MCS Deep Learning Project Proposal (~2-3 pages – max 4 pages)

Name	**Damian Falk, Kaan Aydin – ViTal Visions**
Working Title	**Merging Pixels and Words: A Transformer-Enhanced VQA System**
Basic Research Question	*How do transformer-based model architectures, combined with different pre-training methods and fusion techniques, influence the performance of Visual Question Answering (VQA) in the context of remote sensing data?*
Key paper(s)	Chappuis, C., Lobry, S., Kellenberger, B., Le Saux, B., & Tuia, D. (2021). How to find a good image-text embedding for remote sensing visual question answering? arXiv preprint arXiv:2109.11848. Hackel, L., Clasen, K. N., Ravanbakhsh, M., & Demir, B. (2023). LiT-4-RSVQA: Lightweight Transformer-based Visual Question Answering in Remote Sensing. arXiv preprint arXiv:2306.00758. Lobry, S., Demir, B., & Tuia, D. (2021). RSVQA Meets Bigearthnet: A New, Large-Scale, Visual Question Answering Dataset for Remote Sensing. In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS (pp. 1218-1221). Lobry, S., Marcos, D., Murray, J., & Tuia, D. (2020). RSVQA: Visual Question Answering for Remote Sensing Data. IEEE Transactions on Geoscience and Remote Sensing, 58(12), 8555-8566. Siebert, T., Clasen, K. N., Ravanbakhsh, M., & Demir, B. (2022). Multi-Modal Fusion Transformer for Visual Question Answering in Remote Sensing. arXiv preprint arXiv:2210.04510. Songara, J., Pande, S., Choudhury, S., Banerjee, B., & Velmurugan, R. (2023). Visual Question Answering in Remote Sensing with Cross-Attention and Multimodal Information Bottleneck. arXiv preprint arXiv:2306.14264. Srivastava, Y., Murali, V., Dubey, S. R., & Mukherjee, S. (2021). Visual question answering using deep learning: A survey and performance analysis. In Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5 (pp. 75-86). Springer Singapore. Yuan, Z., Mou, L., Wang, Q., & Zhu, X. X. (2022). From easy to hard: Learning language-guided curriculum for visual question answering on remote sensing data. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-11.

Yuan, Z., Mou, L., & Zhu, X. X. (2023). Overcoming Language Bias in Remote Sensing Visual Question Answering via Adversarial Training. arXiv preprint arXiv:2306.00483.

Yuan, Z., Mou, L., & Zhu, X. X. (2023, May). Multilingual augmentation for robust visual question answering in remote sensing images. In 2023 Joint Urban Remote Sensing Event (JURSE) (pp. 1-4). IEEE.

Zhang, Z., Jiao, L., Li, L., Liu, X., Chen, P., Liu, F., ... & Guo, Z. (2023). A spatial hierarchical reasoning network for remote sensing visual question answering. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-15.

Problem & objective

(~15% of content)

The field of Remote Sensing (RS) has witnessed significant advancements in sensing technology, resulting in the acquisition and processing of a substantial volume of visual data. This data holds immense potential for various applications, including tasks like classifying land cover and detecting objects within images. However, the existing methods employed for RS data analysis are primarily rigid and tailored to specific tasks, posing a significant barrier to access and utilization for those without specialized expertise in the field.

Our proposed solution is the Remote Sensing Visual Question Answering System (RSVQA), which aims to address the limitations of current RS data analysis methods. RSVQA offers a more user-friendly and intuitive approach to extracting information from RS data by allowing users to formulate queries in natural language.

RSVQA has the potential to transform RS data analysis by accommodating a broad spectrum of tasks. Users can ask specific questions, such as "Is there a thatched roof in this image?" or delve into more complex inquiries like "Is there a building located to the right of the river?"

Data

(~20% of content)

(1) What data do you propose to use?

We plan on using the RSVQA dataset created by Lobry, S., Demir, B., & Tuia, D. (2021). The dataset consists of remote sensing images (either from Sentinel-2 for the low-resolution set or from the United States Geological Survey for the high-resolution dataset). Additionally, the data set contains various Q&A pairs for each image, which are derived from Open Street Map.

(2) What sample size do you expect? Cross-sectional? In Time-series/longitudinal?

We plan on using the high-resolution dataset, which contains cross-sectional data in the form of 10’659 images and 955’664 question and answer pairs.

(3) What are the labels? What’s about data cleaning?

The dataset already contains labels in the form of question-and-answer pairs. Example questions include: "What is the area covered by commercial buildings”, “What is the number of circular roads”, “Is it a rural or an urban area?”, and “Is there a commercial building?”. Based on our current knowledge, no data cleaning will be required.

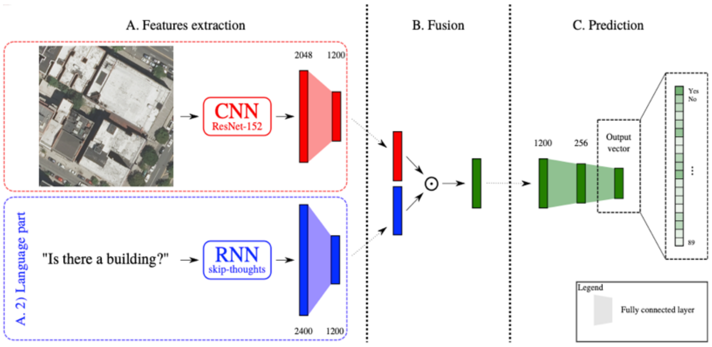
(4) What is the quality/reliability of your data?

The quality and reliability of the RSVQA data can be inferred from its sources and the methodologies used. The images in the high-resolution dataset are orthorectified images from the USGS, which is a reputable source for such data. Moreover, the questions and answers are based on OpenStreetMap (OSM), a widely used and collaborative mapping platform. The project has also been presented in reputable conferences and journals, such as the International Geoscience and Remote Sensing Symposium (IGARSS) and IEEE Transactions on Geoscience and Remote Sensing, which further attest to the quality and reliability of the data.

Approach

(~30% of content)

For the basic project setup, we follow the approach of Lobry, Marcos, Murray & Tuia, (2020), who proposed a model setup with three distinct layers. First, the images and questions are encoded using a pre-trained ResNet-152 in conjunction with a pre-trained RNN model. The representations of the inputs are then fused using a fixed operation with element-wise multiplication of the inputs. This is then passed to a fully connected network for classification, which projects the representation to the answer space.



For the deep learning project, we identified three distinct areas where we would like to modify the existing approach.

Employing alternative vision feature extraction methods

Previous approaches mostly worked with CNNs to encode an image, which was then used together with the question representation (from an RNN) in a fusion and a subsequent classification layer. We propose to use a vision transformer (ViT), a detection transformer (DETR), or both in combination in the first layer to increase accuracy in the various classification tasks by improving vision feature extraction. Specifically, we would like to investigate whether we can improve the accuracy of counting tasks with which current models seem to struggle most. In the scope of this project, we will leave the language module and the classification module largely untouched to ensure comparability to previous results.

Employing various pre-training methods on the same dataset instead of relying on pre-trained models from other datasets

In this part of our approach, we aim to leverage a range of pre-training methods to enhance the capabilities of the RSVQA model. We will explore self-supervised learning (SSL) techniques, training our model on a large remote sensing dataset. This allows our model to learn meaningful representations from the raw data, which might be particularly beneficial when dealing with domain-specific remote sensing information compared to using a pre-trained model from a different image domain (e.g., ImageNet). Specifically, we would like to investigate if using pre-text tasks like inpainting or relative location prediction can improve accuracy on the downstream task.

Exploration of diverse data fusion approaches
In this aspect of our approach, we will delve into various data fusion techniques to effectively integrate information from both the remote sensing images and textual questions in the RSVQA dataset. Previous work employed element-wise multiplication or cross-attention (with information maximization). We would like to include those and potentially other fusion approaches, e.g., multimodal compact bi-linear pooling (MCB), in combination with transformers to further enhance the accuracy of the model.

Experiments

(~20% of content)

To evaluate the performance of our model, we can calculate the accuracy in the following areas:

- Count
- Presence
- Comparison
- Area

as well as the average and overall accuracy. As the baseline, we use the previous research on the topic as outlined before.

We plan to start with the same setup as used in previous research and then successively change parts of the architecture and training procedure. First, we want to incorporate vision transformers (i.e., ViT and DETR), followed by exploration of different pre-training methods and fusion techniques.

What`s new?

(~5% of content)

We propose to use new approaches to image feature extraction as well as the fusion layer to investigate the impact on performance. Therefore, we focus on small modifications to the approach used in previous research, as outlined before.

So what?

(~5% of content)

While VQA in remote sensing is a rather new and unexplored research direction, it has a very high potential for further downstream tasks, allowing end-users to interact with a model more directly. By using an existing approach and trying to improve it, we can incrementally contribute to this field, allowing us to further explore new possibilities. Additionally, this could be used as the foundation for further new approaches, such as incorporating time-series data in VQA systems to enable new sets of questions (e.g., “How has deforestation changed in the last 5 years?”).

Other Considerations

(~5% of content)

In general, the project should be reasonably scoped since we have access to an existing dataset and can build on existing research without implementing something completely new and untested. However, there are still many unexplored approaches to this kind of problem that we can explore, and we will not run out of ideas anytime soon.

We have identified the following moderate risks for this project:

- Limited experience with multi-model and multi-modal training set-ups.
- Risk of obsolescence – LLMs might be able to do this soon for all images.
- Competitors might work on similar approaches.
- Overhead – More data is needed for a transformer to outperform.