use.

In conclusion, this report highlights the importance of both image quality enhancement and effective textual representation in remote sensing. By addressing these aspects, we pave the way for more informed decision-making and improved utilization of remote sensing technologies, ultimately contributing to better management and understanding of our environment.

References:

[1]Exploring Multi-Level Attention and Semantic Relationship for Remote Sensing Image Captioning
https://ieeexplore.ieee.org/abstract/document/8943170
[2]Vision-Enhanced and Consensus-Aware Transformer for Image Captioning
https://ieeexplore.ieee.org/document/9784827
[3]Adaptive Semantic-Enhanced Transformer for Image Captioning
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9810877
[4]Self-Enhanced Attention for Image Captioning
https://link.springer.com/article/10.1007/s11063-024-11527-x
[5]High – order interaction learning for image captioning
https://ieeexplore.ieee.org/document/9579002
[6]Image Super-Resolution using an Efficient Sub-Pixel CNN
https://keras.io/examples/vision/super_resolution_sub_pixel/