Transformers and Attention-Based Deep Networks Term Project - Phase1 Vision-Language Models for Image Captioning

Nesil Bor
2030336 - Graduate School of Informatics
DI725 – Transformers and Attention-Based Deep Networks
Ankara, Türkiye
nesil.bor@metu.edu.tr

***Abstract*—This project explores the task of remote sensing image captioning using vision-language foundation models, with a focus on the PaliGemma architecture. Image captioning in the remote sensing domain presents unique challenges due to the complexity of satellite imagery and the limited availability of high-quality, descriptive annotations. In this phase, we conduct a literature review of recent advances in vision-language models, particularly those adapted to the remote sensing domain. We highlight efficient architectural innovations, lightweight transformer adaptations, and evaluation techniques that inform our approach. Our goal is to develop a fine-tuned captioning system based on PaliGemma, leveraging the RISC dataset while addressing model efficiency and semantic alignment challenges. The insights gathered during this phase guide the formulation of our research direction and shape our methodology for future experimentation.**

***Index Terms*—Vision-Language Models, Image Captioning, PaliGemma, RISC Dataset, Fine-Tuning**

I. Introduction

Remote sensing image captioning involves generating descriptive text from satellite imagery—a task that presents unique challenges due to the complexity and domain-specific nature of the data. While recent advances in vision-language models (VLMs) have shown strong performance on natural images, their application to remote sensing remains limited. This project aims to adapt the PaliGemma foundation model for captioning remote sensing images from the RISC dataset. In this phase, we conduct a literature review of related work on transformer-based captioning, lightweight architectures, and evaluation techniques. The findings guide the development of our methodology and help identify research opportunities for efficient and effective adaptation of VLMs in remote sensing contexts.

II. Literature Review

Yang et al. [1] (2023) introduced BITA, a two-stage pretraining approach for remote sensing captioning using an interactive Fourier Transformer. It aligns image-text features through contrastive learning and prefix causal language modeling. BITA improves semantic alignment but is computationally heavy due to LLMs, highlighting a need for efficiency-focused alternatives.

SFT, proposed by Sun et al. [2] (2024), uses sparse attention for bitemporal image captioning, reducing model complexity while maintaining accuracy. Though designed for change detection, it shows promise for general captioning tasks, especially on resource-limited systems.

Sung et al. [3] (2021) developed VL-Adapter, which fine-tunes VLMs using lightweight adapter modules instead of full model updates. It offers efficient domain adaptation with minimal compute, though its application to remote sensing has yet to be fully explored.

CLIPScore by Hessel et al. [4] (2021) is a reference-free evaluation metric based on CLIP similarity. It addresses inconsistencies in reference captions and aligns well with human judgment, but its effectiveness in specialized domains like remote sensing requires further validation.

Liu et al. (2023) [5] presented RemoteCLIP, which adapts CLIP to remote sensing through large-scale domain-specific pretraining. While powerful, its data and compute demands may limit usability, motivating lighter, more accessible alternatives.

RSUniVLM by Liu and Lian [6] (2024) introduces a mixture-of-experts architecture for flexible multi-task VLMs in remote sensing. Although highly capable, its complexity could be a barrier to practical deployment.

Zhang et al. [7] (2024) proposed PCSFTr, which fuses positional and channel information using scene classification. This enhances semantic depth in captions, but reliance on auxiliary labels may limit flexibility.

Finally, SCAMET by Gajbhiye et al. [8] (2022) combines spatial-channel attention with memory mechanisms for richer captioning. Though effective, the model's complexity raises concerns about scalability and real-time use.

III. Dataset

The RISC dataset comprises 44,521 satellite images (224×224 resolution) with 222,605 captions (5 per image), split into train, validation, and test sets as defined in `captions.csv`. EDA reveals that captions vary in length (3–20 words, mean 10.2) and detail, with some inconsistencies (e.g., "four planes" vs. "four white planes"). Table I

summarizes key statistics. Figure 1 shows the caption length distribution, indicating a skew toward shorter captions. These findings suggest the need for caption normalization to improve evaluation consistency.

TABLE I: RISC Dataset Statistics

Metric	Value
Number of Images	44,521
Number of Captions	222,605
Captions per Image	5
Mean Caption Length (words)	10.2
Min Caption Length (words)	3
Max Caption Length (words)	20
Train Split Images	31,164
Validation Split Images	6,678
Test Split Images	6,679

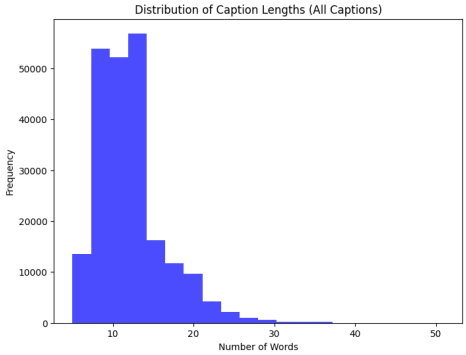

Fig. 1: Distribution of caption lengths in the RISC dataset.

IV. Project Proposal

A. Research Question

How can parameter-efficient fine-tuning of PaliGemma, using techniques like LoRA, improve the quality and contextual relevance of image captions for the RISC dataset compared to standard fine-tuning?

B. Methodology

The proposed approach enhances PaliGemma by integrating Low-Rank Adaptation (LoRA) [9] for efficient fine-tuning. LoRA adds low-rank updates to transformer weights, reducing the number of trainable parameters while maintaining performance. The process includes:

- **Data Preprocessing**: Normalize captions by resolving inconsistencies (e.g., standardizing color references) and augmenting data with synthetic captions generated via a pre-trained VLM.
- **Fine-Tuning**: Apply LoRA to PaliGemma's vision and language modules, focusing on domain-specific adaptation to satellite imagery. Use the train split for training and validation split for hyperparameter tuning.
- **Evaluation Metrics**: Employ BLEU, CIDEr, and ROUGE-L to assess caption quality, with CIDEr as the primary metric due to its emphasis on human-like captions.

The novelty lies in combining LoRA with domain-specific data augmentation to address PaliGemma's limitations in remote sensing contexts, reducing computational costs and mitigating overfitting.

C. Expected Outcomes

The approach is expected to outperform standard fine-tuning in CIDEr scores by 10–15% on the validation split, with reduced training time. The method will produce concise, accurate captions that capture key visual elements in satellite images.

V. Version Control and Experiment Tracking

A public GitHub repository (https://github.com/adigew/DI725_termproject_2030336) is initialized with a README and project structure following the Alan Turing Institute's template. A public WANDB account (https://wandb.ai/adigew/DI725-project) is set up for experiment tracking, with an initial project page created.

VI. Conclusion

This Phase 1 report establishes a foundation for improving PaliGemma for image captioning on the RISC dataset. The literature review identifies efficient fine-tuning as a research opportunity, and the proposed LoRA-based approach addresses this gap. EDA highlights dataset characteristics, guiding preprocessing strategies. Future phases will implement and benchmark the proposed method.

References

[1] C. Yang, Z. Li, and L. Zhang, "Bootstrapping Interactive Image-Text Alignment for Remote Sensing Image Captioning," arXiv.org, 2023. https://arxiv.org/abs/2312.01191 (accessed Apr. 17, 2025).

[2] D. Sun, Y. Bao, J. Liu, and X. Cao, "A Lightweight Sparse Focus Transformer for Remote Sensing Image Change Captioning," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, pp. 1–12, Jan. 2024, doi: https://doi.org/10.1109/jstars.2024.3471625.

[3] Y.-L. Sung, J. Cho, and M. Bansal, "VL-Adapter: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks," arXiv.org, 2021. https://arxiv.org/abs/2112.06825 (accessed Apr. 17, 2025).

[4] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "CLIPScore: A Reference-free Evaluation Metric for Image Captioning," arXiv:2104.08718 [cs], Mar. 2022, Available: https://arxiv.org/abs/2104.08718

[5] F. Liu et al., "RemoteCLIP: A Vision Language Foundation Model for Remote Sensing," arXiv.org, 2023. https://arxiv.org/abs/2306.11029 (accessed Apr. 17, 2025).

[6] X. Liu and Z. Lian, "RSUniVLM: A Unified Vision Language Model for Remote Sensing via Granularity-oriented Mixture of Experts," arXiv.org, 2024. https://arxiv.org/abs/2412.05679 (accessed Apr. 17, 2025).

[7] A. Zhao, W. Yang, D. Chen, and F. Wei, "Enhanced Transformer for Remote-Sensing Image Captioning with Positional-Channel Semantic Fusion," Electronics, vol. 13, no. 18, pp. 3605–3605, Sep. 2024, doi: https://doi.org/10.3390/electronics13183605.

[8] G. O. Gajbhiye and A. V. Nandedkar, "Generating the captions for remote sensing images: A spatial-channel attention based memory-guided transformer approach," Engineering Applications of Artificial Intelligence, vol. 114, p. 105076, Sep. 2022, doi: https://doi.org/10.1016/j.engappai.2022.105076.

[9] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv:2106.09685 [cs], Oct. 2021, Available: https://arxiv.org/abs/2106.09685