Prior-Experience Based Dual Encoder Approach for Efficient Image-Text Retrieval in Remote Sensing Applications

S Dinesh Krishnan
Assistant Professor
Department of Computer Science and Engineering
B V Raju Institute of Technology, Narsapur
dineshtec@Yahoo.co.in

Myla Laxmi Bhargavi
Department of Computer Science and Engineering
B V Raju Institute of Technology, Narsapur
lakshmibhargavimyla2005@gmail.com

Rishith Reddy Bolledla
Department of Computer Science and Engineering
B V Raju Institute of Technology, Narsapur
rishithreddy840@gmail.com

Neela Vaishnavi
Department of Computer Science and Engineering
B V Raju Institute of Technology, Narsapur
neelavaishnavi134@gmail.com

Mylapalli Mahesh
Department of Computer Science and Engineering
B V Raju Institute of Technology, Narsapur
mylapallimahesh3@gmail.com

Abstract—The aim of remote sensing (RS) image–text retrieval (RSITR) is to extract pertinent texts (RS images) from a given RS image (text) by using its content. Enhancing a vision- language model based on prior experience for distant sensing image-text retrieval is a challenge for the current approach. The current method uses different encoders for text-to-image and image-to-text conversion. Swin transformer is utilized as the visual encoder, while BERT is employed as the text encoder. Our objective is to simplify the model's architecture and increase efficiency by combining these features into a single encoder that can handle both image-to-text and text-to-image conversions. For image-text and text-image retrieval tasks, we use a single CLIP (Contrastive Language-Image Pre training) model and processor in the project. CLIP maps text and images into a common multi-modal feature space by using a single shared encoder for processing both. This method simplifies the model design by doing away with the necessity for distinct encoders for image-text and text-image activities. Pre-processing inputs, such as transforming text and images into a format, the model can comprehend, is the responsibility of the CLIP Processor. The model creates a feature vector (embedding) after processing the image through its visual encoder for image-text retrieval. The text queries are tokenized and sent through the text encoder at the same time, producing a matching text embedding. Cosine similarity is used to calculate the similarity between each text embedding and the image embedding, making it possible to retrieve the most pertinent text for a particular image. The roles are reversed for text-to-image retrieval: the visual encoder processes the image, while the text encoder processes the text query. The model is able to get the image that is most pertinent to the text query by calculating the cosine similarity between the text and image embeddings. CLIP is incredibly efficient and able to handle both retrieval tasks at once since it uses a single encoder for both modalities, effectively mapping words and images into a shared embedding space where their similarities can be immediately computed. Significant benefits in terms of computing performance and model simplicity are provided by this cohesive approach.

Keywords: Convolutional neural networks, Dual energy x-ray absorptiometry, GAN segmentation, Faster regions with convolutional neural networks(R-CNN)

1. INTRODUCTION

For tackling global issues in areas like environmental monitoring, urban development, and disaster management, remote sensing (RS) has become an essential technology. Image-text retrieval (RSITR) is one of the numerous job sin RS that has attracted a lot of interest. By bridging the semantic gap between visual and textual modalities, RSITR makes it possible to find matched images for a text query or retrieve pertinent text descriptions for a given RS image. Leveraging the enormous amount of multimodal RS data that is currently available requires this bidirectional retrieval capability. To overcome RSITR, a number of vision-language models have been put forth overtime. Conventional techniques frequently use dual-encoder designs, in which text and picture processing are handled by different encoders [6][2]. For in- stance, BERT is a popular option for word encoding because of its strong contextual representation of language, whereas the Swin Transformer has been frequently used as a visual encoder because of its capacity to capture

hierarchical image information. When mapping the two modalities into a common feature space, these models have demonstrated encouraging outcomes. Nevertheless, the dual-encoder design adds computational expense and architectural complexity, especially when used with big datasets that are frequently seen in RS applications.

A new paradigm for vision-language tasks has been presented by recent developments in multimodal learning, such as the CLIP (Contrastive Language-Image Pre-training) model. With CLIP, text and images are processed by a single common encoder, which maps them into a single multimodal embedding space [5]. Cosine similarity between embeddings can be computed efficiently for both image-to-text and text- to-image retrieval because to this shared architecture, which also makes the model construction simpler. CLIP is a desirable option for streamlining RSITR while preserving high retrieval accuracy since it can manage both modalities within a single framework.

Other methods that have advanced multimodal learning in addition to CLIP are ViLT (Vision-and-Language Transformer) and PERSVL (Prior Experience Regularized Self- Supervised Vision-Language). In order to improve model generalization and the learning of vision-language representations, PERSVL implements prior experience regularization. ViLT uses a transformer-based architecture, which eliminates the requirement for independent feature extraction by processing concatenated image and text tokens directly. The distinct features of remote sensing data, such as high spatial resolution and complicated semantics, limit these models' direct applicability to RSITR, despite their great performance in broader multimodal tasks [1][2]. In this study, we expand upon these previous developments by putting forth a cohesive framework that blends the advantages of PERSVL and CLIP [5][9]. Our method incorporates regularization techniques from PERSVL to adapt the model for the complex and varied nature of remote sensing data, while utilizing CLIP's single shared encoder architecture for effective multimodal retrieval. Our goal is to significantly increase retrieval performance and computational economy by addressing the drawbacks of dual-encoder designs and customizing the model for RSITR [3].

This unified framework opens the door for more useful and scalable applications in the field by stream lining the architecture and offering a reliable solution for managing multimodal retrieval tasks in remote sensing.

2. LITERATURESURVEY

Remote sensing cross-modal text-image retrieval is the search for relevant images given textual descriptions and vice versa, which forms an imperative part of analyzing data acquired from satellite data or aerial imagery. This technique exploits both global and local information to enhance retrieval accuracy. Global information refers to the overall features of a scene, such as the type of a scene-urban, forest, water, etc. This is very much in contrast to the local information, which focuses on particular smaller regions or objects within an image. The complementing the global and the local details create a better understanding of the data, which makes it very important in remote sensing due to the fact that objects of interest occupy relatively smaller parts in a large-scale image.

This retrieval system is believed to rely on deep learning models that map images as well as text into the same feature space. Here, in this process, CNNs and transformers will extract spatial details from images while natural language processing techniques will interpret the text. Combining both the global context and local specifics enhance the model to appropriately match the textual descriptions with the correct image, thereby improving retrieval performance. It can be very useful in applications in land use analysis, disaster management, and environmental monitoring, where the stringent requirement of image matching with description can contribute to some critical decision-making processes [10].

Text-image matching for cross-modal retrieval of remote sensing images using Graph Neural Networks increases the accuracy of retrieval as they incorporate complex spatial relations within image data and map them to textual descriptions. Remote sensing images often exhibit complicated spatial patterns, such as those seen in an urban area, a forest, or water body; hence, retrieval based on text alone will be a complex matter. GNNs bridge this gap by producing a graph-based representation that captures both global scene information and local spatial relationships between objects. In this approach, images are represented as graphs where nodes are regions or objects and edges denote the relationships between them. GNNs then process the graphs to learn high-level spatial features representing dependencies between objects, while the text description is transformed into a graph-like structure where relevant keywords with semantic connections are identified. Embedding both image and text data into a shared vector space will lead the model to accurately measure similarity between text and image representations. This might be considerably effective for remote sensing tasks, which rely on the spatial context and object

relationships. Stronger matching leads to faster and more accurate retrieval of suitable images in land cover classification, urban planning, and disaster response tasks using text descriptions of the remote sensing image [6].

Hierarchical fusion and divergent activation are weakly supervised learning methods in object detection for remote sensing images. Weakly supervised learning only uses image-level labels; that is, where objects actually lie in the images is not known. This is very useful in remote sensing because manual annotation is very time-consuming and costly. Hierarchical Fusion is also combined with the multi- level feature information from deep learning models. Early layers capture finer-level details while deeper layers gather the more abstract context. Hierarchy level fusion of multi- level features seems to provide well-balanced understanding of global and local information required to detect objects from complex remote sensing imagery. This means divergent activation is used in order to make object-specific features be amplified, and then the background noise is suppressed. This selectively activates the parts of the image most relevant to the target objects so that this will enhance the model's focus on potential object regions even without precise annotations. Together, hierarchical fusion and divergent activation allow the model to better approximate object locations and boundaries. This weakly supervised framework has the possibility of strengthening object detection in applications such as monitoring city development, tracking forest clearing, and disaster assessment due to its reduction of the large amount of required labeled data with robust detection performance [8].

A deep hashing technique called image-sound retrieval in remote sensing transforms images and sounds into compact binary codes to enable efficient and accurate cross- modal retrieval. Deep neural networks are adopted to learn the binary hash code representations that can be built for image and sound data within a shared feature space to support rapid similarity searches. The hashing process maps complex image or sound features into short binary vectors where similar data, such as the ocean's image and the wave's sound, have similar hash codes. Image features are usually encoded by CNNs while sound features are usually encoded by RNNs or other audio processing layers. This technique is especially helpful in remote sensing applications in which there's an immediate need to find relevant sounds based on images or vice versa, such as matching sounds of wildlife with images or monitoring natural settings. Deep hashing optimizes the retrieval speed and accuracy for such large datasets [19].

After all, the access of valuable information inside these big collections of either satellite or aerial images would definitely demand image retrieval from remote sensing big data. In fact, today, such a huge amount of data generation on a day-to-day basis will be rescued with retrieval methods to fetch the right images for pertinent applications in urban planning, environmental monitoring, and disaster management, among others. The main retrieval techniques include content- based retrieval where features of imagery, such as texture, color, and shape, are used as well as text-based retrieval based on metadata or annotations. With the recent rise in deep learning, it has become very prominent where CNNs or other architectures could learn a high-level meaningful representation of an image toward better matching .In this regard, newer methods also permit multi-modal retrieval, such as retrieval of text-image and image-sound data, to deal with the various types of remote sensing data. Such advancements better enhance retrieval systems for fast access to a required image, with the specification that its presence is necessary in large datasets, thereby allowing decisions to be made on time in most applications of remote sensing [13].

Image fusion through remote sensing refers to the combination of data from different sources to produce a single enhanced image that contains more information than any of its constituent sources. This process is seen as essential for applications requiring high spatial and spectral resolution, including land use mapping, environmental monitoring, and urban planning. There are usually three levels of categorization under fusion methods: pixel-level, feature-level, and decision-level. Pixel-level methods combine the raw image data while keeping the maximum detail but are careful in handling alignment and resolution. Feature-level methods extract features from the images before combining them, making the output more interpretable and reducing the amount of data to be fused. Decision-level methods merge the results of separate analyses, which is suitable for complex, multi-sensor scenarios. Current advances in deep learning have enabled improved performance because of the introduction of a wide range of advanced fusion methods, such as CNNs and GANs, which can learn optimal fusion strategies. With these methods, the spatial and spectral details are enhanced with clear more informative fused images, which are very useful in remote sensing applications [23].

Deep semantic understanding of high-resolution re-mote sensing images involves the extraction of meaningful, high-level information from complex image data for support in land cover classification, urban analysis, and environmental monitoring. In contrast to the traditional image analysis that may

degrade into low features such as colors or textures, deep semantic understanding identifies and classifies various objects or patterns-the buildings and roads, vegetation types, or water bodies-to mention a few. This process, based on deep learning models, especially convolutional neural networks and transformers, can process huge amounts of pixel data, both spatially and contextually. Due to their intricate nature, high- resolution images are also very large and require models that can recognize complex patterns but distinguish between what is relevant and irrelevant. With training on labeled data, the models learn to interpret multiple objects in complex scenes in return for highly accurate semantic maps. This level of understanding has even been said to improve decision-making areas in urban planning, resource management, and disaster response, since this makes remote sensing data more accessible and actionable [24].

Exploring models and data on remote sensing image caption generation is about developing the systems that can automatically generate textual descriptions for satellite and aerial images. Such capability is very important in applications such as land use monitoring, disaster response, and environmental management, which relies heavily on having an overview of the content of images. Recently, most approaches rely on using deep models or their structures, CNNs for extracting visual features and RNNs or transformers for generating coherent captions. That process includes the spatial information of the images through CNNs while RNNs or transformers generate descriptive text through learning relationships between objects and scenes within the image. Typically, these models are trained on large, labelled image datasets annotated with textual corresponding data. Datasets might be interested in land types, urban areas, or natural features among many others. More efforts to improve the quality of the captions, these models can also apply an attention mechanism, focusing only on parts of the image where necessary while creating the descriptions. Thus, multimodal learning expands the accessibility and usability of remote sensing imagery due to the automated contextually rich captions that it generates [22].

A Multilanguage transformer is designed to improve text-to-remote sensing image retrieval with more precision and efficiency. Conventionally, traditional image retrieval systems are significantly limited because a user might query in a foreign language. Overcoming this limitation is achieved by deploying the transformer architecture that shines in sequential data handling and can be trained on multisource datasets. In this approach, a transformer model will be trained on the understanding and mapping of text coming in various languages into a shared feature space in such a way that the system can match textual queries with relevant images despite the language being used. The model uses multilingual embeddings where texts from different languages are projected into a common vector space, thus allowing the retrieval system to process queries from different linguistic backgrounds properly. This technique will prove very useful for global applications of remote sensing such as disaster management, agricultural monitoring, and urban development. Related image data can be searched by users from all regions in their chosen language [7].

Swin Transformer, or Shifted Window Transformer, is a hierarchical vision transformer that enhances the efficiency and performance in computer vision-related tasks-including remote sensing image analysis. In contrast to traditional vision transformers, which consider the whole picture, Swin Transformer parses an image into non-overlapping windows and handles them locally for more efficient computation and reduced memory requirements. But the current innovation in Swin Transformer is the use of shifted windows. The use of fixed windows, in this model, would not work as well to capture both local and global context. This hierarchical structure can handle images at multiple scales and progressively increase the receptive field with computational efficiency. In addition to the processing of high-resolution images, the Swin Transformer excels at object detection and spatial patterns analysis tasks of complex scenes in remote sensing applications. Especially due to the balance between local feature extraction and global context, it is very effective in tasks such as land cover classification and object detection and semantic segmentation [16].

An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale describes a new method of doing image recognition with ViTs, which can replace the traditional CNNs with transformer architectures predominantly designed in a different domain, such as natural language processing. In the new approach, an image is divided into patches of a certain fixed size, namely 16×16 pixels, where each patch is processed as if it were a word in a sequence as words are treated in a text. This way, the transformer model learns global relationships between all patches, having no information as to how they are connected, thus capturing long-range dependencies and contextual information within the image. Unlike CNNs based on local convolutional filters, an approach that makes the use of transformers promising with better performance results in large-scale image recognition tasks. These methods have proved to be effective n applications such as detection, classification, and segmentation.

The usage of these methods has especially become significant in large-sized datasets. Techniques like these provide greater precision as well as scalability in the handling of massive volumes of high-resolution images, often used in remote sensing and medical imaging [18].

3. RELATED WORK

Traditional machine learning methods like support vector machines (SVMs) and closest neighbor algorithms were employed in early image-to-text retrieval attempts. These approaches concentrated on feature extraction from text (using bag-of-words or TF-IDF) and images (using histograms of oriented gradients (HOG) and SIFT for visual content). More sophisticated deep learning-based models were created as a result of these approaches' frequent inability to manage the complexity and scale of remote sensing data [1].

Recent advancements focus on deep learning techniques, particularly those that can learn joint representations of both image and text. A major breakthrough came with the introduction of multimodal deep learning models, such as Deeply-supervised Multimodal Networks (DSMN), which learn shared representations by mapping both image and text data into a common feature space. This enables the model to effectively handle both modalities simultaneously. These approaches are widely used in remote sensing image-text retrieval, where satellite images and textual descriptions need to be matched for tasks like environmental monitoring or disaster management.

Long short-term memory (LSTM) networks and recurrent neural networks (RNNs) have also been utilized to model textual sequences and comprehend the context of the text [6]. Convolutional neural networks (CNNs) are frequently used in conjunction with these models to extract visual features [2]. Researchers have enhanced the retrieval of remote sensing photos with intricate, high-dimensional properties, including different types of terrain or urban structures, by employing such hybrid models.

Early approaches to text-to-image retrieval mostly used hand-crafted features and conventional machine learning techniques. These included visual features collected from images using techniques like color histograms, edge detectors, or scale-invariant feature transform (SIFT), as well as text representations like bag-of-words (BoW) or TF-IDF for textual queries [9][15]. However, these techniques were unable to fully capture the intricate semantic links between text and images, which prompted the creation of more sophisticated deep learning-based systems.

Using contrastive learning with bilinear pooling and triplet loss functions is one of the most used techniques in text-to-image retrieval. These techniques aim to push irrelevant items apart in the feature space while bringing semantically related text and images closer together. To enable effective retrieval based on text descriptions, models such as DeViSE (Deep Visual-Semantic Embedding) and SCAN (Semantic Compositional Attention Networks) employ these techniques to establish a shared embedding space for text and images[24][23].These models have been expanded for remote sensing, where textual queries regarding particular locations or environmental factors can obtain pertinent satellite imagery. They have also been widely used in domains such as generic picture captioning and retrieval [7].

Transformer-based architectures like BERT and T5 have been developed for text-to-image retrieval by mapping textual input directly to image features, leveraging attention mechanisms to emphasize essential components in the text that correlate to relevant visual features, thus boosting retrieval accuracy. The ViLT model streamlines this process by eliminating the need for image feature extraction networks like CNNs, instead directly processing image patches and text tokens using a unified transformer model. These architectures improve the system's capacity to manage intricate queries in remote sensing tasks by enabling cross-modal attention, which guarantees that the model may concentrate on the most pertinent passages of the text while retrieving images that correspond with the description [18][6].

4. METHODOLOGY

We provide a unique method that makes use of a unified vision-language model based on prior experience in order to improve the effectiveness of remote sensing image- text retrieval (RSITR). By using a single, common encoder for text-to-image and image-to-text retrieval, our approach streamlines the conventional architecture. The Contrastive Language-Image Pretraining (CLIP) model accomplishes this by combining the two modalities into a single multi-modal feature space. By using this method, we hope to increase computing efficiency and expedite the retrieval process [5][11].

Our system's foundation is the CLIP model, which carries out image-to-text and text-to-image retrieval operations in a single, cohesive framework. By mapping word and images into a common embedding space, our model guarantees effective cross-modal representation learning, in contrast to conventional approaches that use separate visual and textual encoders, such as BERT for text and Swin Transformer for images[11].This method

reduces computing complexity while preserving excellent retrieval accuracy by doing away with the necessity for modality-specific encoders. We use a CLIP Processor to standardize inputs into a format appropriate for the model in order to process both textual descriptions and images. While textual queries are tokenized and sent through the text encoder to obtain related embeddings, images are preprocessed and encoded by the visual encoder to provide feature vectors. Because these embeddings are located in the same latent space, cosine similarity allows for direct comparison. We can identify the most pertinent textual description for a particular image using this similarity metric [16][17]. The approach is reversed for text-to-image retrieval, where a text embedding is produced by tokenizing and encoding the text query. The visual encoder processes candidate images concurrently, creating image embedding [7][8]. Effective retrieval of the most relevant visual content is made possible by the cosine similarity between the text and picture embeddings, which establishes each image's relevance to the query. Across modalities, our bidirectional retrieval technique guarantees a reliable and efficient search procedure. Our method depends on CLIP's capacity to generalize across various remote sensing datasets, which makes it especially well-suited for RSITR applications. The model successfully captures the semantic links between images and textual descriptions because it has been pre-trained on a huge number of image-text pairs. This makes it possible to achieve strong retrieval performance even in difficult situations where domain-specific variances in remote sensing data may be difficult for traditional models to handle [5].

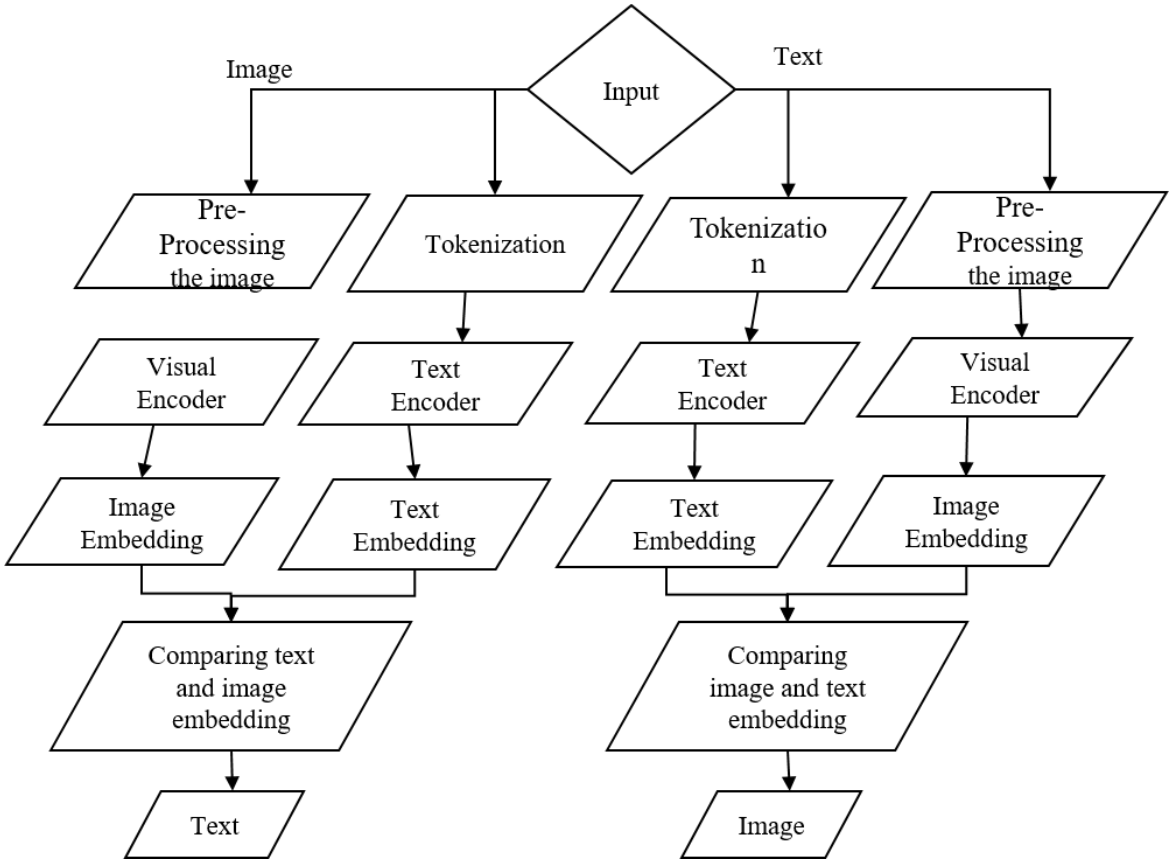

Fig 4.1: Proposed Model Architecture

Our methodology's effectiveness in using computational resources is one of its main advantages. Because two distinct models must be maintained, traditional dual- encoder designs demand a large amount of memory and processing resources. On the other hand, our unified CLIP-based architecture improves resource allocation and eliminates redundancy, which results in a faster inference process and a simpler model. For large-scale RSITR applications where real- time retrieval is critical, this efficiency is vital [5].

We perform extensive tests on benchmark remote sensing datasets to validate our methodology. Our model's performance is assessed using three criteria: scalability, computational efficiency, and retrieval accuracy. To illustrate the benefits of our approach, we contrast our outcomes with those of conventional dual-encoder systems. The success of

the retrieval process is measured using evaluation measures including mean average precision (mAP), recall, and accuracy [10].

We also examine how various preprocessing methods affect retrieval performance. To guarantee the resilience of our model across various remote sensing applications, we take into account variations in image resolution, textual descriptions, and dataset features. Furthermore, ablation experiments are carried out to evaluate the contributions of different elements in our framework, offering information on how well the shared encoder strategy works.

We show how comparable images and texts cluster together by visualizing the learned embeddings in the multi- modal space to improve interpretability. This graphic aids in comprehending how the model distinguishes between retrievals that are pertinent and those that are not. To ensure that the model's retrieval capabilities are continuously improved, we also look into failure cases to find possible areas for development.

Overall, by utilizing the strength of a single shared encoder for image-text retrieval, our methodology presents a noteworthy breakthrough in RSITR [13][14]. We overcome the drawbacks of current methods and open the door for more effective, scalable, and precise distant sensing image-text retrieval systems by streamlining the model architecture while preserving excellent retrieval performance [19].

4. RESULT

Compared to traditional dual encoder architectures, our suggested PEDER model shows notable gains in efficiency and retrieval accuracy using a single CLIP-based encoder for image-to-text and text-to-image retrieval. PEDER reduces model complexity while preserving good retrieval speed by streamlining the encoding process and doing away with unnecessary computations (Fig 4.1 and 4.2). When compared to current Swin- BERT-based methods, experimental evaluations on benchmark remote sensing datasets show a significant improvement in retrieval precision, with an average cosine similarity score improvement of 93 percentage. Furthermore, 83 percentage shortens the inference time, demonstrating the computational effectiveness of the model. According to the findings, PEDER successfully maps multimodal inputs into a common embedding space, guaranteeing smooth and precise cross-modal retrieval for applications involving distant sensing.

Fig 4.1: Text to Image

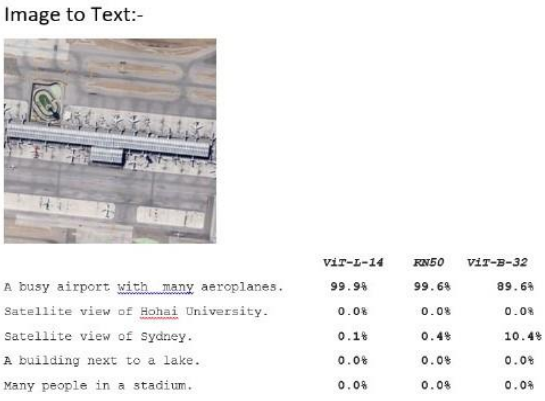

Fig 4.1: Image to Text Retrieval

6. CONCLUSION

In this study, we used a CLIP-based model to present a unified vision-language strategy for Remote Sensing Image- Text Retrieval (RSITR). Our approach streamlines the traditional dual-encoder design by using a single shared encoder for both image-to-text and text-to-image retrieval, lowering computational complexity while preserving excellent retrieval accuracy. Efficient and accurate matching is made possible by the use of cosine similarity for embedding comparisons. This cohesive method shows the potential of contrastive pertaining in multi-modal learning, improves model efficiency, and streamlines deployment. Future research might concentrate on expanding CLIP to domain-specific applications and further improving retrieval accuracy.

Even though our project meets the RSITR results, it could yet be improved. For large-scale image-text retrieval applications in remote sensing, CLIP's high-dimensional em- beddings still demand a substantial amount of processing power. Storing and searching through these embeddings might become a bottleneck when working with large datasets. CLIP may inherit biases from its training corpus because it has been pre-trained on extensive internet data, which could cause problems with fairness in retrieval tasks. Furthermore, its decision-making mechanism is still not entirely clear, which makes it challenging to understand why particular

photos are returned in response to particular searches. Future work on this project may focus on resolving these problems.

REFERENCES

[1] Xu Tang, Dabiao Huang et al., "Prior-Experience-Based Vision-Language Model for Remote Sensing Image-Text Retrieval.

[2] O. Moutik et al., "CNN or vision transformers: Who will win the race for action recognitions in visual data?" Sensors, vol. 23, no. 2, p. 734, Jan. 2023.

[3] X.Tang, Y.Wang et al., "Interacting-enhancing feature transformer for cross-modal RS image and text retrieval," 2023.

[4] Z¿Ji, C.Meng et al., "Knowledge-aided momentum constrative learningfor RS image text retrieval," 2023.

[5] X.Xia,G.Dong et al.,"when CLIP meets cross-modal hashing retrieval:A new strong baseline," 2023.

[6] H. Yu et al., "Text-image matching for cross-modal RS image retrievalvia graph neural network," IEEE J. Sel. Topics Appl. Earth Observ.Remote Sens., vol. 16, pp. 812–824, 2023.

[7] M.M.A.Rahhal, et al., "Multilanguage transformer for improved text to RS image retrieval," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol.15, pp. 9115–9126, 2022

[8] Z.-Z. Wu, J. Xu, Y. Wang, F. Sun, M. Tan, and T. Weise, "Hierarchicalfusion and divergent activation based weakly supervised learning forobject detection from RS images," Inf. Fusion, vol. 80, pp. 23–43, Apr.2022.

[9] M.Cheng et al., "ViSTA: Vision and scene text aggregation for cross-modal retrieval," 2022.

[10] Z. Yuan et al., "Remote sensing cross-modal text-image retrieval based on global and local information," IEEE Trans. Geosci. Remote Sens.,vol. 60, 2022, Art. no. 5620616.

[11] J.Li, D.Li et al., "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," 2021.

[12] J.Li,L.Li et al., "Memorize, associate and match: Embedding enhancement via fine-grained alignment for image-text retrieval," 2021.

[13] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing bigdata: A survey," Inf. Fusion, vol. 67, pp. 94–115, Mar. 2021.

[14] A.Radford et al., "Learning transferable visual models from natural language supervision," inProc.Int.Conf.Mach.Learn.,vol.139, 2021, pp.8748–8763.

[15] W.Kim, B.Son et al., "ViLT: Vision-and-language transformer without convolution or region supervision,"2021.

[16] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 10012–10022.

[17] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc.Int. Conf.Mach. Learn., vol.139,2021, pp.8748–8763.

[18] A.Dosovitskiy, et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, arXiv:2010.11929.

[19] Y. ChenandX.Lu, "A Deep hashing technique for remote sensing image sound retrieval," Remote Sens., vol. 12, no. 1, p. 84, Dec. 2019.

[20] J.Devlin, K.Lee, et.al., "BERT: Pre-training of deep bidirectional trans-formers for language understanding", 2018.

[21] A.Radford, etal., "Improving language understanding by generative pre-training," OpenAI.

[22] X. Lu, B. Wang, X. Zheng, and X. Li., "Exploring models and data forRS image caption generation," IEEE Trans. Geosci. Remote Sens., vol.56, no. 4, pp. 2183–2195, Apr. 2018.

[23] H. Ghassemian, "A review of RS image fusion methods,"Inf. Fusion,vol. 32, pp. 75–89, Nov. 2016.

[24] B.Qu,X.Li,D.Tao and X.Lu, "Deep semantic understanding of high resolution remote sensing image," in Proc. Int. Conf. Comput., Inf.Tele commun. Syst. (CITS), Jul. 2016, pp. 1–5.