UNIVERSITY OF CALIFORNIA, MERCED

Improving Cross-View Remote Sensing Image Retrieval with Images and Captions

A masters project report submitted in partial satisfaction of the requirements for the degree of

Masters of Science

in Electrical Engineering and Computer Science

by

Akshay Bhatia

Committee:
Professor Shawn Newsam, Chair
Professor Ming-Hsuan Yang
Professor Shijia Pan

May 2023

Acknowledgements

I would like to express my sincere gratitude to my advisor, Professor Shawn Newsam, for his guidance, encouragement, and support throughout the course of this project and my Masters's program. His expertise and insights were invaluable in shaping this work and I have learned a great deal from his mentorship.

I would also like to thank the members of our Computer Vision lab - Yuxin, Jianan, Haolin and Shreeshail for their collaboration and feedback. Their enthusiasm for research and dedication to excellence made this project both challenging and rewarding.

Finally, I would like to thank my family and friends for their unwavering support and encouragement. Their love and belief in me have been my constant motivation and inspiration.

Abstract

Cross-View Remote Sensing Image Retrieval or cross-view geo-localization is a fundamental research area in remote sensing image analysis. It is used to determine the position of a ground image query by correlating it with a database of geo-tagged satellite images and is often used for several applications such as disaster and damage assessment and monitoring road and terrain network understanding. But a common challenge for this task is the significant variation in view angles and time differences between the ground-level image and the corresponding aerial image. As a result, it is very difficult to capture global semantics and other relations between the two image pairs. Recent research has made remarkable strides in remote-sensing image retrieval benchmarks but conventional methods often overlook other modalities such as textual captions which describe the entities and other information present in ground-level images.

We propose a new approach to enhance the cross-view image retrieval results by utilizing both images and associated textual captions depicting geographic content that describe the contents of the ground-level image. Our framework extracts geographic content and terrain features from the ground-level image and text caption. Finally, we curate a new dataset based on an existing geographic image captioning dataset, GeoRic. We do this by scraping overhead imagery for the corresponding ground-level images in the dataset using Google Static Maps API. We then demonstrate the effectiveness of our multi-modal approach on the newly created dataset by comparing it with existing unimodal and multi-modal deep learning-based image retrieval methods. Experimental results show that our approach outperforms the traditional image retrieval method and performs competitively with an image-text retrieval model in terms of retrieval accuracy. The incorporation of captions improves retrieval performance, especially in cases where the images have complex and varied visual content. In summary, this project proposes an approach to enhance the remote sensing image retrieval results by utilizing both images and captions. The proposed approach can capture the complex relationships between the images and captions and achieves superior performance compared to traditional unimodal image retrieval methods.

Contents

1. Introduction

Remote sensing has become an essential tool for capturing the earth's surface information for various applications such as agriculture, vegetation mapping, urban planning, disaster management etc. Remote sensing tasks are now utilizing different modalities, which comprise of various types of data sources and perspectives. It has become an essential tool for capturing the earth's surface information for various applications such as agriculture, urban planning, and disaster management. The reason behind this trend is that multi-view data can provide more valuable information than single-source data. However, the quality of the data may vary from view to view, which can limit the potential benefits of multi-view data. To overcome this limitation, it is crucial to merge the multi-view data with other modalities such as text, audio, geolocation etc. to maximize the benefits of using multi-view data in remote sensing tasks.

Cross-view image retrieval is a challenging task that involves retrieving images from a target view given a query image from a source view. In the case of satellite imagery, where the source view is a ground-level image and the target view is a satellite image, traditional cross-view image retrieval approaches based on handcrafted features and similarity metrics may not capture the complex relationships between the source and target views. In addition to this remote sensing satellite imagery presents its own challenges compared to natural images. These include a) *scale in-variance* i.e the scales of the objects and entities in satellite images vary greatly compared to natural images, b) *lack of notable objects*: they cover many types of land-cover objects like agricultural fields, cities, etc and other common objects may be missing, c) *directional dissimilarity*: since the images are captured from aerial views, they are significantly different from natural images.

To this end, we propose a geographic image-text framework comprising a geographic content extraction module and a geographic terrain extraction module. The idea is to extract the channel statistics using the terrain extraction module and local features from the content extraction module of the ground-level reference image.

Our method has a few advantages over traditional cross-view image retrieval approaches. First, it leverages both visual and textual information to better capture the semantic content of the images. Second, it is end-to-end trainable, allowing us to learn joint representations that are optimized for the retrieval task. Third, it is flexible and can be applied to various types of satellite imagery retrieval tasks, such as land cover classification:

The contribution of this project is as follows:

- We modify an existing remote sensing image captioning dataset and curate a new multi-modal image retrieval dataset for remote sensing applications.
- We propose a multi-modal method for cross-view remote sensing image retrieval. Our approach involves utilizing both image and its associated captions to describe the content of the image by combining a geographic content extraction module and a geographic terrain extraction module to retrieve the corresponding aerial image.
- We show that the proposed method is better than traditional uni-modal image retrieval methods and compares competitively with existing multi-modal methods.
- To show the effectiveness of our approach in terms of its practicality and real-world application, we perform an extensive study detailing how manually created captions can produce better retrieval results compared to existing captions sourced from the previous dataset.

The remainder of this report is organized as follows. In Section 2, we review related work on cross-view image retrieval and multi-model learning. In Section 3, we describe our proposed method in detail. In Section 4, we present experimental results on a cross-view image retrieval dataset for satellite imagery and discuss also discuss some challenges and how to deal with them in a pragmatic way. Finally, we conclude in Section 5 with a summary of our contributions and directions for future work.

2. Related Work

Cross-view image retrieval is an important problem in computer vision, particularly in the context of satellite imagery, where the same location can be observed from different viewpoints. Classical machine learning techniques such as bag-of-words (BoW), Support Vector Machine (SVM), Gaussian Trees, and spatial pyramid matching (SPM) have been widely used for this problem, while more recent deep learning techniques such as convolutional neural networks (CNNs) have shown promising results.

Classical Machine Learning Techniques

BoW and SPM are two commonly used techniques in cross-view image retrieval. The frequency of each visual word in an image is used as its feature vector. SPM is an extension of BoW that divides an image into subregions and applies BoW to each subregion. The resulting feature vectors are concatenated and normalized to obtain the final feature vector. Gao et al. [2] proposed a cross-view image retrieval method that combines BoW and SPM. They extracted BoW features from each view and concatenated them into a joint feature vector. They also applied SPM to each view and concatenated the resulting feature vectors into another joint feature vector. Finally, they used a linear SVM to learn a mapping from the joint feature vectors to a common subspace.

BoW is a representation method that treats an image as a collection of visual words, which are obtained by clustering local features such as SIFT or SURF. Karami et.al [11], evaluated the efficacy of three image matching techniques, namely SIFT, SURF, and ORB, under varying conditions of transformations and deformations such as scaling, rotation, noise, fish-eye distortion, and shearing.

SVM for Image Retrieval was used by "CBIR by cascading features SVM" [9] where they utilized an SVM classifier to train and classify five sets of features. These features were obtained from various sources, including histograms, GLCM (gray-level co-occurrence matrix) textures, wavelets, Gabor filters, and statistical measures. By combining global and local features, our approach aimed to improve the accuracy and effectiveness of the classification process.

[12] employs Latent Dirichlet Allocation (LDA) as a method for encoding visual features, which has proven effective in modeling medical images and investigates both early and late fusion methods for combining these visual features with textual features.

Deep Learning Techniques

Deep learning has shown promising results in various computer vision tasks, including cross-view image retrieval. CNNs, in particular, have been widely used for this problem.

Cross-view image retrieval has been extensively studied in the computer vision community. Traditional cross-view image retrieval approaches are typically based on handcrafted features and similarity metrics. For example, in Zhang et.al [16], the authors used a sparse coding method to combine multiple feature types for the purpose of large-scale image retrieval. Subsequently, they adopted feature pooling strategies in image retrieval, using a probabilistic approach within the framework of sparse coding, and used a modified sum pooling technique, which considerably enhances the accuracy of image retrieval.

Hoang et.al [4] proposed a framework to achieve better retrieval performance by utilizing 3 masking schemes, namely SIFT-mask, SUM-mask, and MAX-mask, and use of embedding and aggregating techniques to choose a representative subset of local convolutional features and eliminate redundant features which can effectively address the issue of feature burstiness and enhance retrieval accuracy.

Siamese Networks have also been used for image retrieval tasks. For example in Wiggers et. al [14], rather than relying on manual feature engineering, the authors employ a Siamese Neural Network to learn a similarity-based representation based on a subset of image pairs from the ImageNet dataset. This learned representation is subsequently employed to create similarity-based feature maps, which are used to identify pertinent image candidates in the data collection in response to an image query. Similarly, Jiang et. al. [5] propose Dual Attention Triplet Hashing Network (DATH) with a two-stream ConvNet architecture that involves two neural networks. The first network emphasizes spatial semantic relevance, while the second network emphasizes channel semantic correlation. Additionally, a triplet likelihood loss and classification loss are used to better utilize label information during network optimization.

More recently, the success of Transformers in natural language understanding and image classification has prompted them to be widely used for cross-view image retrieval tasks. For example, in "Training Vision Transformers for Image Retrieval" [1] the authors leverage a transformer-based approach for image retrieval to generate image descriptors and train the resulting model using a metric learning objective that combines a contrastive loss with a differential entropy regularizer.

In summary, cross-view image retrieval for remote sensing is a challenging task that has been addressed using a variety of classical machine learning techniques and more recent deep learning techniques. While traditional approaches are based on handcrafted features and similarity metrics, recent approaches have focused on using deep learning models to learn joint representations of images and text. Transformers have also been shown to be effective for cross-view image retrieval tasks, particularly in handling multiple source views. Our proposed method is inspired by these recent developments in deep learning and multimodal learning and aims to improve upon existing methods by utilizing both ground-level images and captions to retrieve relevant satellite images.

3. Proposed Method

Problem Formulation: Consider a set of ground-level images, text captions and the corresponding aerial images $\{(I_i^g, T_i^g, I_i^a)\}, i = 1, \ldots, N$ where I_i^g and I_i^a refer to the ground level and aerial image respectively, T_i^g refers to the text caption and N refers to the number of pairs. Given the query pair I_q^g, T_q^g, the task is to retrieve the best target image I_q^a. In this image retrieval problem, each reference image and caption has one and only one target image i.e a 1 to 1 correspondence.

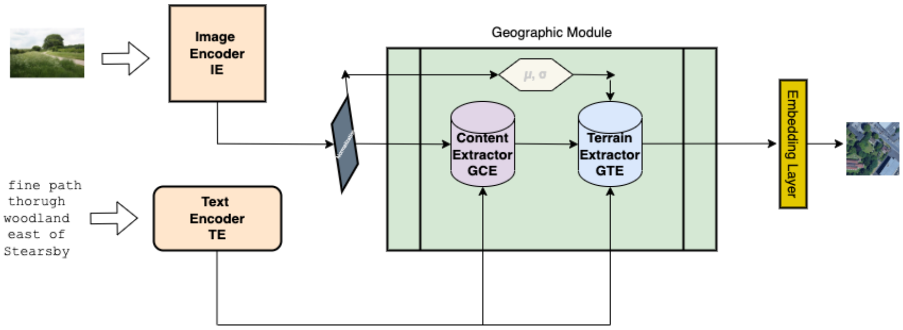


Figure 3.1: An overview of our approach. The geographic module consists of a geographic content extractor and a geographic terrain extractor and is followed by an embedding layer

We first extract high-level representations from the ground-level image and caption. For our image encoder module IE, we choose ResNet50 [3] pre-trained on Imagenet. The output, which we refer to as im_i^g, is a 3-d matrix representation of shape HxWxC where the number of channels C is 2048. More specifically we get the output from layer 4 since it is responsible for extracting high-level features that are useful for recognizing complex patterns in the input image. This is because the bottleneck blocks in layer 4 are designed to increase the number of channels from 512 to 2048, which allows them to capture more complex and abstract features. For our text encoder TE, we use LSTM with 1024 hidden units followed by a linear layer to get a 512 fixed dimensional vector which we refer to as tx_i^g.

$$im_i^g = IE(I_i^g) \tag{3.1}$$

$$tx_i^g = TE(T_i^g) \tag{3.2}$$

3.1 and 3.2 exhibit the process of extracting high-level representations from the ground-level image and text captions - the Image encoder IE takes the ground-level image as input I_i^g and outputs im_i^g of shape HxWxC. The text encoder TE accepts a sequence of words T_i^g and outputs a 1-D representation tx_i^g.

$$geo_i^g = GE(im_i^g, tx_i^g) \tag{3.3}$$

Next, we pass these representations to the geographic module GE which combines them in a way that the text and image cohere to form a compact model of the image-text features. We refer to this representation as geo_i^g according to 3.3.

$$im_i^t = IE(I_i^t) \tag{3.4}$$

The target aerial image is also encoded using the image encoder IE to get im_i^a as shown in 3.4. The idea is that geo_i^g and im_i^ta will be similar to each other in some embedding space. To this end, we build a final embedding layer EM with a global average pooling layer followed by an MLP linear layer which outputs a fixed-size 512-dimensional representation.

$$f_i^g = EM(geo_i^g) \tag{3.5}$$

$$f_i^a = EM(im_i^a) \tag{3.6}$$

3.5 and 3.6 show the final outputs for the ground-level image and caption pair and target image respectively. We choose the popular batch-based classification loss to calculate the difference between the final representation from the embedding layer. The batch-based classification loss is typically calculated using a softmax function, which converts the raw outputs into a probability distribution over the possible classes. The softmax function is applied to the outputs of the embedding layer for each image in the batch, resulting in a set of probability distributions. The loss function $\mathcal{L}_{\text{geo}}$ used for batch-based classification typically takes the form of the negative log-likelihood of the true labels given the predicted probabilities.

$$\mathcal{L}_{\text{geo}} = \frac{1}{B} \sum_{i=1}^{B} - \log \frac{\exp\left(cosine\left(\mathbf{f}^{g,i}, \mathbf{f}^{a,i}\right)\right)}{\sum_{j=1}^{N} \exp\left(cosine\left(\mathbf{f}^{g,i}, \mathbf{f}^{a,j}\right)\right)} \tag{3.7}$$

In the sections below we discuss the geographic module consisting of the Geographic Content Extractor and Geographic Terrain Extractor.

3.0.1 Geographic Content Extractor

The geographic content extractor *GCE* works on the following principle: We are initially interested in searching for all the geographic entities in the ground-level image - vehicles, buildings etc. To do this, we first apply a technique known as instance normalization to the image encoder representation. Instance normalization is a technique used in deep learning to normalize the output of intermediate layers of a neural network. In instance normalization, the mean and standard deviation of each feature map in the output of a layer are computed independently for each image or instance in the batch. This means that each image is normalized independently, as opposed to batch normalization which normalizes the features across the entire batch. This allows us to focus on the entities of the ground-level image rather than the overall geographic terrain information.

$$in_i^g = (im_i^g - mean(im_i^g))/sqrt(var(im_i^g) + \epsilon) \tag{3.8}$$

We use a simplified version of instance normalization where the shift and scale parameters gamma γ and beta β are set to 1 and 0 respectively. We do this due to the complexity of the model and fear that additional learning parameters may lead to overfitting. Removing these parameters can help the normalization operation to become more rigid, and the model may become more robust to changes in the input distribution. We also add a small positive constant ϵ to the denominator in our instance normalization formula. This ensures that the denominator in the normalization formula is always positive, even if the variance of the feature map is very small or zero, and helps prevent numerical instability and division by zero errors during training and thus stabilizes the computation.

The geographic content extractor directly employs a feature map of the Disentangled Non=Local Block introduced in Disentangled Non-Local Neural Networks [15]. The Disentangled Non-Local(DNL) Block is a building block introduced in the paper "Disentangled Non-Local Neural Networks". This block is designed to capture long-range dependencies in multi-modal data by disentangling the spatial and channel-wise information of the input feature maps. The DNL Block first applies a disentangled non-local operation to the spatial feature maps, which allows it to capture long-range dependencies in the spatial domain. The disentangled non-local operation is performed separately for each spatial position, and it computes a weighted sum of the values of all the spatial positions based on their similarities with the query position. After the disentangled non-local operation is applied to the spatial feature maps, the DNL Block applies a channel-wise attention mechanism to the channel-wise feature maps. This attention mechanism allows the block to selectively focus on the most relevant channels for the task at hand. The output feature map of the DNL Block is a combination of the disentangled non-local features and the channel-wise attention features. This output feature map is then passed to the next layer in the network for further processing. In our work, the DNL block takes in the instance normalization feature in_i^g and the text feature obtained from the LSTM encoder tx_i^g and outputs a combined image-text feature dnl_i^g

$$d_i^g = conv_{1x1}(dnl_i^g) + in_i^g \tag{3.9}$$

Here d_i^g is the output of our geographic content extractor GCE which combines the feature map of the DNL block and the instance normalization vector using a convolutional layer. The result is then passed onto the geographic terrain extractor GTE module.

3.0.2 Geographic Terrain Extractor

The idea behind the terrain extractor module is to preserve geographic context such as the terrain information and what landscape is the image from like city, rural, urban, ocean, mountains, forests, etc. We borrow the idea of geometric transformations from Euclidean geometry. These are mathematical transformations that preserve lines and parallelism to handle translations, rotations, scaling, shearing, and reflections. These can help to improve the robustness of the network to variations in input images, such as changes in viewpoint, scale, or orientation.

$$GE(im_i^g, tx_i^g) = geo_i^g = G1 * d_i^g + G2 \tag{3.10}$$

3.10 shows how we apply the geometric transformation to individual channels of the output d_i^g of geographic Content extractor GCE. This is the final output from the geographic module GE. Here $G1$ and $G2$ are geometric parameters computed using simple linear layers and the mean of the image encoder representation im_i^g computed in 3.8.

4. Experimental Results and Discussion

In this section, we first discuss how we curate a new dataset from an existing publicly available remote sensing image captioning dataset. We then discuss and demonstrate the effectiveness of the proposed approach on this dataset. We use the popular Recall@K metric on the test set to evaluate and compare all methods. Recall@K is the percentage of test queries where the target image is among the top K retrieved images. We run all experiments 3 times to ensure a mean and standard deviation are reported.

4.0.1 Dataset

The GeoRic [7] dataset is a geo-aware image captioning dataset sourced from Geograph, a project that collects satellite imagery covering large parts of Great Britain and Ireland. The GeoRic dataset [7] comprises 29,038 images from the Geograph project website, each with a text caption and location coordinates. As per the authors, the captions that are only one sentence long and contain at least one spatial expression, such as "near," "north of," or "across," to ensure they contain enough geographic referencing.

To adapt the GeoRic dataset for remote sensing image retrieval, we leveraged the Google Static Maps API to scrape overhead images. The aim of this modification was to enrich the dataset with more visual content and enhance its geographic information. We achieved this by providing the API with the latitude and longitude coordinates of the image locations in the GeoRic dataset and specifying the desired zoom level and image size. The API then returned a static image of the location, which we added to our modified dataset. By doing so, we aimed to create a more comprehensive dataset that could be used for remote sensing image retrieval tasks. The modified dataset still contained the same captions and location coordinates as the original GeoRic dataset, ensuring that it remained consistent and comparable. In summary, our modification of the GeoRic dataset using the Google Static Maps API allowed us to create

Ground Image	Aerial Image	Caption
		Leafy lane near Roughton

Table 4.1: An example entry in the dataset

	Total	Train	Validation	Test
Number of captions	29,038	21,778	3,630	3,630
Number of tokens	289,028	215,883	36,409	36,736
Average caption length	9.95	9.91	10.03	10.12
Geographic entities/caption	2.05	2.04	2.05	2.06

Table 4.2: Statistics from the GeoRic dataset

a more visually informative and geographically rich dataset for use in remote sensing image retrieval research.

4.0.2 Results

In this section, we present our experimental results. First, we discuss the baseline methods used to compare our work.

The idea is to choose methods with different approaches including classical and more recent methods. We decided to compare our model with 4 methods as explained below:

- **CBIR-SVM**: Based on Content Based-Image Retrieval Using Support Vector Machine [10], this method uses the popular Support Vector Machine algorithm which classifies data samples into N classes using an optimal hyperplane and kernel.
- **ResNet50**: This is a simple unimodal baseline where we encode both ground-level and aerial images with the ResNet50 pre-trained model and retrieve the with the cosine similarity vector search.
- **ViT**: Same as above except the model is a Vision Transformer(ViT)
- **CNN-IR**: A unimodal CNN model based on CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples [8]

- **Early Fusion using ResNet50(EF)**: Early Fusion is a technique that involves utilizing the initial feature extraction layers of a target neural network as a backbone. This backbone is then duplicated and applied to both aerial and ground images. To fuse the features, a concatenation layer is applied to the low-level features, resulting in a tensor with twice the number of kernels as the original. For the implementation, we follow [6] and accordingly, concatenate the feature vectors prior to the first convolutional layer that doubles the number of kernels in the target network.
- **Late Fusion using ResNet50(LF)**: Late fusion or decision-based algorithms integrate results after each modality has made a prediction. These algorithms use decision values from each modality and combine them using various fusion mechanisms, such as averaging, voting schemes, or weighting based on specific criteria. This approach differs from early fusion, which combines features from multiple modalities before making predictions. We again use the same implementation as detailed in [6]
- **TIRG**: A multi-modal image-text retrieval model based on Composing Text and Image for Image Retrieval - An Empirical Odyssey [13]. This method uses a residual and gating module to compose image features with text features.

Here is a summary of the experimental settings for training the model. We used PyTorch to train the model and Weights and Biases API for logging all metrics and statistics. The choice of optimizer was rectified Adam, the learning rate was set to 0.0002 with a decay factor of 30 for every 25 epochs, the batch size was set to 8 and the model was trained for 100 epochs on a single NVIDIA Titan V. GPU.

Table 4.3 shows the results of all models on the modified version of the dataset. Overall TIRG performs the best across all recall@K scores except Recall@100 where our method has a better performance compared to it.

Method	**Metric**				
	R@1	**R@5**	**R@10**	**R@50**	**R@100**
CBIR-SVM	0.096	0.112	0.131	0.145	0.149
ResNet50	0.0007	0.0038	0.0079	0.0355	0.0672
ViT	0.0052	0.0155	0.0266	0.0873	0.1423
Early Fustion	0.131	0.155	0.164	0.223	0.247
Late Fustion	0.126	0.164	0.192	0.289	0.291
TIRG	**0.256**	**0.3045**	**0.3322**	**0.409**	0.498
Ours	0.2168	0.2867	0.3151	0.379	**0.512**

Table 4.3: Quantitative results of our method compared with other popular models on our modified version of GeoRic dataset

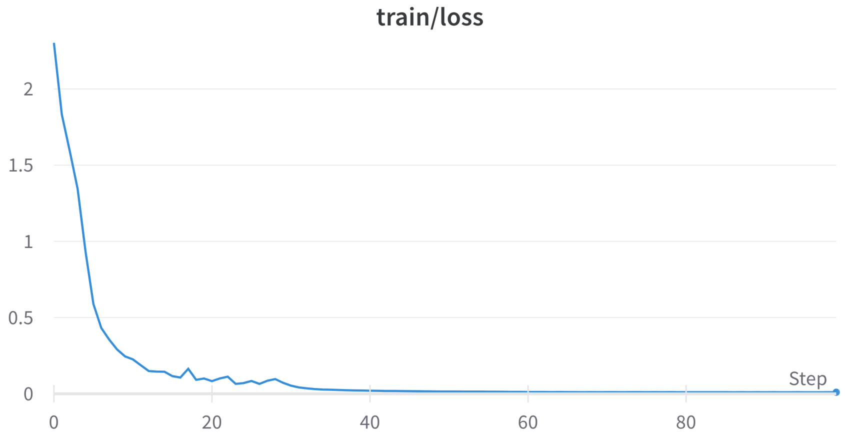


Figure 4.1: Loss

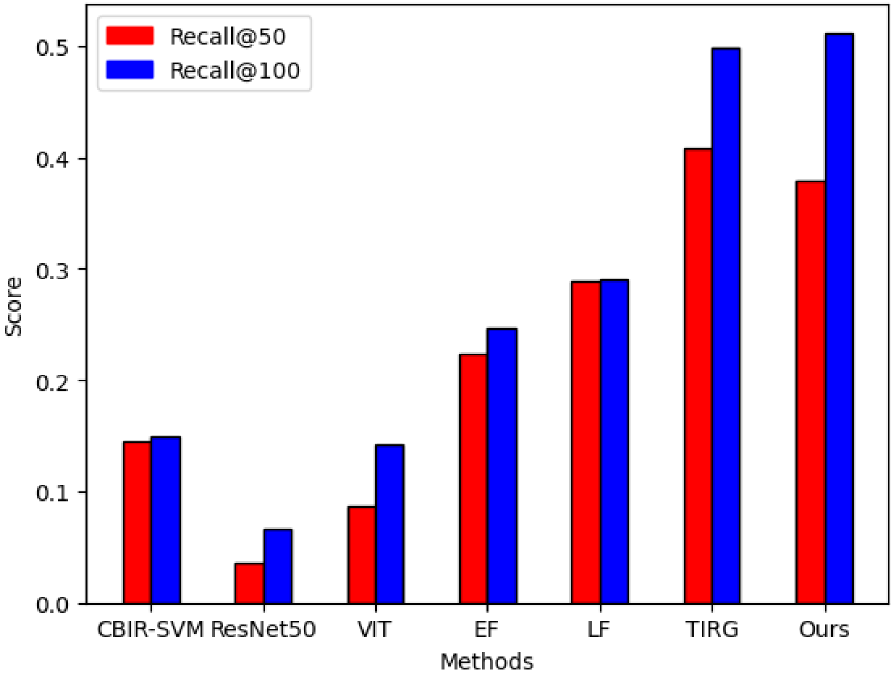


Figure 4.2: Comparison of all methods across Recall@50 and Recall@100

Discussion

Table 4.4 shows examples of some captions from the GeoRic dataset. Even though results from the previous section show that the captions can help improve the retrieval results, we can observe that captions in the GeoRic dataset do not necessarily provide the required geographic context for the image retrieval task and are not always extremely useful.

Captions
Lolham Fisheries, near Maxey.
Featureless area near Caldercruix
Autumn is well underway in Woodcote.
At Monkrigg near Haddington.
Landscape near Lowdham.

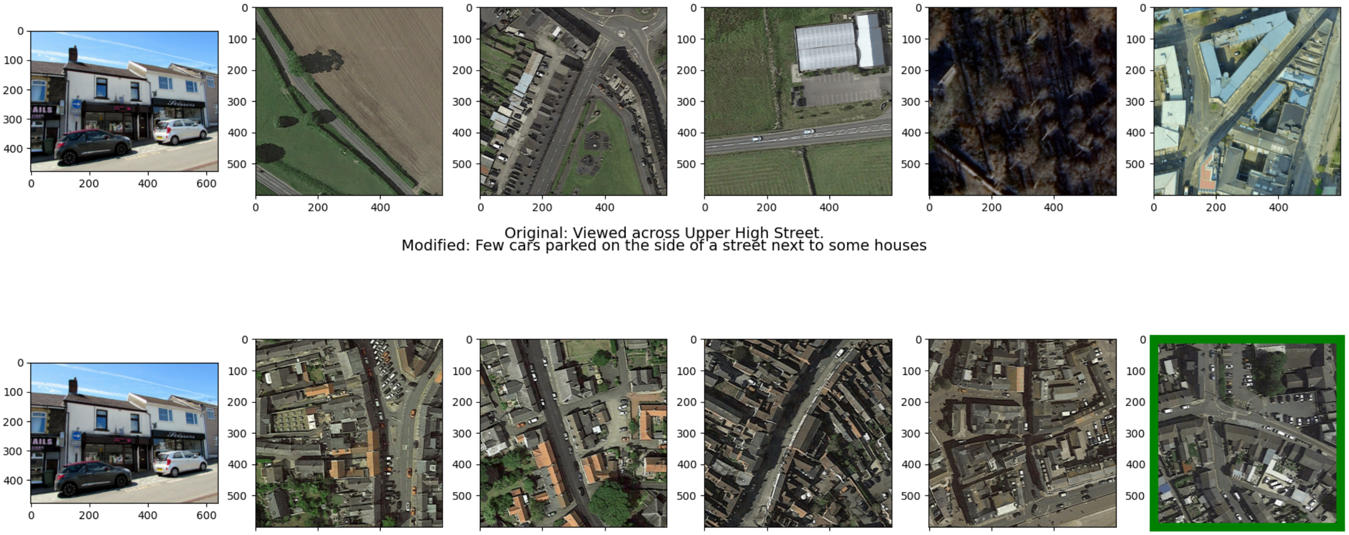
Table 4.4: Examples of captions from the GeoRic dataset

There are a few reasons for this - first, the GeoRic dataset is a remote sensing image captioning dataset meant to improve upon the generic captions by standard image caption generation systems using geographic contexts such as the location where a photograph is taken or relevant geographic objects around an image location. Second, the captions in the dataset may not be optimized for the specific requirements of the image retrieval task. For example, the captions may not include the relevant geographic features or attributes that are important for accurate retrieval. Third, these captions may not be designed to capture the unique characteristics of remote sensing satellite imagery, such as its large scale, complex terrain, and diverse environmental conditions.

In a nutshell, the captions may not provide sufficient geographic details that are necessary for accurate image retrieval. This can be a critical issue in remote sensing applications, where the geographic context of an image can significantly affect its interpretation and utility. This can lead to inaccurate retrieval results, which can be a significant problem in applications such as environmental monitoring, disaster response, and urban planning.

To address this issue, we conduct an experiment where we manually provide the captions for the test set images and compare the results of the models that use the existing captions. The goal of this experiment is to demonstrate the importance of accurate captions in cross-view image retrieval and to show that manually provided captions can significantly improve the accuracy of retrieval results. The process of manually providing the captions involves carefully analyzing each image and identifying the geographic features and attributes that need to be included in the caption. This requires extensive knowledge of the geography and terrain represented in the images, as well as an understanding of the specific features that are relevant to the retrieval task.

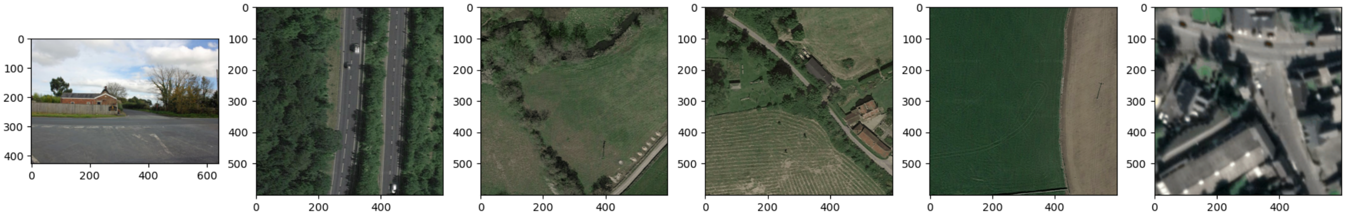
By comparing the results of the model that uses the manually provided captions with the results of the model that uses the existing captions, we can evaluate the effectiveness of the new captions in improving the accuracy of image retrieval. This will provide valuable insights into the importance of accurate captions in cross-view image retrieval for remote sensing satellite imagery and demonstrate the potential benefits of using manually provided captions to enhance the accuracy of retrieval models.



Original: Viewed across Upper High Street.
Modified: Few cars parked on the side of a street next to some houses

Original: Viewed across Upper High Street.
Modified: Few cars parked on the side of a street next to some houses

Figure 4.3: Example 1 - Top is Late Fusion ResNet50, below is Ours



Original: The junction of Willaston Road and Raby Mere Road at For Lanes End near Raby.
Modified: Field road along a farm or village with a house nearby.

Figures 4.3, 4.4 and 4.5 signify the importance of incorporating additional details about the geographic Content. The 3 examples show that modifying the existing captions can improve the retrieval results. Here we compare our method against the Late Fusion ResNet50 model. We show the top 5 retrieved images(aerial) for the ground-level image in the left corner and the original and modified versions of the captions at the bottom. The top row shows the results for the Late Fusion

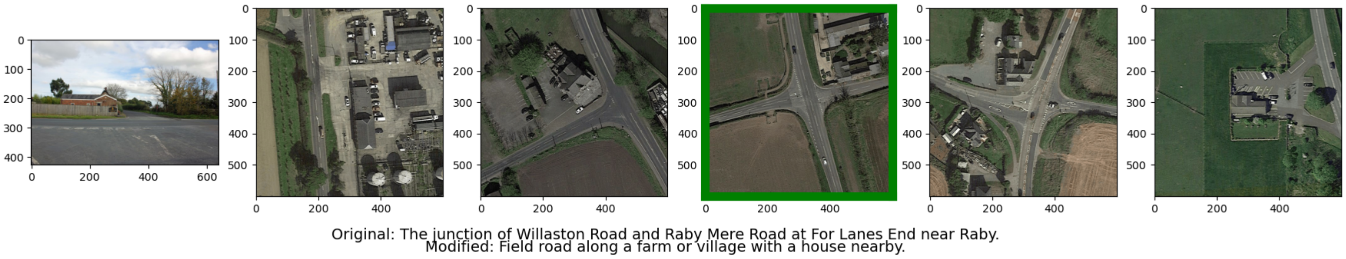

Original: The junction of Willaston Road and Raby Mere Road at For Lanes End near Raby.
Modified: Field road along a farm or village with a house nearby.

Figure 4.4: Example 2 - Top is Late Fusion ResNet50, below is Ours

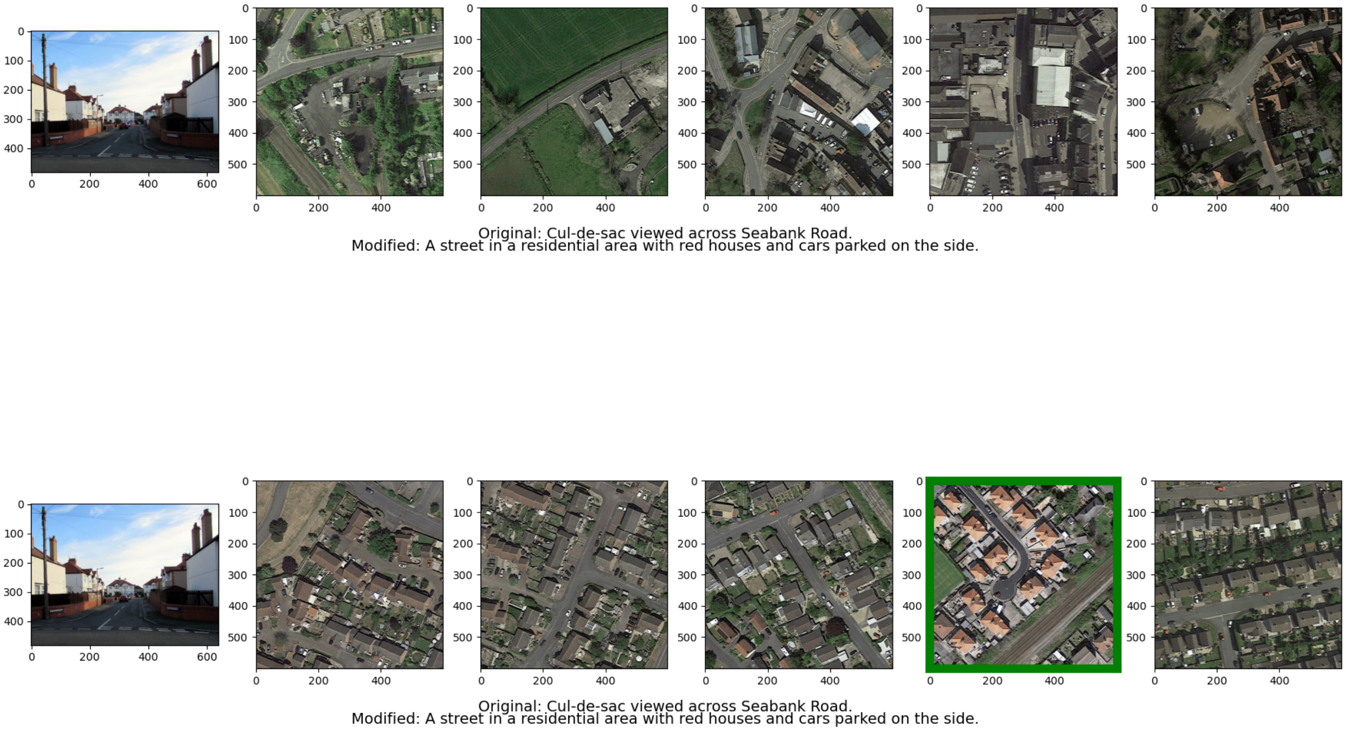

Original: Cul-de-sac viewed across Seabank Road.
Modified: A street in a residential area with red houses and cars parked on the side.
Original: Cul-de-sac viewed across Seabank Road.
Modified: A street in a residential area with red houses and cars parked on the side.

Figure 4.5: Example 3 - Top is Late Fusion ResNet50, below is Ours

model and the bottom row is Ours. In each example, our model can retrieve the correct image(highlighted with a green border) within the top 5 retrieved images whereas the Late Fusion model still fails.

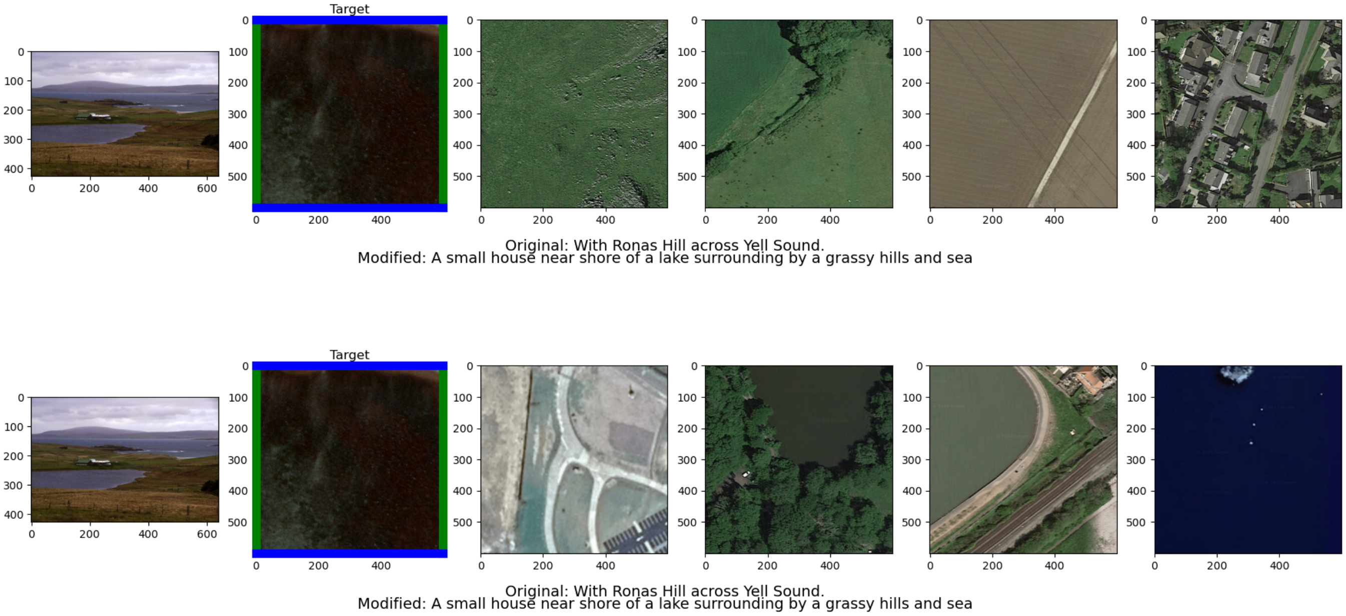

Target
Original: With Ronas Hill across Yell Sound.
Modified: A small house near shore of a lake surrounding by a grassy hills and sea
Target
Original: With Ronas Hill across Yell Sound.
Modified: A small house near shore of a lake surrounding by a grassy hills and sea

Figure 4.6: Here GT is shown in blue and green in both figures but is not retrieved by any method. The top 4 retrieved images are showed.

Overall we still see that the dataset still has many limitations. This not only pertains to the textual captions but also the aerial images extracted using the Google Static Maps API. Some of the issues are time-range duration between the ground level and the corresponding aerial image since the GeoRic dataset was built with images taken around 5-7 years ago whereas we extract the aerial images in March 2023. In addition to this, the quality of the images extracted from Google Static Maps API is often influenced by factors not in our control. This is further impacted by other parameters such as zoom level, image dimensions, and accuracy of the GPS latitude-longitude coordinates provided in the GeoRic dataset. Figure 4.6 shows such an example where the ground-level and aerial images have a complete mismatch in terms of the terrain and geographic features. As a result, none of the methods in our experimentation is able to correctly retrieve the corresponding aerial image even with the modified version of the caption. Hence, in the future, it would be beneficial to extensively devote time to improve the quality of both the aerial image and captions. Possible solutions include crowdsourcing more descriptive cations and extracting aerial imagery using a dedicated commercially licensed remote sensing imagery API such as Maxar or EOS.

5. Conclusion

In remote sensing applications, cross-view image retrieval is a challenging task due to the differences in viewpoint and scale between ground-level images and aerial/satellite images. This task has important applications in disaster response and damage assessment, as well as monitoring infrastructure and land use changes over time. However, conventional cross-view retrieval methods often rely on handcrafted features and similarity metrics, which can be limited in their ability to capture complex relationships between different views of the same location.

To address these limitations, we propose a multi-modal approach that combines both visual and textual information to enhance cross-view remote sensing image retrieval. Specifically, we utilize both ground-level images and associated textual captions to better capture the semantic content of the images. Our approach involves extracting both geographic content and terrain features from the ground-level image and text caption, which are then combined to retrieve the corresponding aerial image.

To demonstrate the effectiveness of our approach, we curate a new multi-modal image retrieval dataset based on an existing geographic image captioning dataset, GeoRic. We use the Google Static Maps API to scrape overhead imagery for the corresponding ground-level images in the dataset and then evaluate the performance of our approach against existing unimodal and multi-modal deep learning-based image retrieval methods. Our experimental results show that our approach outperforms traditional image retrieval methods and performs competitively with an image-text retrieval model in terms of retrieval accuracy.

One of the key advantages of our approach is its flexibility and potential for real-world applications. Our multi-modal framework can be applied to various types of satellite imagery retrieval tasks, such as land cover classification, and can help to overcome some of the challenges associated with remote sensing tasks. For example, the incorporation of textual captions can help to provide additional context and information about the ground-level image, which may be particularly useful in cases where images have complex and varied visual content.

Furthermore, our approach has the potential for use in other fields beyond remote sensing. Multi-modal image retrieval is a rapidly growing area of research, and there are numerous potential applications in fields such as medical imaging, art history, and cultural heritage preservation. By leveraging both visual and textual information, our approach can help to capture more comprehensive representations of images and enhance the performance of image retrieval systems in various domains.

In conclusion, this project proposes an unexplored approach to enhance cross-view remote sensing image retrieval by utilizing both images and captions. Our multi-modal framework extracts geographic features from the ground-level image and text captions and combines them to retrieve the corresponding aerial image. We demonstrate the effectiveness of our approach on a new dataset and show that it outperforms traditional unimodal image retrieval methods. The flexibility and potential for real-world applications of our approach make it a promising direction for future research in multi-modal image retrieval.

Bibliography

[1] Alaaeldin El-Nouby et al. "Training Vision Transformers for Image Retrieval". In: *CoRR* abs/2102.05644 (2021). arXiv: `2102.05644`. URL: `https://arxiv.org/abs/2102.05644`.

[2] Yang Gao et al. "Cross-view gait recognition using transfer learning from deep convolutional networks". In: *IEEE Transactions on Information Forensics and Security* 9.11 (2014), pp. 1844–1857.

[3] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: `1512.03385`. URL: `http://arxiv.org/abs/1512.03385`.

[4] Tuan Hoang et al. "Selective Deep Convolutional Features for Image Retrieval". In: *CoRR* abs/1707.00809 (2017). arXiv: `1707.00809`. URL: `http://arxiv.org/abs/1707.00809`.

[5] Lian Zhichao Jiang Zhukai and Jinping Wang. "Dual Attention Triplet Hashing Network for Image Retrieval". In: *Frontiers in Neurorobotics 15*. 2021. DOI: `https://doi.org/10.3389/fnbot.2021.728161`.

[6] Gabriel Machado et al. "AiRound and CV-BrCT: Novel Multiview Datasets for Scene Classification". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), pp. 488–503. DOI: `10.1109/JSTARS.2020.3033424`.

[7] Sofia Nikiforova et al. "Geo-Aware Image Caption Generation". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 3143–3156. DOI: `10.18653/v1/2020.coling-main.280`. URL: `https://aclanthology.org/2020.coling-main.280`.

[8] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. "CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples". In: *CoRR* abs/1604.02426 (2016). arXiv: `1604.02426`. URL: `http://arxiv.org/abs/1604.02426`.

[9] Savita, Sandeep Jain, and K.K. Paliwal. “CBIR by cascading features SVM”. In: *2017 International Conference on Computing, Communication and Automation (ICCCA)*. 2017, pp. 93–97. DOI: `10.1109/CCAA.2017.8229778`.

[10] Syazwani Izzati Shahrom et al. “Content Based-Image Retrieval Using Support Vector Machine”. In: *2021 11th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*. 2021, pp. 40–45. DOI: `10.1109/ICCSCE52189.2021.9530873`.

[11] Anuj Sharma et al. “Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images”. In: *Journal of Computer Science* 11.4 (2015), pp. 389–399.

[12] Mandikal Vikram, Aditya Anantharaman, and Suhas BS. “An Approach for Multimodal Medical Image Retrieval using Latent Dirichlet Allocation”. In: *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*. 2019, pp. 44–51.

[13] Nam Vo et al. “Composing Text and Image for Image Retrieval - An Empirical Odyssey”. In: *CoRR* abs/1812.07119 (2018). arXiv: `1812.07119`. URL: `http://arxiv.org/abs/1812.07119`.

[14] Kelly L. Wiggers et al. “Image Retrieval and Pattern Spotting using Siamese Neural Network”. In: *2019 International Joint Conference on Neural Networks (IJCNN)*. 2019, pp. 1–8. DOI: `10.1109/IJCNN.2019.8852197`.

[15] Minghao Yin et al. “Disentangled Non-Local Neural Networks”. In: *CoRR* abs/2006.06668 (2020). arXiv: `2006.06668`. URL: `https://arxiv.org/abs/2006.06668`.

[16] Yunchao Zhang et al. “A Probabilistic Analysis of Sparse Coded Feature Pooling and Its Application for Image Retrieval”. In: *PLOS ONE* 10.7 (July 2015), pp. 1–18. DOI: `10.1371/journal.pone.0131721`. URL: `https://doi.org/10.1371/journal.pone.0131721`.