remote sensing

MDPI

Article

RS-LLaVA: A Large Vision-Language Model for Joint Captioning and Question Answering in Remote Sensing Imagery

Yakoub Bazi [1,*], Laila Bashmal [1], Mohamad Mahmoud Al Rahhal [2], Riccardo Ricci [3] and Farid Melgani [3]

[1] Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia; 439204359@student.ksu.edu.sa
[2] Applied Computer Science Department, College of Applied Computer Science, King Saud University, Riyadh 11543, Saudi Arabia; mmalrahhal@ksu.edu.sa
[3] Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy; riccardo.ricci-1@unitn.it (R.R.); farid.melgani@unitn.it (F.M.)
* Correspondence: ybazi@ksu.edu.sa

Abstract: In this paper, we delve into the innovative application of large language models (LLMs) and their extension, large vision-language models (LVLMs), in the field of remote sensing (RS) image analysis. We particularly emphasize their multi-tasking potential with a focus on image captioning and visual question answering (VQA). In particular, we introduce an improved version of the Large Language and Vision Assistant Model (LLaVA), specifically adapted for RS imagery through a low-rank adaptation approach. To evaluate the model performance, we create the RS-instructions dataset, a comprehensive benchmark dataset that integrates four diverse single-task datasets related to captioning and VQA. The experimental results confirm the model's effectiveness, marking a step forward toward the development of efficient multi-task models for RS image analysis.

Keywords: remote sensing (RS); large language models (LLMs); Large Language and Vision Assistant Model (LLaVA); instruction tuning; captioning; visual question answering (VQA)

check for updates

Citation: Bazi, Y.; Bashmal, L.; Al Rahhal, M.M.; Ricci, R.; Melgani, F. RS-LLaVA: A Large Vision-Language Model for Joint Captioning and Question Answering in Remote Sensing Imagery. *Remote Sens.* **2024**, *16*, 1477. https://doi.org/10.3390/rs16091477

Academic Editor: Hossein M. Rizeei

Received: 6 March 2024
Revised: 8 April 2024
Accepted: 17 April 2024
Published: 23 April 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Remote sensing (RS) represents a vital source of information for observing the Earth's surface. Driven by advancements in satellite and aerial technologies, the volume of Earth observation data has experienced exponential growth, creating an urgent demand for the development of sophisticated analysis strategies. However, traditional RS visual analysis techniques, such as scene classification and semantic segmentation, while useful, often struggle to capture the complexity embedded within RS scenes due to their limited expressivity and interactive capabilities. Natural language, with its inherent semantic richness, captures not just objects and properties within the scene but also their intricate relationships, offering a precise, and human-centric perspective on RS data analysis. Recognizing this potential, researchers are increasingly turning their attention to the application of natural language processing (NLP) techniques [1]. These techniques have notably enhanced the analysis of RS data, leading to enhanced efficiency, accuracy, and accessibility.

The RS community has already made significant strides in utilizing NLP's potential through tasks like image captioning [2] and visual question answering (VQA) [3]. Image captioning algorithms automatically generate human-like descriptions of RS scenes, while VQA enables machines to answer natural language questions based on visual information present in the RS scene. Beyond image captioning and VQA, the RS community has explored various tasks that harness the potential of NLP, such as text-based image retrieval [4], RS image generation [5–7], visual grounding [8], change captioning [9,10], change VQA [11] of multi-temporal images, and even the generation of natural language questions based on visual cues [12].

Despite the progress in the field, current research efforts often rely on designing and training separate models for each task. This approach overlooks the potential commonalities among tasks as well as the shared information across datasets. Extracting meaningful insights from RS images demands innovative approaches that can go beyond single-task analysis. Performing multiple joint tasks has several advantages over single-task models. It improves efficiency by reducing the development and resource burden of training dedicated models for each task. A user may, for instance, desire both a natural language description of an RS image and relevant information extracted through VQA, all from a single model. Additionally, building one model that can jointly perform multiple tasks is particularly important for domains with scarce annotated datasets like RS and has many advantages over task-specific models in reducing the risk of over-fitting. However, handling multiple tasks with one model poses its own challenges, such as ensuring the accurate and reliable results across diverse tasks.

In recent times, the field of NLP has witnessed a remarkable surge in the development of large language models (LLMs), exemplified by prominent examples like ChatGPT [13]. These models, equipped with billions of parameters, demonstrate exceptional capabilities in comprehending and generating text that closely resembles human language. They excel in various linguistic tasks, such as text generation, translation, summarization, and question answering. Their proficiency in multi-tasking stems from their comprehensive understanding of language patterns and their ability to generate human-like text in various contexts. Leveraging the multi-tasking strengths of LLMs offers promising opportunities for efficient and insightful applications in RS domains.

While LLMs demonstrate mastery in text processing and generating, their counterparts, large vision-language models (LVLMs) such as and GPT-4 [14] and the open-source LLaVA [15] further enhance this capability by combining vision and language processing. LVLMs can seamlessly integrate visual information with natural language understanding and generation, enabling a holistic comprehension of both visual and textual data. This ability empowering them to tackle complex tasks such as image captioning and VQA and opens up new possibilities in RS domains, where the fusion of visual and language understanding yields valuable insights and efficient solutions. However, despite the impressive capabilities of LVLMs in the general domain, their performance tends to be suboptimal when applied to RS data. This performance gap stems from fundamental differences between RS images and natural images, which can be attributed to the high resolution, diverse scales, and unique acquisition angles of RS images. As a result, the interpretations provided by LVLMs may lead to inaccurate or even fabricated interpretations when faced with RS-specific queries. An additional challenge lies in the scarcity of a comprehensive instruction dataset specifically designed for the RS domain. Such dataset is crucial for effectively customizing LVLMs for RS applications through instruction tuning. Thus, in this paper, we present Remote Sensing Large Language and Vision Assistant (RS-LLaVA), a multi-modal model specifically tailored for RS image analysis. RS-LLaVA accepts an RS image and text as inputs and jointly performs image captioning or VQA. The model is trained in a two-step process, pre-training and fine-tuning through low-rank adaptation (LoRA) [16]. In the pre-training step, the layer that connects between the image encoder and the language decoder is pre-trained. Then, RS-LLaVA is fine-tuned through the LoRA approach. In this way, the model integrates RS image understanding with language processing, enabling it to excel in both captioning and VQA tasks in the RS domain. To rigorously train RS-LLaVA's, we developed a multi-tasking instructional dataset. The dataset is constructed by blending various captioning and VQA datasets, and it is further enhanced by formatting them as training instructions. Experimental results demonstrate that RS-LLaVA outperforms previous state-of-the-art methods in both single-task and multi-task scenarios.

Specifically, the main contributions of this paper can be summarized as follows.

(1) We propose RS-LLaVA based on the LLaVA model [15], a large vision-language model that jointly performs captioning and question answering for RS images. The model is specifically adapted for RS data through LoRA fine-tuning.

(2) We develop the RS-instructions dataset, a multi-task instruction-following dataset by integrating diverse image-text pairs from captioning and VQA datasets.
(3) We demonstrate the RS-LLaVA's effectiveness in multi-task mode compared to single-task state-of-the-art models. This model marks a promising step towards developing universal, multi-task models for RS data analysis.

The outline of this paper is as follows. The related works are introduced in Section 2. The RS-LLaVA model is explained in detail in Section 3. Section 4 presents the proposed RS-instructions dataset. Section 5 displays the experimental results. Finally, the conclusions are summarized in Section 6.

2. Related Works

2.1. NLP in Remote Sensing

Language integration in RS has showcased impressive capabilities across various tasks, including image captioning [2,17–28], VQA [3,29–32], and text–image retrieval [4]. A comprehensive review of NLP applications in RS can be found at [1].

The literature on image captioning presents different methods to address this task. These methods can be categorized as template-based [18], retrieval-based [19], or encoder–decoder-based [2] approaches. Among these, the encoder–decoder framework is widely adopted, where an encoder extracts visual features, and a decoder utilizes these features to generate captions. Qu et al. [2] explored the encoder–decoder architecture using different combinations of convolutional neural networks (CNN) and recurrent neural networks (RNNs). Lu et al. [20] compared deep encoder–decoder methods with handcrafted visual features and introduced an attention layer to focus on visual contents most relevant to the textual description. Other advancements in image captioning include models that summarize multiple captions into one during training [21]. Ramos et al. [22] used continuous word vector representations in the decoder instead of discrete representations. Hoxha et al. [23] employed a decoder based on multiple Support Vector Machines (SVMs) to alleviate overfitting. The authors of [24] trained the captioning model on a truncation cross-entropy loss to reduce overfitting. Huang et al. [25] extracted multi-scale features from CNN and fused them with a denoising operation.

More recent image captioning models have adopted transformer architectures [33] to establish relationships between tokens in the text and image through dot-product attention. In [26], multi-scale visual features are extracted by a CNN, which are decoded using a language transformer. Another proposed approach incorporates the caption type into the caption features within an encoder–decoder based on the transformer, enabling the generation of more controlled captions [27]. In [28], visual features extracted by a CNN are fed into a transformer encoder–decoder trained with a self-critical sequence strategy. In [17], features extracted by a CNN are fed into multiple transformer encoders, and the aggregated output features are then passed through an LSTM and a transformer decoder to generate captions.

VQA is another interesting vision-language task that has received an increasing attention recently. A typically VQA model comprises an image encoder, a question encoder, a multi-modal fusion module, and an answer predictor. Lobry et al. [3] introduced this task into RS and employed a CNN and an RNN for encoding the image and the question, respectively. The features from these encoders are fused by point-wise multiplication and fed to a classifier for predicting the answer. Zheng et al. [30] proposed an alternative model consisting of a CNN and a Gated Recurrent Unit (GRU), incorporating a mutual attention mechanism during the fusion step. Yuan et al. [31] put forth a curriculum learning-based approach, where the model learns questions progressively from easy to hard. Bazi et al. [32] utilized a transformers-based model to encode both the image and the question, employing two decoders to capture the dependencies between them. The answer is then predicted using two classifiers. For addressing open-ended questions, Al Rahhal et al. [29] proposed a model that leverages transformers-based encoders and a decoder with a cross-attention mechanism. This model generates the answer in an autoregressive manner.

2.2. Multi-Tasking in NLP for RS

Limited progress has been made in the realm of multi-task learning for vision-language tasks in the context of RS image analysis. Previous works predominantly focus on training models on auxiliary tasks to enhance the performance of the main task. For example, in [34], the proposed model aims to improve RS image captioning by transferring visual features from a CNN pre-trained on classifying natural and RS images. Wang et al. [35] combined features from a multi-labeling model with CNN-extracted features to train a captioning model. Murali et al. [36] proposed a captioning model that initially undergoes training on VQA as an auxiliary task, followed by leveraging the acquired knowledge to generate more accurate captions. In [37], a multi-label classifier is employed to generate labels from the image, which are then utilized, along with ground truth captions, to train the captioning model. In [38], a CNN predicts the label of the input image, and the label embedding guides the calculation of the attention mask, aiding in caption generation. In [39], a semantic segmentation mask is generated using weakly supervised learning, serving as an attention mask during caption generation. Wang et al. [40] fine-tuned a pre-trained model to perform RS image classification, utilizing two classification heads to consider features of different scales. These features are subsequently passed into a decoder to generate captions. Yuan et al. [41] proposed a model with two branches: a CNN to extract multi-level features, and a multi-label classifier with a graph convolutional network. The extracted features from both branches are fed into a language decoder for caption generation.

The aforementioned methods break the training process into two stages, where the model is trained first on the auxiliary task and then the obtained knowledge is utilized by the main task. Alternatively, another line of work proposes training on the auxiliary task jointly with the main task to better exploit the synergy between them. In [42,43], a captioning model with an auxiliary multi-labeling classifier is proposed, where the classifier is jointly trained with the captioning model to generate better descriptions. Hoxha et al. [44] proposed a text-based image retrieval model that utilizes a captioning model to generate textual descriptions from RS images and the retrieval is performed by matching the query to all texts. Ma et al. [45], proposed a captioning model that employs object detection as an auxiliary task to obtain better representations of the image for captioning model.

Despite the progress made in applying NLP to RS image analysis, there has been limited effort to integrate vision-language tasks into a unified, multi-tasking model.

2.3. Vision-Language Models in General Computer Vision

Vision-language models represent a category of models that integrate computer vision and NLP techniques to achieve a comprehensive understanding of visual and textual data. The field of research in this area has witnessed significant advancements, encompassing a wide range of techniques and architectures that aim to fuse vision and language, thereby enhancing performance and capabilities across various tasks.

One approach adopted by these models involves learning generic multi-modal representations from a large corpus of image-text pairs. Models such as CLIP [46] and ALIGN [47] employ dual encoders trained with a contrastive objective on extensive image-text datasets, enabling them to model the interaction between modalities. However, these models do not natively support multi-modal generation tasks (e.g., image captioning and VQA). Other models, like BLIB [44] and CoCa [45], combine contrastive and generative pre-training to support both cross-modal alignment and multi-modal generation tasks.

Another paradigm of vision-language models are the generalized models that are trained to jointly perform multiple tasks without task-specific fine-tuning. For instance, UniT [48] unifies different tasks in a single model by designing a decoder with per-task query embedding and task-specific output heads. OFA [49] formulates different vision-language tasks as sequence-to-sequence tasks and trains a transformer model on them without task-specific layers. GIT [50] is a multi-tasking transformer-based model that conditions the decoder on vision inputs and text to generate the output text. Flamingo [51] extends sequence-to-sequence models to support interleaved image, video and text inputs for generating the output text.

Building upon the success of GPT models, GPT-4 [14] has exhibited impressive capabilities in engaging in multi-modal dialogues with humans. The open-source project LLaVA [15] aims to replicate this performance by aligning visual representations with the input space of the LLM. By leveraging the original self-attention mechanism within the LLM, LLaVA enables effective processing of visual information alongside textual input. Multimodal-GPT [52] utilizes gated cross-attention layers to facilitate seamless interactions between images and text. These layers are fine-tuned using carefully constructed instruction data, aimed at enhancing user interaction and improving the model's performance in multimodal tasks.

2.4. Vision-Language Models in RS

The literature has witnessed some initiatives to employ vision-language models for RS data analysis. Some researchers have focused on harnessing the zero-shot classification capabilities of these models, leveraging the aligned representations learned from both the vision and language domains. For instance, Qiu et al. [53] proposed the utilization of the pre-trained CLIP [46] model for extracting RS image features and classifying images into classes that were not encountered during training, achieving promising results in scene classification despite limited labeled data. Xiang Li et al. [54] enhanced zero-shot classification in the RS domain by adapting the CLIP model using a pseudo-labeling technique and curriculum fine-tuning strategy. Additionally, Al Rahhal et al. [55] conducted experiments to determine the most effective prompt for querying the language backbone of CLIP for zero-shot classification of RS images. Furthermore, Bazi et al. [32] employed the pre-trained CLIP model for VQA. Their approach involved utilizing CLIP-based vision and language transformers to extract visual and textual representations. To capture the dependencies within and between these representations, they incorporated two decoders with two classifiers to predict the answer. Ricci et al. [56] utilized ChatGPT and Blip [57] models to initiate a machine-to-machine dialogue to extract information from the RS image. This dialogue is concluded by generating a summary of the discussion, resulting in the final description of the RS image's content.

3. The RS-Instructions Dataset

Instruction tuning is a training technique used to adapt LLMs or LVLMs models to better understand and generate responses based on specific instructions [15]. It involves training the model to align with the desired behavior or task by providing explicit instructions during the training process. During instruction tuning, the model is exposed to examples that include both the input and the desired response. To adapt LVLMs for RS tasks, it is essential to have an instruction dataset specifically tailored to RS. This dataset should consist of image, instruction, and output text triplets. However, currently, there is no comparable instructional data available for the RS domain. To address this gap, we have developed the RS-instructions dataset, which is a multi-task RS vision-language instruction dataset created from existing RS datasets by transforming the information present in these datasets into instructional format. This enables the model to grasp and comprehend the complexities of language and vision within the context of RS analysis.

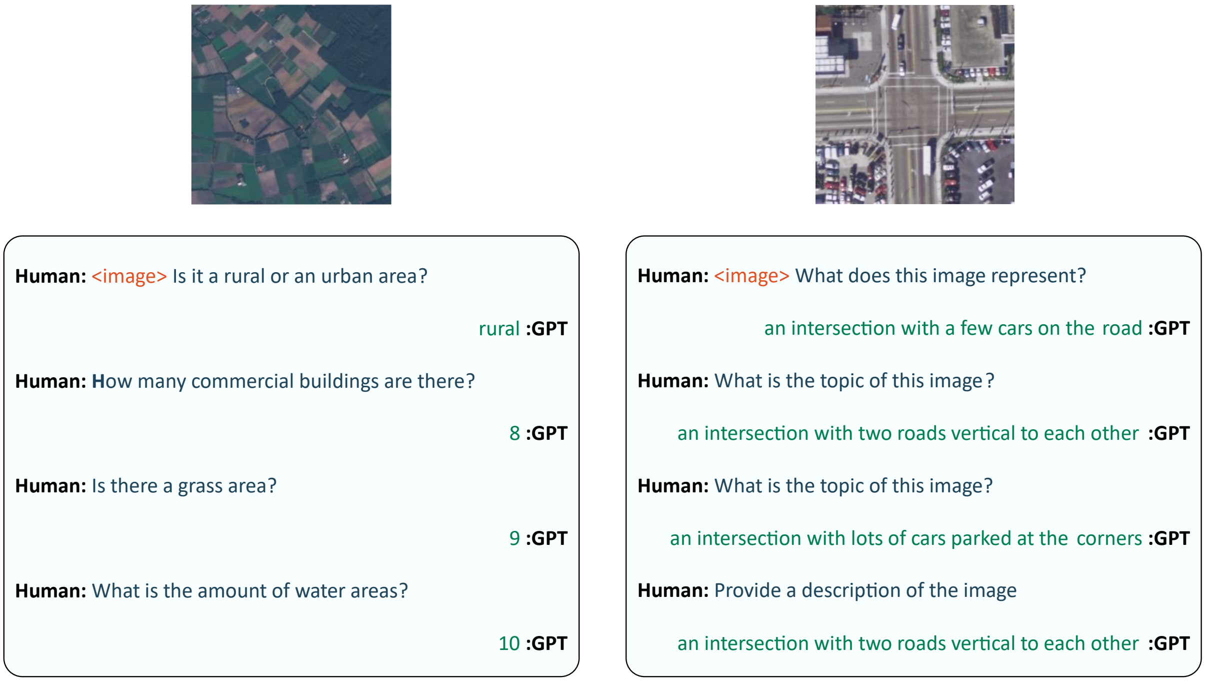
Since RS-LLaVA is trained to perform both captioning and VQA based on the instruction given to the model, the RS-instructions dataset is constructed by mixing four captioning and VQA datasets. Specifically, we leverage two existing captioning datasets UCM-caption [2], and UAV [23], as well as two VQA datasets, RSVQA-LR [3], and RSIVQA-DOTA [30]. We followed the same training and testing split as the original datasets. This results in a dataset comprising 7058 samples, with 5506 samples in the training set and 1552 samples in the test set. A summary of these datasets can be found in Table 1, and more detailed information about each dataset used to build the RS-instructions dataset is provided in the following:

- The UCM-caption [2] is a captioning dataset derived from the University of California Merced land-use (UCM) dataset [58], which was initially designed for scene classification purposes. Each image in the dataset is assigned to one of 21 land-use classes. The dataset comprises a total of 2100 RGB images, with 100 images per class. The UCM-caption images have a size of 256 × 256 pixels and a spatial resolution of 0.3048 m. Each image is associated with five distinct captions. Consequently, the dataset encompasses a collection of 10,500 sentences. To facilitate experimentation and evaluation, the dataset is split into three subsets: the training set encompasses 80% of the images, amounting to 1680 images; the evaluation dataset encompasses 10% of the images, totaling 210 images; and the remaining 10% of images, also amounting to 210 images, are designated for the test dataset.
- UAV [23] is a captioning dataset that was captured near the city of Civezzano, Italy, on 17 October 2012, using an unmanned aerial vehicle equipped with an EOS 550D camera. It comprises a total of ten RGB images, each with a resolution of 2 cm and a size of 5184 × 3456 pixels, resulting in a spatial resolution of 2 cm. Among the ten images, six are allocated for training purposes, one for validation, and three for testing. From these images, crops of size 256 × 256 pixels are extracted. Specifically, the training images yield a total of 1746 crops, while the testing image provides 882 crops. Each crop is associated with three descriptions, authored by different annotators.
- RSVQA-LR [3] consists of 772 low-resolution images. This dataset was curated using seven tiles captured by the Sentinel-2 satellite, covering an area of 6.55 km^2 in the Netherlands. Each image in the dataset has dimensions of 256 × 256 pixels and consists of RGB spectral channels, with a spatial resolution of 10 m. The dataset comprises a total of 772 images, which are split into 572, 100, and 100 images for training, validation, and testing, respectively. The total number of questions in the dataset is 77,232, with each image annotated with approximately 100–101 questions. The questions in the dataset cover four categories: object presence (answer: yes/no), comparisons between objects (answer: yes/no), rural/urban classification (answer: rural/urban), and object counting.
- RSIVQA-DOTA [30] is a VQA dataset is based on the DOTA [59] object detection dataset. It includes questions about scenes, objects, relative locations, color, and shape. The total number of image/question/answer triplets in the dataset is 16,430. The questions are of three types: presence, counting and other. The dataset is split into three sets: the training set which represents 80% of the entire set, the testing set that comprises 10%, and the validation set that comprises 10%.

Table 1. Datasets used to build the RS-instructions dataset.

Dataset	Task	#Images	Image Size	Text per Image	Resolution (m)
UCM-captions [2]	Captioning	2100	256 × 256	5	0.3048
UAV [23]	Captioning	2628	256 × 256	3	0.02
RSVQA-LR [3]	VQA	772	256 × 256	100–101	10
RSIVQA-DOTA [30]	VQA	1868	Varies	3–24	Varies
RS-instructions dataset	Captioning + VQA	7058	Varies	Varies	Varies

To construct the RS-instructions dataset, questions, and answers in the two VQA datasets have been formatted in a conversation format as shown in Figure 1a. For captioning datasets, we use a set of instructions that simply asks for a description of the image such as ‘Describe the image’ and ‘What does this image represent?’ to transform the original datasets into the instruction–answer format as shown in Figure 1b.



Human: <image> Is it a rural or an urban area?

rural **:GPT**

Human: How many commercial buildings are there?

8 **:GPT**

Human: Is there a grass area?

9 **:GPT**

Human: What is the amount of water areas?

10 **:GPT**

Human: <image> What does this image represent?

an intersection with a few cars on the road **:GPT**

Human: What is the topic of this image?

an intersection with two roads vertical to each other **:GPT**

Human: What is the topic of this image?

an intersection with lots of cars parked at the corners **:GPT**

Human: Provide a description of the image

an intersection with two roads vertical to each other **:GPT**

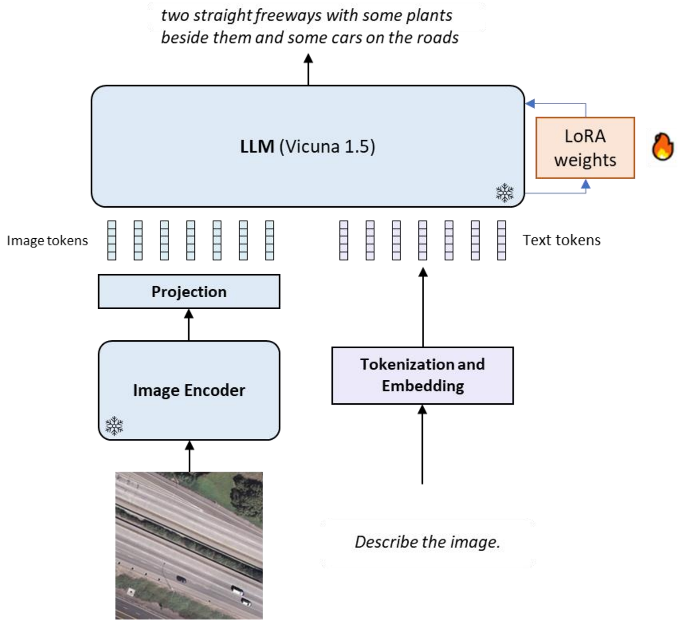
(**a**) (**b**)

Figure 1. Samples from the RS-instructions Dataset: (**a**) Image from RSVQA-LR dataset and (**b**) image from UCM dataset.

4. The RS-LLaVA Model

4.1. Model Architecture

The architecture of RS-LLaVA, which is shown in Figure 2, consists of a pre-trained visual backbone to encode the image, a chat-based LLM to generate the response, and a projection network that connects the visual backbone to the language model.



two straight freeways with some plants beside them and some cars on the roads

Describe the image.

Figure 2. Architecture of RS-LLaVA: In this model, the image encoder and language decoder are frozen, while we fine-tune the model using the LoRA method. The LoRA method adds extra weight to the original LLM.

Given a sample $\{X,\ I,\ T\}$ from the RS-instructions dataset, where X represents the image, I denotes the instruction, and T represents the response to the instruction. Initially, the image encoder is employed to extract visual tokens from the input image $X \in \mathbb{R}^{H\times W\times C}$, where H, W, and C rerepresent he height, the width, and the number of channels, respectively. The encoder encodes the image into the image tokens $Z_x \in \mathbb{R}^{N\times D}$, where N is length of the sequence of tokens, and D is the dimension of the image encoder.

Subsequently, the resulting sequence is passed through the projection network, which is a two-layer network with GELU activation, which maps the visual tokens to the embedding space dimension S, forming the sequence $F_x \in \mathbb{R}^{N\times S}$. The mapped image features are then concatenated with textual instruction tokens $F_I \in \mathbb{R}^{M\times S}$, forming the input for the LLM $F \in \mathbb{R}^{K\times S}$, where $K = M + N$.

The LLM is a chat-based language model based on the transformer architecture. The model takes a sequence F of visual and language tokens as input and starts to generate the response in an auto-regressive manner. This involves maximizing the probability distribution of generating the correct response given the image-instruction tokens. This probability distribution can be represented as follows:

$$P(T|X,\ I) = \prod_{k=1}^{K} P\left(T_k \middle| T_{1,\ldots,(k-1)},\ I, X\right) \tag{1}$$

where K represents the length of the response sequence, and $P\left(T_k \middle| T_{1,\ldots,(k-1)},\ I, X\right)$ denotes the probability of the k-th token given the previous tokens, instruction, and image.

4.2. Model Training

The training process model consists of two steps: (1) pre-training and (2) fine-tuning. During the pre-training phase, the image encoder and the LLM weights are kept frozen, and only the projection network is trained using a general image-language dataset for text–image pairs. In the subsequent step, the projection network and the image encoder are frozen, while the LLM is fine-tuned.

Fine-tuning LLMs can be challenging and computationally expensive due to their large number of parameters. To address this, we employ LoRA [16], which is a fine-tuning technique that facilitates the fine-tuning of large models. The key idea of LoRA is to decompose the large weight matrix of the LLM into two smaller matrices through low-rank decomposition. This decomposition creates trainable pairs of rank decomposition matrices that run in parallel with the existing weight matrices, and only these new matrices are fine-tuned to adapt to the RS data.

Formally, given a pre-trained weight matrix $W_0 \in \mathbb{R}^{u\times v}$, the update is represented with a low-rank decomposition of that matrix $W_0 + \triangle W = W_0 + BA$, with $B \in \mathbb{R}^{u\times r},\ A \in \mathbb{R}^{r\times v}$ and the rank $r \ll \min(u, v)$. During training, W_0 is frozen and does not receive gradient updates, while A and B contain the fine-tuned weights that represent the differences to be added to the original weights of the LLM. During inference, the fine-tuned weights are combined with the original pre-trained weights. Both W_0 and $\triangle W = BA$ are multiplied with the same input, and their respective output vectors are summed coordinate-wise. For an input x and $h_0 = W_0 x$, the modified forward pass can be expressed as:

$$h = W_0 x + \triangle W x = W_0 x + BAx \tag{2}$$

To initialize the parameters, A is randomly initialized using Gaussian initialization, while B is initialized with zeros. Thus, at the beginning of training, $\triangle W = BA$ is zero. To scale $\triangle W x$, it is multiplied by $\frac{\alpha}{r}$, where α is a constant related to r. When optimizing with Adam, tuning α is approximately equivalent to tuning the learning rate, provided the initialization is appropriately scaled. Therefore, α is typically set to the first r value tested and is not further tuned. Specifically, all weight matrices of the LLMs are frozen and the LoRA technique is implemented on the W_q, W_k, and W_v, weights in the attention layers.

5. Experimental Results

5.1. Experimental Settings

The RS-LLaVA model is based on the architecture of the LLaVA model [15]. In our experiments, we explore two variants of pre-trained Vicuna-v1.5 [60] LLM variants, ranging in size from 7B to 13B, to initialize the language model for RS-LLaVA. Vicuna 1.5 is an open-source large language model developed by LMSYS. It is a fine-tuned version of the Llama 2 model, trained on user conversations collected from ShareGPT. Vicuna is licensed under the Llama 2 Community License Agreement. For image encoding, the model adopts the pre-trained vision backbone of CLIP-ViT (large) [46], which utilizes an image resolution of 336 × 336.

To facilitate fine-tuning, we employ LoRA with a rank (r) set to 64 and α set to 16 as suggested by the original paper. We utilize the Adam optimizer, with a learning rate of 1×10^{-4}. Figures 3–5 display the training loss during the fine-tuning process. It is evident from all the figures that the loss experiences a substantial decrease during the initial stages of training. However, as the training progresses, the rate of loss reduction gradually slows down.

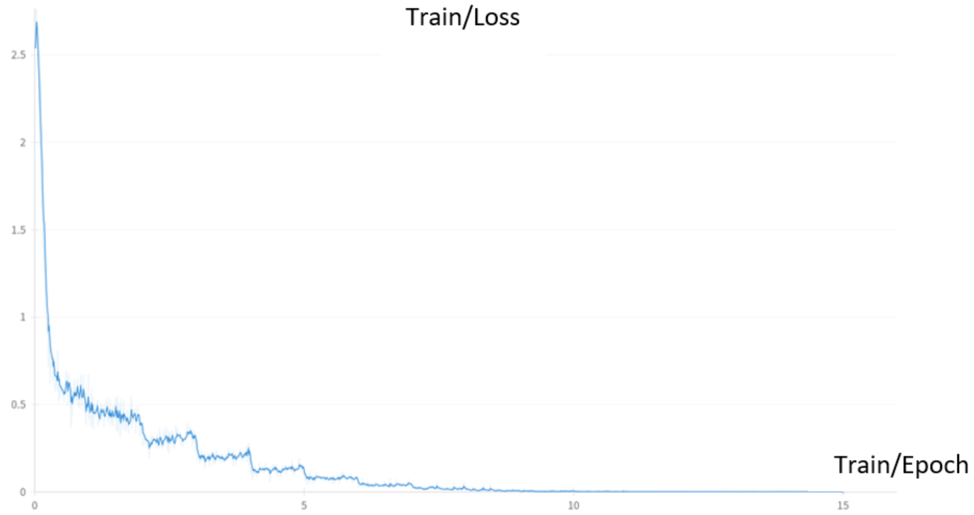


Figure 3. Loss during the fine-tuning process on the UAV dataset.

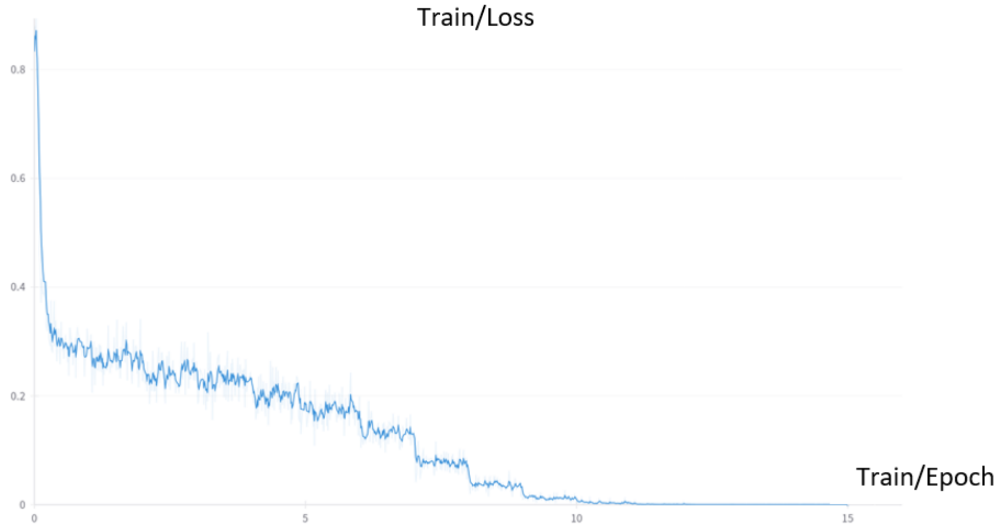


Figure 4. Loss during the fine-tuning process on the LR-VQA dataset.

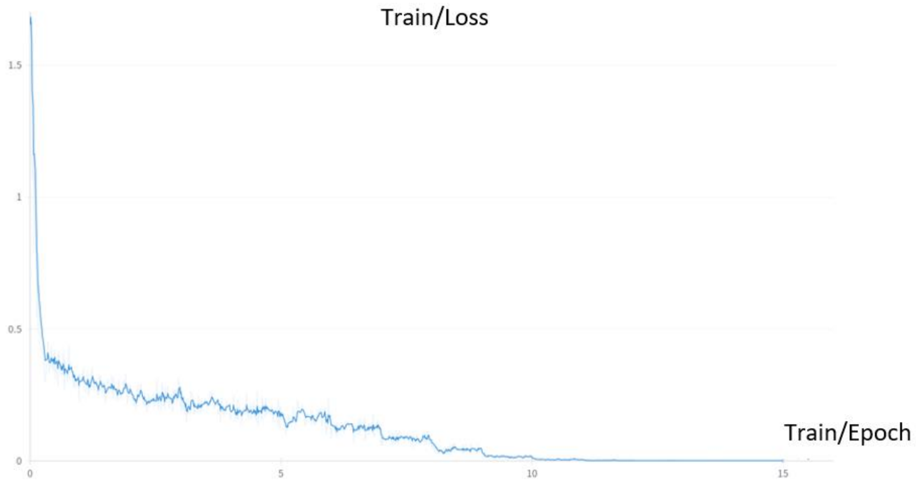


Figure 5. Loss during the fine-tuning process on the RS-instructional dataset.

5.2. Evaluation Metrics

We utilize different metrics to evaluate the performance of the model in the captioning and VQA tasks. In the captioning task, we assess the performance using the following metrics: BLEU score with n-gram values ranging from 1 to 4 [61], METEOR [62], ROUGE [63], and CIDEr [64]. The BLEU (Bilingual Evaluation Understudy) score compares the n-grams (from 1 to 4 words) for the generated text to the reference text, providing an objective measure of the text's quality. METEOR (Metric for Evaluation of Translation with Explicit Ordering) considers synonyms and paraphrases when comparing the generated text to references. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) focuses on evaluating text summarization by measuring the overlap between generated summaries and reference summaries. Finally, CIDEr (Consensus-based Image Description Evaluation) is specifically designed for image captioning and text generation tasks, assessing both the semantic similarity and n-gram overlap between the generated output and reference captions.

For the VQA task, we evaluate the answers for all question types in VQA-LR based on accuracy. In the case of VQA-DOTA, the presence questions are evaluated using three metrics: precision, recall, and the F1 score. As for count questions, which elicit numerical responses, we evaluate them using the root mean square error (RMSE). We use Pytorch for training the model on a station with an Intel Core i9-14900K 14th Gen Processor, 192 GB RAM, and 2 NVIDIA RTX A6000 GPUs with 48 GB of memory each. We use also an open-source DeepSpeed library that enables efficient and scalable training of very large models by providing advanced memory, speed, and parallelism optimizations.

5.3. Results

5.3.1. Results on the Captioning Task

In this experiment, we assess the quality of the generated captions using different variants of the model. The evaluation is conducted in two scenarios: when the model is trained on the constructed RS-instructions dataset to perform multiple tasks, and when the model is trained on the respective dataset itself. The results are presented for two sizes of the pre-trained Vicuna-v1.5 model [60] (7B and 13B).

Table 2 presents the model's performance on the UCM-caption dataset. As observed, both scenarios demonstrate that fine-tuning a larger language model leads to higher performance, with the cost of longer training time. Additionally, there is a minor difference between the results of the language model trained on the RS-instructions dataset and the UCM-caption dataset only, particularly for the smaller language model. This difference represents a noticeable improvement in terms of the BLEU4, METEOR, and CIDEr metrics.

Table 2. Captioning results on the UCM-captions dataset.

Training	Merced	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGH	CIDEr	Training Time
Single	CLIP336-Vicuna7B	88.77	82.84	77.62	72.78	46.75	83.72	343.21	2.10 h
	CLIP336-Vicuna13B	90.41	84.54	79.39	74.50	48.62	86.09	355.06	3.73 h
Joint	CLIP336-Vicuna7B	88.70	82.88	77.70	72.84	47.98	85.17	349.43	11.04 h
	CLIP336-Vicuna13B	90.00	84.88	80.30	76.03	49.21	85.78	355.61	17.04 h

The results on the UAV dataset shown in Table 3 illustrate that when fine-tuning RS-LLaVA solely on the UAV dataset, it exhibits better performance compared to fine-tuning the model on the RS-instructions dataset. We also observe that the smaller Vicuna language model outperforms the larger model, which can be attributed to the limited size of the UAV dataset.

Table 3. Captioning results on the UAV dataset.

Training	UAV	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGH	CIDEr	Training Time
Single	CLIP336-Vicuna7B	82.60	72.67	63.04	53.27	42.03	78.21	427.17	1.83 h
	CLIP336-Vicuna13B	82.64	72.51	62.57	52.79	42.77	79.15	423.18	3.81 h
Joint	CLIP336-Vicuna7B	79.81	69.27	59.00	49.02	40.46	76.80	404.54	11.04 h
	CLIP336-Vicuna13B	79.82	69.60	58.84	49.24	40.14	76.28	390.30	17.04 h

5.3.2. Results of the RS-LLaVA on the VQA Task

In this experiment, we assess the ability of the model to answer questions about given RS images. The results on the RSVQA-LR dataset, presented in Table 4, reflect the model's performance in terms of accuracy, categorized by question type. The overall accuracy is computed by summing the correct classification samples over the total number of samples, while the average accuracy is computed by averaging the individual accuracies. Notably, fine-tuning the model solely on the single RSVQA-LR dataset leads to superior results compared to training on the RS-instructions dataset. Additionally, when trained on a single dataset, the larger model tends to exhibit slightly better performance.

Table 4. VQA results on the RSVQA-LR dataset.

Training	VQA_LR	Count	Presence	Comparisons	Urban/Rural	Average	Overall	Training Time
Single	CLIP-336-Vicuna-7B	75.05	92.97	91.23	95.00	88.56	87.20	3.25 h
	CLIP-336Vicuna-13B	75.87	92.32	91.37	95.00	88.64	87.22	7.10 h
Joint	CLIP-336-Vicuna-7B	74.38	92.80	91.33	94.00	88.13	86.95	11.04 h
	CLIP-336Vicuna-13B	73.76	92.27	91.37	95.00	88.10	86.58	19.40 h

Results on the RSIVQA-DOTA dataset are provided in Table 5. As previously mentioned, presence questions are evaluated based on precision, recall, and the F1 score. We observe that fine-tuning the Vicuna-13B model exclusively on the VQA-DOTA dataset achieves the highest F1 score of 85.80. Conversely, fine-tuning the Vicuna-7B model on the RS-instructions dataset yields the best-balanced score between precision and recall. For assessing the counting capability, we employ the RMSE. The Vicuna-7B model trained on the RS-instructions dataset achieves the lowest RMSE for counting questions.

Table 5. VQA results on the RSIVQA-DOTA dataset.

Training	VQA_DOTA	Count	Yes/No			Training Time
		RMSE	Pr	Re	F1	
Single	CLIP-336-Vicuna-7B	221.40	91.49	72.07	80.63	1.95 h
	CLIP-336-Vicuna-13B	226.79	89.03	82.79	85.80	3.88 h
Joint	CLIP-336-Vicuna-7B	209.47	85.26	86.15	85.70	11.04 h
	CLIP-336-Vicuna-13B	232.75	100	33.28	49.94	19.40 h

5.3.3. Comparison with State-of-the-Art Methods

In this section, we present a comparison of the model with state-of-the-art methods in RS image captioning and VQA. Here, we compare the results of the state-of-the-art methods on the RS-instructions dataset. Table 6 compares the results of RS-LLaVA and other methods on the UCM-caption dataset. Table 6 demonstrates that RS-LLaVA with Vicuna13B outperforms the other methods in all metrics, except for CIDEr, where it achieves the second position.

Table 6. Results of different RS image captioning methods on the UCM-caption dataset.

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGH	CIDEr
CSMLF [19]	43.61	27.28	18.55	12.10	13.20	39.27	22.27
VGG19+LSTM [2]	63.80	53.60	37.70	21.90	20.60	-	45.10
GoogLeNet -hard att. [20]	83.75	76.22	70.42	65.62	44.89	79.62	320.01
VAA [65]	81.92	75.11	69.27	63.87	43.80	78.24	339.46
Yuan et al. [41]	83.30	77.12	71.54	66.23	43.71	77.63	316.84
ResNet18 MSF [25]	83.06	75.98	69.72	63.45	-	73.18	329.56
SD-RISC [21]	74.80	66.40	59.80	53.80	39.00	69.50	213.20
Hoxha et al. [23]	76.53	69.47	64.17	37.02	37.02	68.77	292.28
Li et al. [24]	82.10	76.22	71.40	67.00	47.75	75.67	285.47
MSA [45]	83.37	78.22	74.06	70.21	45.04	79.18	325.71
Word-sentence [37]	79.31	72.37	66.71	62.02	43.95	71.32	278.71
Structured att. [39]	85.38	80.35	75.72	71.49	46.32	81.41	334.89
Zia et al. [26]	83.90	76.90	71.50	67.50	44.60	-	323.10
Li et al. [66]	85.18	79.25	74.32	69.76	45.71	80.72	338.87
SCST [28]	83.40	77.60	72.30	67.60	-	76.00	336.00
Wang et al. [40]	84.30	77.50	71.10	65.10	45.30	78.50	338.10
MLCA [67]	82.60	77.00	71.70	66.80	43.50	77.20	324.00
Ye et al. [43]	86.96	82.24	77.88	73.76	49.06	83.64	371.02
CLIP336-Vicuna7B (Joint)	88.70	82.88	77.70	72.84	47.98	85.17	349.43
CLIP336-Vicuna13B (Joint)	90.00	84.88	80.30	76.03	49.21	85.78	355.61

Table 7 provides the results of different captioning methods on the UAV dataset. The table shows that the performance of RS-LLaVA, whether with Vicuna7B or Vicuna13B, surpasses other methods, demonstrating higher overall performance in all metrics.

Table 7. Results of different RS image captioning methods on the UAV dataset.

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGH	CIDEr
Hoxha et al. [23]	68.84	58.05	48.33	39.22	32.81	69.63	391.31
Hoxha et al. [23]	65.13	56.53	48.15	39.69	32.17	69.31	389.45
Bashmal et al. [12]	77.11	66.45	55.99	45.17	38.18	75.19	390.27
CLIP336-Vicuna7B (Joint)	79.81	69.27	59.00	49.02	40.46	76.80	404.54
CLIP336-Vicuna13B (Joint)	79.82	69.60	58.84	49.24	40.14	76.28	390.30

Finally, in Table 8, we compare the performance of the model in answering questions about RS scenes using the RSVQA-LR dataset. The results show that both sizes of Vicuna offer more accurate answers compared to state-of-the-art methods across all question types. These experiments validate the ability of the model to effectively comprehend textual instructions and accomplish diverse RS visual understanding tasks.

Table 8. Results of different RS VQA methods on the RSVQA-LR dataset.

Method	Count	Presence	Comparisons	Urban/Rural	Average	Overall
Lobry et al. [3]	67.01	87.46	81.50	90.00	81.49	79.08
Yuan et al. [31]	68.53	90.13	86.91	92.00	84.39	82.50
Bazi et al. [32]	72.22	91.06	91.16	92.66	86.78	85.56
CLIP336-Vicuna7B (Joint)	74.38	92.80	91.33	94.00	88.13	86.95
CLIP336-Vicuna13B (Joint)	73.76	92.27	91.37	95.00	88.10	86.58

5.3.4. Qualitative Results

In this section, we present the qualitative results obtained from our experiments, which provide visual evidence of the performance and capabilities of the model. Figures 6–9 showcase outputs from different samples of the RS-instructions dataset. By visually examining the responses generated by the model in the image captioning task, we observe a high degree of similarity between the model's responses and the ground truth captions in both the UCM-caption dataset and the UAV dataset. This alignment indicates the model's proficiency in generating captions that accurately correspond to the given instructions and the image content.

Instruction: Provide a description of the image.

Response: There are many airplanes stopped at the airport.

Ground Truth: There are many airplanes at the airport.

Instruction: What does this image represent?

Response: There are two straight freeways in the desert.

Ground Truth: There are two straight freeways in the desert.

Figure 6. Sample of RS-LLaVA results from UCM-captions dataset.

Instruction: Provide a description of the image.

Response: Some asphalt on the left and some trees on the right .

Ground Truth: Small road on the left and several trees on the right.

Instruction: What does this image represent?

Response: Soil field with building facade on the top.

Ground Truth: Soil ground with building facade on the top.

Figure 7. Sample of RS-LLaVA results from UAV dataset.

Instruction: how many small vehicles are there in this picture?

Response: 20

Ground Truth: 2

Instruction: how many large vehicles are there in this picture?

Response: 1

Ground Truth: 16

Instruction: how many roundabouts are there in this picture?

Response: 2

Ground Truth: 1

Instruction: does this picture contain small vehicle?

Response: yes

Ground Truth: yes

Instruction: how many planes are there in this picture?

Response: 3

Ground Truth: 3

Instruction: how many small vehicles are there in this picture?

Response: 8

Ground Truth: 2

Instruction how many large vehicles are there in this picture?

Response: 3

Ground Truth: 2

Instruction: how many helicopters are there in this picture?

Response: 3

Ground Truth: 2

Figure 8. Sample of RS-LLaVA results from RSIVQA-DOTA dataset.

Instruction: what is the amount of large water areas?

Response: 6

Ground Truth: there are 6 large water areas

Instruction: are there more water areas than roads?

Response: no

Ground Truth: no there are not more water areas than roads

Instruction: is there a large road?

Response: no

Ground Truth: no there is no large road.

Instruction: what is the number of large forests?

Response: 7

Ground Truth: there are 6 large forests.

Instruction: is it a rural or an urban area?

Response: urban

Ground Truth: It is an urban area.

Instruction: is a commercial building present?

Response: yes

Ground Truth: yes a commercial building is present.

Instruction is a residential building present?

Response: yes

Ground Truth: yes a residential building is present

Instruction: what is the amount of grass areas?

Response: 326

Ground Truth: there are 268 grass areas

Figure 9. Sample of RS-LLaVA results from the RSVQA-LR dataset.

The results of VQA task are shown in Figures 8 and 9. We can observe that the model encounters some challenges in answering counting questions, which is recognized as a complex task. However, the model demonstrates improved performance in answering other types of questions, such as presence-related inquiries.

6. Conclusions

This paper explored the promising capabilities of LLMs and their extension, LVLMs, in the field of RS, specifically by investigating their multi-tasking potential for tasks like image captioning and VQA. We introduced RS-LLaVA, an enhanced version of LLaVA adapted for RS imagery. To train this model, we developed the RS-instructions dataset by leveraging existing four single-task datasets. Then, we fine-tuned the architecture using the LoRA method that adds extra-tunable weights to the large language model. We have demonstrated the capability of the proposed architecture using two different LLMs, namely vicuna-7B and vicuna-13B. While the experiments demonstrated the notable performance of the proposed RS-LLaVA architecture, it is important to mention the computational challenges posed by large parameter sizes. Indeed, LLMs often require extensive computational resources for training and inference, limiting their accessibility and scalability. To address this issue in future research, efforts should focus on exploring techniques for model compression, such as knowledge distillation or parameter pruning, to reduce the computational burden while maintaining performance. Additionally, one can plan to integrate additional datasets and tasks, such as visual grounding and change detection in multi-temporal images to further enhance the versatility and applicability of RS-LLaVA in RS applications.

Author Contributions: Methodology, Y.B., M.M.A.R. and F.M.; Software, Y.B.; Formal analysis, R.R.; Writing—original draft, Y.B. and L.B.; Writing—review & editing, L.B., M.M.A.R., R.R. and F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the researchers supporting project number RSPD2024R995, King Saud University, Riyadh, Saudi Arabia.

Data Availability Statement: The codes and datasets will be released at: https://github.com/BigData-KSU/RS-LLaVA.

Acknowledgments: This research was supported by the researchers supporting project number (RSPD2024R995), King Saud University, Riyadh, Saudi Arabia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bashmal, L.; Bazi, Y.; Melgani, F.; Al Rahhal, M.M.; Al Zuair, M.A. Language Integration in Remote Sensing: Tasks, datasets, and future directions. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 63–93. [CrossRef]
2. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding of high resolution remote sensing image. In Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), Kunming, China, 6–8 July 2016; pp. 1–5. [CrossRef]
3. Lobry, S.; Marcos, D.; Murray, J.; Tuia, D. RSVQA: Visual Question Answering for Remote Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8555–8566. [CrossRef]
4. Abdullah, T.; Bazi, Y.; Al Rahhal, M.M.; Mekhalfi, M.L.; Rangarajan, L.; Zuair, M. TextRS: Deep Bidirectional Triplet Network for Matching Text to Remote Sensing Images. *Remote Sens.* **2020**, *12*, 405. [CrossRef]
5. Zhao, R.; Shi, Z. Text-to-Remote-Sensing-Image Generation with Structured Generative Adversarial Networks. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
6. Bejiga, M.B.; Melgani, F.; Vascotto, A. Retro-Remote Sensing: Generating Images from Ancient Texts. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 950–960. [CrossRef]
7. Bejiga, M.B.; Hoxha, G.; Melgani, F. Improving Text Encoding for Retro-Remote Sensing. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 622–626. [CrossRef]
8. Zhan, Y.; Xiong, Z.; Yuan, Y. RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–13. [CrossRef]
9. Liu, C.; Zhao, R.; Chen, H.; Zou, Z.; Shi, Z. Remote Sensing Image Change Captioning with Dual-Branch Transformers: A New Method and a Large Scale Dataset. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [CrossRef]

10. Hoxha, G.; Chouaf, S.; Melgani, F.; Smara, Y. Change Captioning: A New Paradigm for Multitemporal Remote Sensing Image Analysis. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
11. Yuan, Z.; Mou, L.; Xiong, Z.; Zhu, X.X. Change Detection Meets Visual Question Answering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
12. Bashmal, L.; Bazi, Y.; Melgani, F.; Ricci, R.; Al Rahhal, M.M.; Zuair, M. Visual Question Generation From Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 3279–3293. [CrossRef]
13. OpenAI. ChatGPT. OpenAI API, 2023. Available online: https://openai.com/blog/chatgpt (accessed on 1 April 2024).
14. OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
15. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. *arXiv* **2023**, arXiv:2304.08485.
16. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
17. Liu, C.; Zhao, R.; Shi, Z. Remote-Sensing Image Captioning Based on Multilayer Aggregated Transformer. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
18. Shi, Z.; Zou, Z. Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image? *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3623–3634. [CrossRef]
19. Wang, B.; Lu, X.; Zheng, X.; Li, X. Semantic Descriptions of High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1274–1278. [CrossRef]
20. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2183–2195. [CrossRef]
21. Sumbul, G.; Nayak, S.; Demir, B. SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6922–6934. [CrossRef]
22. Ramos, R.; Martins, B. Using Neural Encoder-Decoder Models with Continuous Outputs for Remote Sensing Image Captioning. *IEEE Access* **2022**, *10*, 24852–24863. [CrossRef]
23. Hoxha, G.; Melgani, F. A Novel SVM-Based Decoder for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
24. Li, X.; Zhang, X.; Huang, W.; Wang, Q. Truncation Cross Entropy Loss for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5246–5257. [CrossRef]
25. Huang, W.; Wang, Q.; Li, X. Denoising-Based Multiscale Feature Fusion for Remote Sensing Image Captioning. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 436–440. [CrossRef]
26. Zia, U.; Riaz, M.M.; Ghafoor, A. Transforming remote sensing images to textual descriptions. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *108*, 102741. [CrossRef]
27. Chen, Z.; Wang, J.; Ma, A.; Zhong, Y. TypeFormer: Multiscale Transformer with Type Controller for Remote Sensing Image Caption. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
28. Zhuang, S.; Wang, P.; Wang, G.; Wang, D.; Chen, J.; Gao, F. Improving Remote Sensing Image Captioning by Combining Grid Features and Transformer. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
29. Al Rahhal, M.M.; Bazi, Y.; Alsaleh, S.O.; Al-Razgan, M.; Mekhalfi, M.L.; Al Zuair, M.; Alajlan, N. Open-ended remote sensing visual question answering with transformers. *Int. J. Remote Sens.* **2022**, *43*, 6809–6823. [CrossRef]
30. Zheng, X.; Wang, B.; Du, X.; Lu, X. Mutual Attention Inception Network for Remote Sensing Visual Question Answering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
31. Yuan, Z.; Mou, L.; Wang, Q.; Zhu, X.X. From Easy to Hard: Learning Language-Guided Curriculum for Visual Question Answering on Remote Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
32. Bazi, Y.; Rahhal, M.M.A.; Mekhalfi, M.L.; Zuair, M.A.A.; Melgani, F. Bi-Modal Transformer-Based Approach for Visual Question Answering in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
34. Yang, Q.; Ni, Z.; Ren, P. Meta captioning: A meta learning based remote sensing image captioning framework. *ISPRS J. Photogramm. Remote Sens.* **2022**, *186*, 190–200. [CrossRef]
35. Wang, S.; Ye, X.; Gu, Y.; Wang, J.; Meng, Y.; Tian, J.; Hou, B.; Jiao, L. Multi-label semantic feature fusion for remote sensing image captioning. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 1–18. [CrossRef]
36. Murali, N.; Shanthi, A.P. Remote Sensing Image Captioning via Multilevel Attention-Based Visual Question Answering. In *Innovations in Computational Intelligence and Computer Vision*; Roy, S., Sinwar, D., Perumal, T., Slowik, A., Tavares, J.M.R.S., Eds.; Springer Nature: Singapore, 2022; pp. 465–475.
37. Wang, Q.; Huang, W.; Zhang, X.; Li, X. Word–Sentence Framework for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 10532–10543. [CrossRef]
38. Zhang, Z.; Diao, W.; Zhang, W.; Yan, M.; Gao, X.; Sun, X. LAM: Remote Sensing Image Captioning with Label-Attention Mechanism. *Remote Sens.* **2019**, *11*, 2349. [CrossRef]

39. Zhao, R.; Shi, Z.; Zou, Z. High-Resolution Remote Sensing Image Captioning Based on Structured Attention. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
40. Wang, Y.; Zhang, W.; Zhang, Z.; Gao, X.; Sun, X. Multiscale Multiinteraction Network for Remote Sensing Image Captioning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 2154–2165. [CrossRef]
41. Yuan, Z.; Li, X.; Wang, Q. Exploring Multi-Level Attention and Semantic Relationship for Remote Sensing Image Captioning. *IEEE Access* **2020**, *8*, 2608–2620. [CrossRef]
42. Kandala, H.; Saha, S.; Banerjee, B.; Zhu, X.X. Exploring Transformer and Multilabel Classification for Remote Sensing Image Captioning. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
43. Ye, X.; Wang, S.; Gu, Y.; Wang, J.; Wang, R.; Hou, B.; Giunchiglia, F.; Jiao, L. A Joint-Training Two-Stage Method for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
44. Hoxha, G.; Melgani, F.; Demir, B. Toward Remote Sensing Image Retrieval Under a Deep Image Captioning Perspective. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4462–4475. [CrossRef]
45. Ma, X.; Zhao, R.; Shi, Z. Multiscale Methods for Optical Remote-Sensing Image Captioning. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 2001–2005. [CrossRef]
46. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the International Conference on Machine Learning PMLR, Virtual, 18–24 July 2021.
47. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision. *arXiv* **2021**, arXiv:2102.05918.
48. Hu, R.; Singh, A. UniT: Multimodal Multitask Learning with a Unified Transformer. *arXiv* **2021**, arXiv:2102.10772.
49. Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; Yang, H. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. *arXiv* **2022**, arXiv:2202.03052.
50. Wang, J.; Yang, Z.; Hu, X.; Li, L.; Lin, K.; Gan, Z.; Liu, Z.; Liu, C.; Wang, L. GIT: A Generative Image-to-Text Transformer for Vision and Language. *arXiv* **2022**, arXiv:2205.14100.
51. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A Visual Language Model for Few-Shot Learning. *arXiv* **2022**, arXiv:2204.14198.
52. Gong, T.; Lyu, C.; Zhang, S.; Wang, Y.; Zheng, M.; Zhao, Q.; Liu, K.; Zhang, W.; Luo, P.; Chen, K. MultiModal-GPT: A Vision and Language Model for Dialogue with Humans. *arXiv* **2023**, arXiv:2305.04790.
53. Qiu, C.; Yu, A.; Yi, X.; Guan, N.; Shi, D.; Tong, X. Open Self-Supervised Features for Remote-Sensing Image Scene Classification Using Very Few Samples. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
54. Li, X.; Wen, C.; Hu, Y.; Zhou, N. RS-CLIP: Zero shot remote sensing scene classification via contrastive vision-language supervision. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103497. [CrossRef]
55. Rahhal, M.M.; Bazi, Y.; Elgibreen, H.; Zuair, M. Vision-Language Models for Zero-Shot Classification of Remote Sensing Images. *Appl. Sci.* **2023**, *13*, 12462. [CrossRef]
56. Ricci, R.; Bazi, Y.; Melgani, F. Machine-to-Machine Visual Dialoguing with ChatGPT for Enriched Textual Image Description. *Remote Sens.* **2024**, *16*, 441. [CrossRef]
57. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
58. Yang, Y.; Newsam, S. Bag-of-visual-words and Spatial Extensions for Land-use Classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, in GIS ’10, San Jose, CA, USA, 3–5 November 2010; ACM: New York, NY, USA, 2010; pp. 270–279. [CrossRef]
59. Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983. [CrossRef]
60. Chiang, W.L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J.E.; et al. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. March 2023. Available online: https://lmsys.org/blog/2023-03-30-vicuna/ (accessed on 2 March 2024).
61. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics—ACL ’02, Philadelphia, PA, USA, 6–12 July 2002; p. 311. [CrossRef]
62. Denkowski, M.; Lavie, A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, ML, USA, 26–27 June 2014; pp. 376–380. [CrossRef]
63. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; p. 10.
64. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based image description evaluation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4566–4575. [CrossRef]
65. Zhang, Z.; Zhang, W.; Diao, W.; Yan, M.; Gao, X.; Sun, X. VAA: Visual Aligning Attention Model for Remote Sensing Image Captioning. *IEEE Access* **2019**, *7*, 137355–137364. [CrossRef]

66. Li, Y.; Zhang, X.; Gu, J.; Li, C.; Wang, X.; Tang, X.; Jiao, L. Recurrent Attention and Semantic Gate for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
67. Cheng, Q.; Huang, H.; Xu, Y.; Zhou, Y.; Li, H.; Wang, Z. NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–19. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.