Faculty of Mathematics and Informatics

Change Detection in Satellite Imagery Using Transformer Models and Machine Learning Techniques: A Comprehensive Captioning Dataset

Kürşat Kömürcü, Linas Petkevičius

Institute of Computer Science, Vilnius university
Corresponding Author E-Mail: kursat.komurcu@mif.vu.lt

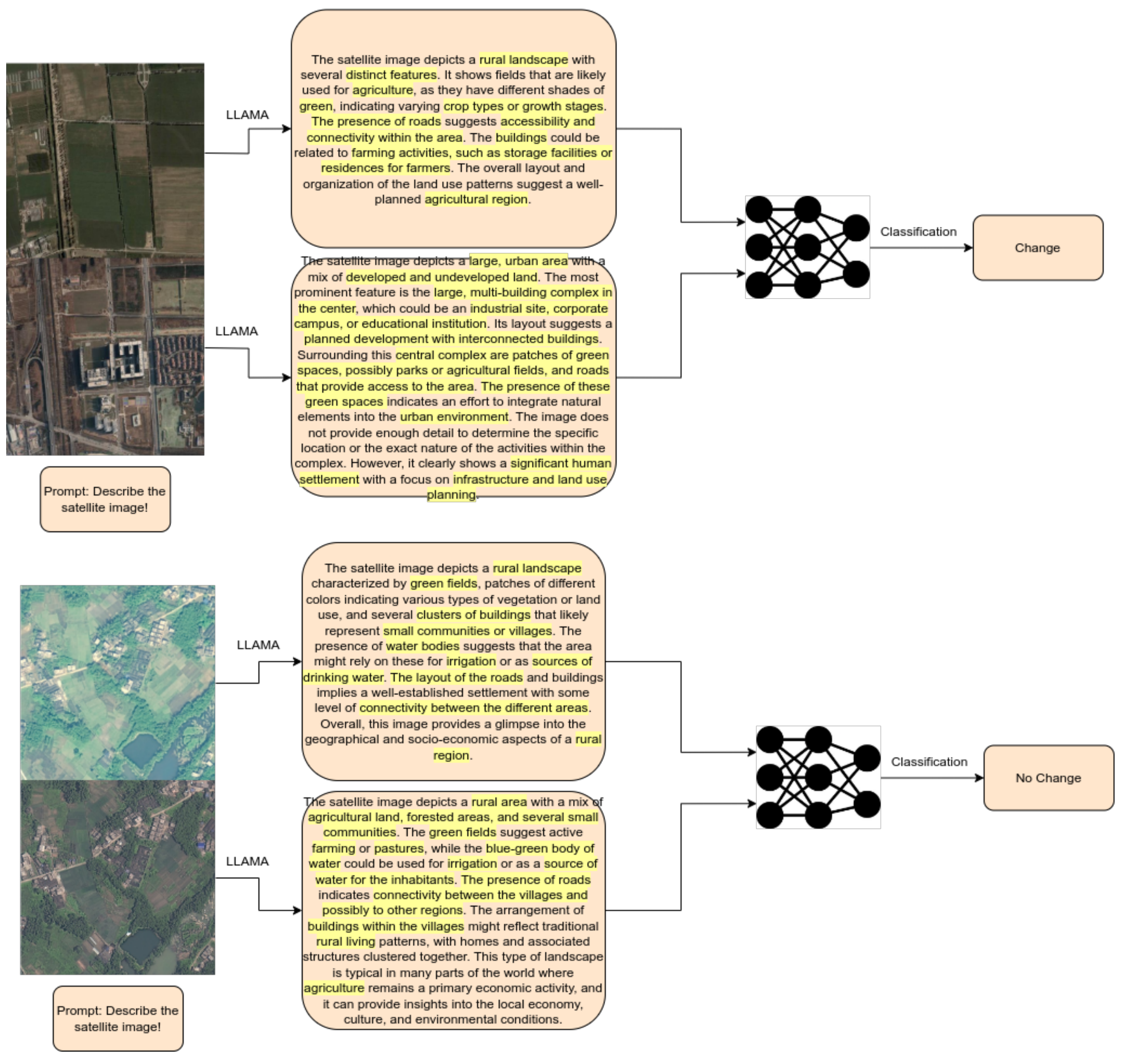
Abstract

The increasing availability of high-resolution satellite imagery enables detailed monitoring of Earth's surface, making change detection a critical task for urban planning, disaster management, and environmental monitoring. This study presents a satellite image captioning dataset, generated using the CLCD, LEVIR-CD, DSIFN, and S2Looking datasets, and leverages the Llama model for caption generation. Transformer models like BERT and RoBERTa, alongside machine learning techniques, were evaluated for change detection performance. The results highlight the effectiveness of these methods in translating complex satellite data into meaningful captions, providing valuable tools for remote sensing applications. The dataset, containing captions for 16,753 image pairs, is publicly available on Kaggle.

The Dataset

Figure 1: Data augmentation examples for a) CLCD, b) LEVIR-CD, c) DSIFN, d) S2Looking datasets

This study utilizes four benchmark datasets for satellite image change detection: CLCD, LEVIR-CD, DSIFN, and S2Looking. The CLCD dataset focuses on multi-sensor land cover change detection, offering valuable insights into land use dynamics. LEVIR-CD is specialized in detecting building changes in urban environments, providing high-resolution image pairs. DSIFN supports detailed change detection through its dual-stream approach, capturing subtle differences in high-resolution images. S2Looking, built on Sentinel-2 imagery, enables large-scale monitoring of diverse landscapes. To ensure a balanced and robust dataset, data augmentation techniques such as random rotation and scaling were applied, addressing dataset imbalances and increasing the diversity of samples. Using these datasets, 16,753 annotated image pairs were generated with detailed captions through the Llama model.



The satellite image depicts a rural landscape with several distinct features. It shows fields that are likely used for agriculture, as they have different shades of green, indicating varying crop types or growth stages. The presence of roads suggests accessibility and connectivity within the area. The buildings could be related to farming activities, such as storage facilities or residences for farmers. The overall layout and organization of the land use patterns suggest a well-planned agricultural region.

The satellite image depicts a large, urban area with a mix of developed and undeveloped land. The most prominent feature is the large, multi-building complex in the center, which could be an industrial site, corporate campus, or educational institution. Its layout suggests a planned development with interconnected buildings. Surrounding this central complex are patches of green spaces, possibly parks or agricultural fields, and roads that provide access to the area. The presence of these green spaces indicates an effort to integrate natural elements into the urban environment. The image does not provide enough detail to determine the specific location or the exact nature of the activities within the complex. However, it clearly shows a significant human settlement with a focus on infrastructure and land use planning.

Prompt: Describe the satellite image!

The satellite image depicts a rural landscape characterized by green fields, patches of different colors indicating various types of vegetation or land use, and several clusters of buildings that likely represent small communities or villages. The presence of water bodies suggests that the area might rely on these for irrigation or as sources of drinking water. The layout of the roads and buildings implies a well-established settlement with some level of connectivity between the different areas. Overall, this image provides a glimpse into the geographical and socio-economic aspects of a rural region.

The satellite image depicts a rural area with a mix of agricultural land, forested areas, and several small communities. The green fields suggest active farming or pastures, while the blue-green body of water could be used for irrigation or as a source of water for the inhabitants. The presence of roads indicates connectivity between the villages and possibly to other regions. The arrangement of buildings within the villages might reflect traditional rural living patterns, with homes and associated structures clustered together. This type of landscape is typical in many parts of the world where agriculture remains a primary economic activity, and it can provide insights into the local economy, culture, and environmental conditions.

Prompt: Describe the satellite image!

Figure 2: Change and no change examples

Methodology

- **Dataset Preparation:** The CLCD, LEVIR-CD, DSIFN, and S2Looking datasets were used to establish a solid foundation for change detection. To address data imbalances and enhance sample diversity, data augmentation techniques such as random rotation ($-90°$ to $+90°$) and scaling (0.5x to 1.5x) were applied.
- **Caption Generation Using Llama Model:** The Llama model, a transformer-based architecture, was employed to generate captions for satellite image pairs. Each pair was processed with the prompt: *"Describe the satellite image!"*, guiding the model to produce coherent descriptions. The model optimized the probability of word sequences:

$$P(w_t|I, w_1, w_2, \ldots, w_{t-1})$$

The objective was to maximize the likelihood of accurate caption generation:

$$\mathcal{L} = \sum_{t=1}^{n} \log P(w_t|I, w_1, w_2, \ldots, w_{t-1}).$$

- **Evaluation Using Machine Learning Models:** Traditional machine learning models, including Logistic Regression, Naive Bayes, Support Vector Machines (SVM), and K-Nearest Neighbors, were trained on the generated captions. Text data was vectorized using the Tf-Idf method before feeding it into the models.
- **Evaluation Using Transformer Models:** Transformer-based models (BERT, DistilBERT, RoBERTa, and XLNET) were fine-tuned to classify changes based on the generated captions. These models utilized self-attention mechanisms to effectively capture dependencies and contextual relationships within the text data.

Table 1: Accuracy Results for Train Data

Models	CLCD	LEVIR-CD	DSIFN	S2Looking
Machine Learning Models				
Logistic Regression	0.8521	0.8644	0.8087	0.8263
Naive Bayes	0.8183	0.7704	0.7348	0.7406
Support Vector Machine	**0.9760**	**0.9838**	**0.9430**	**0.9443**
K-Nearest Neighbors	0.7169	0.8040	0.7910	0.7769
Transformer Models				
BERT	**0.7394**	**0.6846**	**0.7970**	**0.7485**
DistilBERT	0.7113	0.6779	0.7848	0.7438
RoBERTa	0.7183	0.6443	0.7856	0.7438
XLNET	0.7042	0.6644	0.7939	0.7400

Table 2: Accuracy Results for Validation Data

Models	CLCD	LEVIR-CD	DSIFN	S2Looking
Machine Learning Models				
Logistic Regression	**0.7727**	0.7105	0.7762	0.7859
Naive Bayes	0.7545	0.6754	0.7254	0.7181
Support Vector Machine	0.7681	**0.7192**	**0.8084**	**0.7893**
K-Nearest Neighbors	0.6090	0.6447	0.7050	0.6592
Transformer Models				
BERT	**0.6773**	0.6930	0.8169	0.7522
DistilBERT	0.6727	**0.7018**	0.8102	0.7383
RoBERTa	0.6455	0.6360	**0.8288**	**0.7628**
XLNET	0.6455	0.6711	0.8000	0.7357

Conclusions

- This study proposed an analyze for satellite image change detection using caption generation.
- The Llama model successfully translated satellite images into descriptive captions.
- Support Vector Machines and transformer models achieved high accuracy in change detection.

Link to data: Here