Towards Unsupervised Remote Sensing Image Captioning and Retrieval with Pre-Trained Language Models

Megha Sharma, Yoshimasa Tsuruoka
The University of Tokyo
{meghas,yoshimasa-tsuruoka}@g.ecc.u-tokyo.jp

Abstract

Captioning and retrieval is an important yet underdeveloped task for large remote sensing databases due to limited training resources. We propose a captioning framework that does not require human-made annotations. Using land cover maps as statistical prompts, we use ChatGPT for natural caption generation. Our experiments with OpenEarthMap reveal that pre-trained language models are capable of inferring and describing high level environment information using land cover statistics of the image.

1 Introduction

Recent improvements in remote sensing technology have led to an unprecedented growth in the applications of Remote Sensing Images (RSI) for Geo-spatial Information Systems. One such application is known as Land Usage / Land Cover Mapping (LULC), which aims to classify and understand the terrain and usage of the piece of land captured in the RSI [1]. Well-maintained LULC can aid in land cover monitoring, resource planning, and land management [1] [2].

With more data available than ever, accurate and robust image captioning and retrieval methods have become important to maintain such large databases. OpenEarthMap (OEM) is one such benchmark LULC dataset that offers diverse locations and high quality land cover annotations [2]. Historically, retrieval systems for these databases are queried by image or image patterns [3]. However, it is more intuitive for humans to annotate or query with textual descriptions instead [4]. Cross-modal text-image retrieval and captioning systems present an interesting opportunity to combine Geo-spatial information with Natural Language Processing (NLP) techniques. The opportunity explores the challenge of measuring cross-modal semantic similarity. Intuitively, this can be approached as a two-step process of converting and comparing the database mode to the query mode. Hoxha et al.[5] successfully used deep learning networks to measure similarities between captions generated from the RSI in the database with the queried text. However, these frameworks depend on a large amount of labelled training data to successfully annotate images and measure similarity with query text. Existing caption datasets, such as [6] [7], are limited by the resources required to generate captions. Moreover, the captions are often based on lower-level knowledge such as identified objects like airplanes and stadiums, which is difficult to evaluate without further expert knowledge on the location. On the other hand, LULC information is usually predefined and easier to distinguish even to a common eye.

We are motivated to study unsupervised approaches for RSI captioning to reduce the overhead of manually generating such annotations to train for retrieval systems. Template-based methods are popular to extract visual descriptions into a rule-based output. The annotations by this approach are arguably limited for describing higher level semantic information [5]. However, language models have since made considerable leaps, one such being the release of ChatGPT[1)], a pre-trained language model trained to perform like a chat bot.

In this paper, we aim to explore if it is possible for such pre-trained language models to recognize higher-level information such as landscape (e.g., rural, urban, and forest) with controlled and uncontrolled generation based on statistical LULC information. We found that pre-trained models could successfully generate natural language captions for the RSI, bringing us one step closer to developing a framework of cross modal retrieval and captioning systems without labelled data. The code is available on GitHub[2)].

1) https://chat.openai.com/chat
2) https://github.com/ms3744/OEM-Land-Cover-Experiments

2 Method

Let $I = \{I_1, I_2, \cdots, I_N\}$ be a database of N images. For any image I_i in the database, there is a corresponding LULC cover map for the image, I_i^{LC}. We can divide the unsupervised caption generation into three main steps:

1. Land Cover Mapping
2. Statistical Image Understanding
3. Natural Language Generation

The framework, illustrated in Fig. 1, can be used as is for annotation-based retrievals [5] or used to generate ground truth for few shot learning models for supervised image to text retrieval models, such as those proposed in [8].

2.1 Land Cover Mapping

To create a truly end-to-end model for unlabelled captioning, we first predict land cover maps from remote sensing images. We experimented with three different types of UNet models, in particular the recent UNetFormer, which gained traction for semantic segmentation [9]. An advantage of these models is that it can be adapted using many different types of vision transformers as its backbone encoder. The network is trained using labelled image pairs of RSI and LULC images. The predictions are a 2D array where each cell's location (row, column) corresponds to the predicted class of the pixel at the same location.

2.2 Statistical Image Understanding

Feature extraction plays a key role to project one mode of information to another. Historically, cross-modal text-image systems extract semantic topics using a statistical and semantic analysis of the captions [10]. We translate the image composition and land cover structure into a template statistical prompt for the next step. The land cover classes are mapped to their text labels pre-defined by the training dataset in the prompt. Examples of the statistical prompt is shown in Fig. 2. The image understanding steps are explained in subsections below.

2.2.1 Image Composition

Suppose class C is one of the classes present in a predicted LULC image $I_i^{LC'}$, we calculate the percentage of the composition for class C as follows.

$$Compostion(C \mid I_i^{LC'}) = \frac{\text{\# pixels predicted as } C}{\text{Total \# of pixels}} \quad (1)$$

2.2.2 Land Cover Structure

To support image composition, we also describe the locations of prominent land classes. Using the median location of top 2 classes in the LULC, we locate the centroid of each class. Unlike calculating mean, median is more robust to outliers. If the centroid falls into a circle with a radius of 10% of image size from the center, it is classified with *Center* location. This threshold accounts as a reasonable margin to describe the location of the LULC class. Any centroid outside of this area is classified as either *Upper left, Upper Right, Lower Left*, or *Lower Right* by dividing the image into four quadrants with the center of the image as the center of the quadrant.

2.3 Natural Language Generation

In the natural language generation stage, we use the inference of large pre-trained models to convert low level statistical text to a natural sounding caption. In this paper, we use ChatGPT, which was capable of summarising radiology reports without jargon for patients [11]. The statistical prompt is input to ChatGPT for either controlled or uncontrolled generation. In uncontrolled generation the ChatGPT is only asked to "write a natural image caption in 2 lines without numbers" for the given statistical prompt without a ground truth. Restrictions are placed on the length and usage of numbers to reflect annotations found in existing datasets [6]. To test the model against some ground truth, we created four captions as positive examples for ChatGPT. The motivation for controlled generation stems from supervised few-shot training, where we give ChatGPT a few positive examples for generated captions and then ask it to generate given novel statistical prompts, in an attempt to raise behavior expected from the model. We also used five examples from a popular dataset with annotations, RSICD dataset [6] for a more robust ground truth. The images selected from the RSICD dataset were selected to have annotations related to the environment or land cover, as most of the images in the dataset actually describe specific objects or buildings.

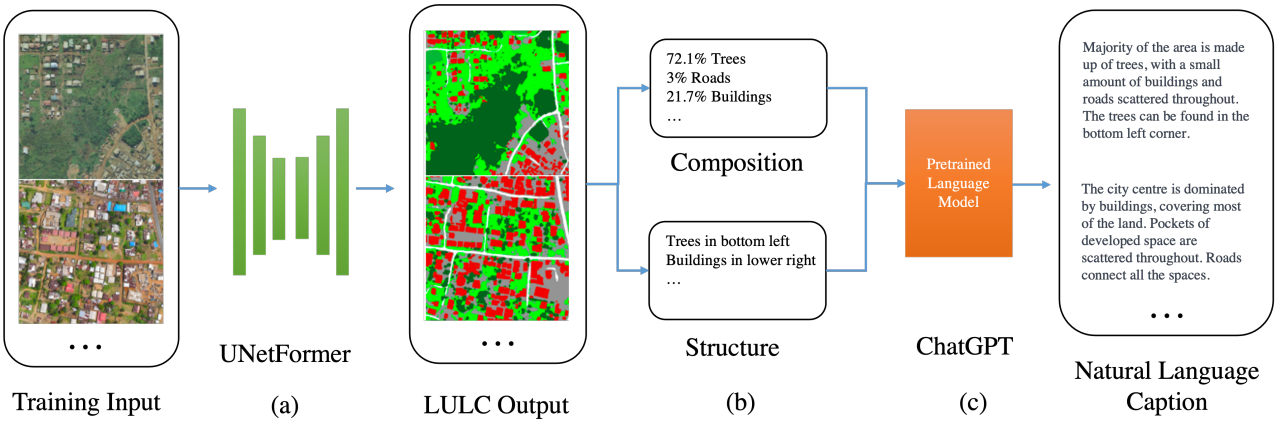


Figure 1 The 3-step framework for unsupervised natural language caption generation. In (a), the training data is used to train a model (e.g. UNetFormer) to generate land cover maps of the input. The inferred LULC outputs are used to extract visual information including composition and structure in (b). Finally, in (c), the template is sent as an input to a language model (here we use ChatGPT), to generate natural language captions inferring higher level semantics such as environment (e.g., city center).

3 Experiment with OpenEarthMap and ChatGPT

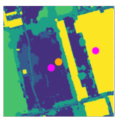
To test the framework, we train several UNet models on the OEM Mini, a smaller version of the OEM dataset with 1068 examples with sub-meter resolution classified into 8 land cover labels. We used a 70-30 train-test split of the dataset for training, and trained with a batch size of 16. The RGB images were divided into three input channels, and each image was transformed into a 512 * 512 size image. The learning rate was set at 10^{-3} with a weight decay 10^{-6}. The model is trained with the Dice loss function [12] and Adam optimisation. We objectively evaluated the land cover model using pixel-wise mean Intersection over Union (mIoU), a standard metric to measure the coverage between the true and predicted classes. Our experiments found that the UNetFormer model with SE Res-Net encoder performs best with an mIoU of 53.4% on the validation dataset and 63.13% on the test dataset. Using the trained UNetFormer with SE Res-Net land cover model, we retrieved 6 land cover maps from the test split of the dataset, intentionally selecting images with diverse land composition to study the performance across a variety of environments. The extracted statistical information was fed with controlled and uncontrolled prompts into the ChatGPT, and we found that ChatGPT was able to successfully convert the prompts to natural sounding captions.

3.1 Uncontrolled Generation

Without seeing any ground truth, ChatGPT is still able to infer urban environments in the generation captions. As seen in the left-hand example in Fig.2, ChatGPT describes the image as urban (city-like) when "buildings" is one of the top two classes. This presents the inherent potential of language models to infer higher level information, such as understanding a high concentration of buildings and grass reflects a "mix or urban and natural elements". The language model is also able to describe quantitative information using relative descriptors, such as converting "31.3% buildings" to "buildings taking up a significant portion of the frame" and "1.2% Water" to "a hint of water". All of the generated captions were found to be complete, i.e., they used all information provided in the prompt.

The generated captions particularly featured prominent use of creative adjectives. These adjectives seem to be inferred from the language model's training data. For example, "lush trees", "tall buildings", "bustling cityscape" were used in the captions, when no such descriptions were used in the prompt. In the best case scenario, these are positive descriptors of the image; however at worst, these are classified as hallucinations. Hallucination is a common problem with ChatGPT [11]. Albeit not dangerous in this application, it can still raise false positives when querying the annotations. The quality of the land cover map also raises false artefacts in the description when predicting incorrect land classes. This is shown in the right hand example in Fig. 2. The buildings are described as "towering", and a non-existent "body of water" is also described in the caption.

Statistical Prompt:
4.3% Bareland, 3.4% Grass, 24.5% Pavement, 8.1% Road, 1.1% Tree, 27.3% Water, 31.3% buildings, buildings in upper left. Water in lower right.
Uncontrolled Generation:
The image shows a bustling cityscape, with buildings taking up a significant portion of the frame. In the upper left corner, there are numerous buildings, while the lower right corner is dominated by a body of water.
Controlled Generation:
A bustling city with towering skyscrapers rising up in the upper left, while a tranquil body of water stretches out in the lower right. Pavement, roads, and a few scattered trees can be seen throughout the city, adding a touch of greenery to the urban landscape



Statistical Prompt:
32.1% Grass, 7.8% Pavement, 2.4% Road, 17.2% Tree, 11.9% Water, 28.7% buildings, Grass in lower left. buildings in upper right.
Uncontrolled Generation:
A mix of urban and natural elements can be seen in this image, with lush grass covering a large portion in the lower left corner. Buildings take up a significant portion in the upper right, while trees and a small section of pavement and road can be seen scattered throughout. A body of water takes up a portion of the image as well.
Controlled Generation:
A cityscape with lush green grass in the lower left, and towering buildings in the upper right. A few scattered roads and pavement can be seen throughout the city, and a small body of water adds a touch of tranquillity to the scene.

Figure 2 A positive (left hand side) and negative (right hand side) run of the framework with controlled and uncontrolled generation. In the land cover maps (below original image), the orange dot in the center represents the center of the image and the magenta dots represent the median of the two most prominent classes. In the generations, yellow highlights the higher level environment information inferred from the statistics, green highlights the creative adjectives used by ChatGPT that may or may not be accurate to the image, and red highlights the incorrect inferences in the generated captions.

3.2 Controlled Generation

Despite a promising start, uncontrolled generation struggled to infer non-urban environments such as rural or forest areas. Hence, we gave ChatGPT examples of four positive examples of captions, using diverse statistical prompts and asked it to re-generate captions for the previous six land cover maps. One of the most important discovery was that controlled generation managed to solve the biggest issue at hand, and ChatGPT correctly inferred environment descriptors such as "forest", "countryside", "metropolis", and "city" in the controlled generation. In general, this also helps control the hallucinations in the generated captions, and strengthens the proposal of the methodology.

In order to truly test the framework against existing ground truth, we also explore results of controlled generations using the RSICD dataset [13]. We found that ChatGPT was able to successfully mimic the style of the RSICD dataset, with short and generic sentences about the images. This also highlights another ability of large pre-trained language models to mimic styles when generating captions. The generated captions contain similar keywords as the ground truth, and this was made apparent with the 50.1% BLEU1 score of the generated captions against ground truth of the RSICD and the 56.6% BLEU1 score with positive examples from OEM Mini as described in Tab.1. The BLEU1 and BLEU2 scores refer to overlapping 1-gram and 2-gram in the reference and candidate sentences.

3.3 Future Opportunities

ChatGPT and large pre-trained language models present a creative opportunity towards natural image captioning.

Table 1 BLEU1 and BLEU2 scores with the ground truth of RSICD and positive examples for OpenEarthMap

Dataset	BLEU1	BLEU2
OpenEarthMap	56.6%	20.4%
RSICD	50.1%	14%

From our preliminary findings we propose the following opportunities and challenges ahead

- **Few-shot Modeling** - Using generated captions as ground truth for a few-shot learning captioning models such as those proposed in [14].
- **Autonomous Caption and Retrieval Systems** - ChatGPT managed to recover higher level information about the environment using low level land class information. This presents an unprecedented opportunity with autonomous captioning and retrieval systems. Although there are risks of hallucination and misinformation, with controlled generation, we can explore a framework where a user can search for an RSI image using vague text descriptions. We can also expand the framework with multi-level information [15] [16].

4 Conclusion

We found that pre-trained models like ChatGPT are able to infer information about the image with just a statistical description of remote sensing images. Although the scope of the experiments was limited, the results spark an exciting conversation of using large pre-trained models to automate labour intensive tasks in the field of Remote Sensing Image Analysis. In future work, we plan to further elaborate the framework for caption and retrieval databases for land cover map images with more robust experiments and future opportunities.

Acknowledgements

This paper would have not been possible without the work of Naoto Yokoya. Also extending our gratitude to Qiyu Wu, who helped us gain an interesting take on our experiments.

References

[1] John Rogan and DongMei Chen. Remote sensing technology for mapping and monitoring land-cover and land-use change. **Progress in planning**, Vol. 61, No. 4, pp. 301–325, 2004.

[2] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision**, pp. 6254–6264, 2023.

[3] Jaroslaw Jasiewicz and Tomasz F Stepinski. Example-based retrieval of alike land-cover scenes from nlcd2006 database. **IEEE geoscience and remote sensing letters**, Vol. 10, No. 1, pp. 155–159, 2012.

[4] Mohamad M Al Rahhal, Yakoub Bazi, Taghreed Abdullah, Mohamed L Mekhalfi, and Mansour Zuair. Deep unsupervised embedding for remote sensing image retrieval using textual cues. **Applied Sciences**, Vol. 10, No. 24, p. 8931, 2020.

[5] Genc Hoxha, Farid Melgani, and Begüm Demir. Toward remote sensing image retrieval under a deep image captioning perspective. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, Vol. 13, pp. 4462–4475, 2020.

[6] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. **IEEE Transactions on Geoscience and Remote Sensing**, Vol. 56, No. 4, pp. 2183–2195, 2017.

[7] Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. Deep semantic understanding of high resolution remote sensing image. In **2016 International conference on computer, information and telecommunication systems (Cits)**, pp. 1–5. IEEE, 2016.

[8] Qimin Cheng, Yuzhuo Zhou, Peng Fu, Yuan Xu, and Liang Zhang. A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, Vol. 14, pp. 4284–4297, 2021.

[9] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. **ISPRS Journal of Photogrammetry and Remote Sensing**, Vol. 190, pp. 196–214, 2022.

[10] Binqiang Wang, Xiangtao Zheng, Bo Qu, and Xiaoqiang Lu. Retrieval topic recurrent memory network for remote sensing image captioning. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, Vol. 13, pp. 256–270, 2020.

[11] Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, et al. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. **arXiv preprint arXiv:2212.14882**, 2022.

[12] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In **Deep learning in medical image analysis and multimodal learning for clinical decision support**, pp. 240–248. Springer, 2017.

[13] Shan Cao, Gaoyun An, Zhenxing Zheng, and Qiuqi Ruan. Interactions guided generative adversarial network for unsupervised image captioning. **Neurocomputing**, Vol. 417, pp. 419–431, 2020.

[14] Haonan Zhou, Xiaoping Du, Lurui Xia, and Sen Li. Self-learning for few-shot remote sensing image captioning. **Remote Sensing**, Vol. 14, No. 18, p. 4606, 2022.

[15] Zhenwei Shi and Zhengxia Zou. Can a machine generate humanlike language descriptions for a remote sensing image? **IEEE Transactions on Geoscience and Remote Sensing**, Vol. 55, No. 6, pp. 3623–3634, 2017.

[16] Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Xuee Rong, Zhengyuan Zhang, Hongqi Wang, Kun Fu, and Xian Sun. Remote sensing cross-modal text-image retrieval based on global and local information. **IEEE Transactions on Geoscience and Remote Sensing**, Vol. 60, pp. 1–16, 2022.

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 7132–7141, 2018.

A Appendix

A.1 Results from the LULC Modeling

Tab. 2 covers the results from training 8 different types of models for the land cover mapping. The results are shown on the validation dataset only. The SE Res-Net module uses Squeeze-and-Excitation blocks to model inter-dependencies between channels [17]. The ability for SE Res-Net to re-calibrate its features works well with semantic segmentation tasks.

Table 2 Results from LULC modeling experiments

Model	Backbone	mIoU (%)
UNet	vanilla CNN	16.8%
UNetFormer	efficientnet-b0	43.5%
	resnetv2-101x1-bitm	41.9%
	res2net101-26w-4s	43.5%
	res2next50	43.4%
	seresnet152d	**53.4%**
	swsl-resnet18	40.7%
FTUNetFormer	SwinTransformer	42.4%

A.2 Example Generations

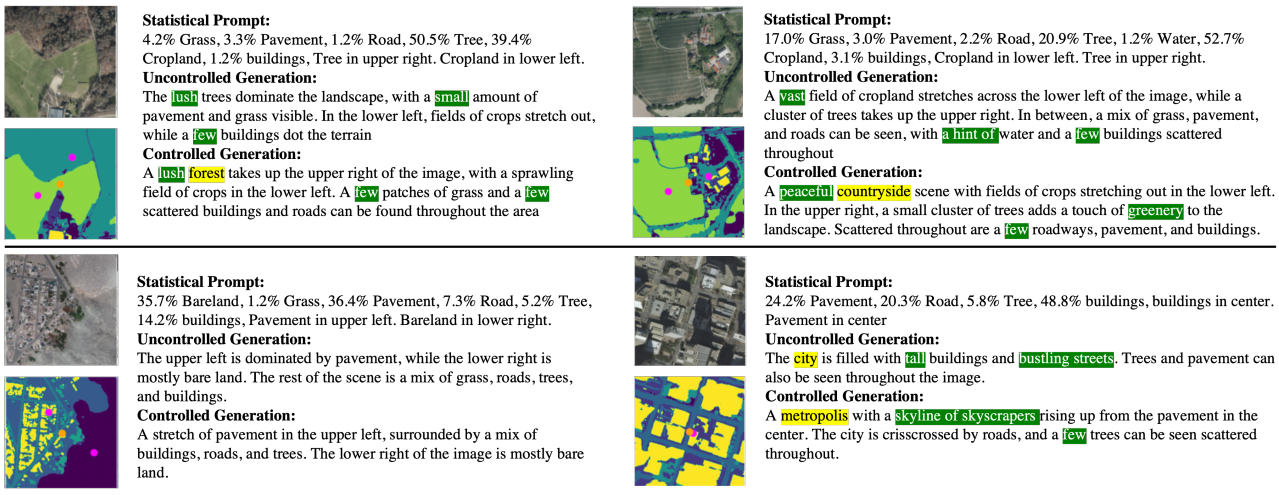


Figure 3 Uncontrolled and controlled generations of the OEM Mini dataset. In the land cover maps, the orange dot in the center represents the center of the image and the magenta dots represent the median of the two most prominent classes.

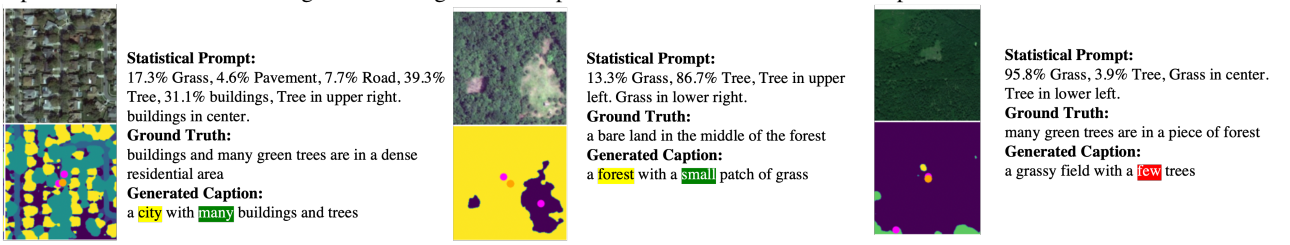


Figure 4 Generations and ground truth of the RSICD dataset. In the land cover maps, the orange dot in the center represents the center of the image and the magenta dots represent the median of the two most prominent classes.