Exposé Master Thesis

Utilizing Text Modality to Uncover Insights from Remote Sensing Images

Yuduo Wang
Humboldt University of Berlin, Germany

May 7, 2024

Supervisors

Prof. Dr. Ulf Leser
Humboldt University of Berlin, Germany

Prof. Dr. Pedram Ghamisi
Helmholtz-Zentrum Dresden-Rossendorf, Germany

1 Introduction

Change Captioning (CC) [17] is a challenging task in the fields of computer vision (CV) and natural language processing (NLP), which focuses on generating descriptive captions for images that describe temporal changes or transformations. Unlike traditional image captioning, which describes a single static image, CC requires understanding and describing dynamic evolutions or transitions captured across multi-temporal images.

The goal of CC is to generate accurate and informative textual descriptions that effectively convey the temporal dynamics and changes depicted in multi-temporal images. This task is particularly important in various fields such as surveillance [2], video analysis [9], environmental monitoring [22], and remote sensing imaging [14], where understanding temporal changes in visual data is critical for decision-making and analysis. In this master thesis, we mainly focus on the change captioning task based on remote sensing images. Compared to natural images, CC tasks on remote sensing images have broader applications. It enables us to better understand urban planning, analyze changes in urban construction, and assess the potential of development areas. Additionally, it can be used for damage detection [5], analyzing the severity of damage caused by natural disasters such as earthquakes in a region, facilitating the implementation of subsequent measures, and so on.

CC presents some unique challenges compared to traditional image captioning tasks. First, it requires modeling the temporal relationships between bi-temporal images to accurately capture the change of the scene. Additionally, the generated captions must be coherent and contextual throughout the sequence. To address these challenges, researchers have explored various approaches [3, 8, 14], including recurrent neural networks (RNN), convolutional neural networks (CNN), and Transformer-based models, suitable for efficient processing of bi-temporal data and capturing temporal dependencies. Additionally, techniques such as attention mechanisms, which allow models to focus on relevant parts of bi-temporal images, have been used to improve the quality and relevance of generated captions.

The objective of this master thesis is to design a new model for better CC understanding. At first, a theoretical background and an overview of the current technological landscape for RS CC will be provided. Then a more robust model tailored to address the characteristics lacking in existing models will be established, thereby generating better captions. The proposed new model will be compared with the current state-of-the-art (SOTA) models to analyze their strengths and weaknesses, providing useful insights for subsequent researchers.

2 Related Work

2.1 Change Captioning

Change captioning is a new subtask of image captioning. In this task, each bi-temporal image pair is manually annotated with corresponding change sentences as ground truth. The model needs to generate a more accurate image representation vector to guide the decoder in producing change captions for bi-temporal image pairs. Unlike single-image captioning, this task is more challenging as it requires understanding the content of two images and further discovering and describing their specific differences, such as their objectives. As a pioneering work, Jhamtani et al. [10] proposed a Spot-the-Diff dataset, where image pairs are extracted from surveillance cameras. Similarly, Tan et al. [19] released an Image Editing Requests dataset, which consists of image pairs and corresponding editing instructions.

However, the image pairs in the aforementioned two datasets are pre-matched, and the image pairs are always taken from the same angle. In order to make the image pairs more consistent with scenes in the real world, Park et al. [17] proposed a larger dataset called CLEVER-Change. They also introduced a baseline model called DUDA for generating captions for image pairs with viewpoint changes. The model first utilizes CNN to extract features from the changed image pairs. Subsequently, it computes the difference representation of bi-temporal image features. Then, this representation, along with the features extracted by CNN, is used to compute spatial attention. The generated results guide the RNN-based caption module to generate change captions. In many subsequent studies [7, 11], the focus of the models is mainly on how to better model the difference representation and then guide the decoder model to generate captions. In addition, some studies have proposed methods [6, 21] that involve pre-training followed by fine-tuning. These methods pre-train the model using contrastive learning-based self-supervised learning and then fine-tune the model in downstream tasks, allowing the model to better adapt to the downstream tasks.

2.2 Remote Sensing Change Captioning

Remote sensing change captioning (RSCC) is a recent multimodal task emerging in the remote sensing community. It requires detecting changes in a given set of remote sensing image pairs and generating corresponding captions describing the specific changes. Hoxha et al. [8] first introduced change captioning to the remote sensing community, using pre-trained CNNs for feature extraction from change image pairs, followed by using RNNs and SVMs to generate corresponding textual descriptions. Due to the lack of large change captioning datasets in the remote sensing community, Liu et al. [14] introduced a large-scare LEVIR-CC dataset, which contains 10,077 bi-temporal image pairs, and proposed a transformer-based model called RSICC-former, which is based on a dual-branch architecture. In subsequent studies [3, 13],

models primarily focus on using attention mechanisms to better model the difference representation, thereby improving the performance of the model. Meanwhile, Liu et al. [15] proposed a new paradigm based on prompt tuning to introduce large language models (LLMs) into the RSCC task.

3 Goals, Dataset, and Evaluation

In this master's thesis, we will first analyze and discuss the background of RSCC and existing models. Subsequently, we will attempt optimizations based on the shortcomings of the corresponding models to construct a more robust and effective RSCC model. Specifically, current RS CC models include two fusion modules, making the models relatively complex. Can we design a model that contains only simple fusion modules while maintaining good performance? Additionally, most current models only focus on spatial modeling, enabling them to capture features of different regions in bi-temporal image pairs. However, these models lack temporal and channel modeling, failing to capture the semantic features of specific objects. This deficiency leads to captions that lack detailed descriptions of changes in specific objects. We can address this shortcoming by designing a model with joint modeling to enhance the understanding of specific objects and generate more accurate captions.

We will primarily validate the models on the LEVIR-CC dataset. The LEVIR-CC dataset is derived from a building change detection dataset comprising 637 very high-resolution (0.5 m/pixel) bitemporal images of size 1,024 × 1,024, acquired over 20 regions of Texas, USA [4]. The image pairs, with a time span of 5-14 years, were obtained from the Google Earth API. To facilitate its use in RSCC, the LEVIR-CC dataset was constructed by exploiting 10,077 small bitemporal tiles of size 256 × 256 pixels, with each tile annotated as containing changes or no changes. The dataset consists of 5038 image pairs with changes and 5039 without changes, with each image pair having five different sentence descriptions describing the nature of changes between the two acquisitions. The maximum sentence length is 39 words, with an average of 7.99 words.

To evaluate the performance of the captions generated by the model regarding the changes in bi-temporal images, we will adopt several evaluation metrics commonly used in previous image captioning studies. These metrics are as follows:

- BLEU-N [16]: The BLEU-N metric measures the n-gram similarity between generated sentences and ground-truth sentences. Here, we will adopt n=1, 2, 3, and 4 to compute the BLEU-N score.
- METEOR [1]: The METEOR computes the harmonic mean of precision and recall of single words. Besides, it incorporates a penalty factor to consider the fluency of generated sentences.

Table 1: Current state of the art models on the LEVIR-CC dataset.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Capt-Rep-Diff [17]	72.90	61.98	53.62	47.41	34.47	65.64	110.57
Capt-Att [17]	77.74	67.40	59.24	53.15	36.58	69.73	121.22
Capt-Dual-Att [17]	79.51	70.57	63.23	57.46	36.56	70.69	124.42
DUDA [17]	81.44	72.22	64.24	57.79	37.15	71.04	124.32
MCCFormer-S [18]	79.90	70.26	62.68	56.68	36.17	69.46	120.39
MCCFormer-D [18]	80.42	70.87	62.86	56.38	37.29	70.32	124.44
RSICC [14]	84.72	76.27	68.87	62.77	39.61	74.12	134.12
PSNet [13]	83.86	75.13	67.89	62.11	38.80	73.60	132.62
Prompt-CC [15]	83.66	75.73	69.10	63.54	38.83	73.72	136.44

- ROUGE-L [12]: The ROUGE-L metric measures the similarity of the longest common subsequence between generated sentences and ground-truth sentences.
- CIDEr-D [20]: The CIDEr metric treats each sentence as a document and represents it in the form of Term Frequency Inverse Document Frequency vectors.

The performance of many state-of-the-art (SOTA) models has been collected and presented in Table 1.

References

[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[2] Louisa R Beck, Bradley M Lobitz, and Byron L Wood. Remote sensing and human health: new sensors and new opportunities. *Emerging infectious diseases*, 6(3):217, 2000.

[3] Shizhen Chang and Pedram Ghamisi. Changes to captions: An attentive network for remote sensing change captioning. *IEEE Transactions on Image Processing*, 2023.

[4] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10):1662, 2020.

[5] Laigen Dong and Jie Shan. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. *ISPRS Journal of Photogrammetry and Remote Sensing*, 84:85–99, 2013.

[6] Zixin Guo, Tzu-Jui Julius Wang, and Jorma Laaksonen. Clip4idc: Clip for image difference captioning. *AACL-IJCNLP 2022*, page 33, 2022.

[7] Mehrdad Hosseinzadeh and Yang Wang. Image change captioning by learning from an auxiliary task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2725–2734, 2021.

[8] Genc Hoxha, Seloua Chouaf, Farid Melgani, and Youcef Smara. Change captioning: A new paradigm for multitemporal remote sensing image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[9] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 958–959, 2020.

[10] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4024–4034, 2018.

[11] Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyunsung Park, and Gunhee Kim. Agnostic change captioning with cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2095–2104, 2021.

[12] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[13] Chenyang Liu, Jiajun Yang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. Progressive scale-aware network for remote sensing image change captioning. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 6668–6671. IEEE, 2023.

[14] Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022.

[15] Chenyang Liu, Rui Zhao, Jianqi Chen, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. A decoupling paradigm with prompt learning for remote sensing image change cap-

tioning. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[17] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4624–4633, 2019.

[18] Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. Describing and localizing multiple changes with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1971–1980, 2021.

[19] Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. Expressing visual relationships via language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1873–1883, 2019.

[20] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[21] Linli Yao, Weiying Wang, and Qin Jin. Image difference captioning with pre-training and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2022.

[22] Qiangqiang Yuan, Huanfeng Shen, Tongwen Li, Zhiwei Li, Shuwen Li, Yun Jiang, Hongzhang Xu, Weiwei Tan, Qianqian Yang, Jiwen Wang, et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, 241:111716, 2020.