Image Captioning with PaliGemma Project Phase 1

İbrahim Orcan Ön
Multimedia Informatics
METU
Ankara, Turkey
orcan.on@metu.edu.tr

***Abstract*—This project proposes an efficient fine-tuning approach to enhance the image captioning capability of the transformer-based vision-language model PaliGemma, specifically targeting remote sensing imagery. We conduct a comprehensive literature review highlighting state-of-the-art captioning techniques, vision-language transformer adaptations, and existing methodologies addressing domain-specific challenges in satellite imagery. Leveraging insights from this analysis, we propose introducing lightweight adapter modules to efficiently adapt PaliGemma to the RISC dataset, which contains 44,521 satellite images and corresponding descriptive captions. Initial exploratory data analysis reveals key characteristics of the dataset, including caption length, vocabulary distribution, and frequent spatial descriptors. Appropriate evaluation metrics—BLEU, METEOR, CIDEr are identified to rigorously assess the model's performance. This work lays a structured foundation for subsequent implementation and benchmarking phases, addressing crucial domain-adaptation challenges in remote sensing image captioning.**

Github:https://github.com/orcanon/MMI725_Final_Project.git
Wandb:https://wandb.ai/orcanon1-metu/MMI725_Final_Project

***Index Terms*—ViT, Vision-Language Model, Image Captioning**

I. Literature Review

Research Question: How can we efficiently adapt the PaliGemma vision-language model to generate high-quality captions for remote sensing images? Specifically, we investigate whether a parameter-efficient fine-tuning approach (such as adapter modules) can improve caption accuracy on the RISC dataset, while being computationally feasible and preserving the model's general vision-language knowledge.

Image Captioning Models: Early image captioning systems adopted an encoder–decoder framework: a convolutional neural network (CNN) encodes the image, and a recurrent neural network (RNN) decodes a caption, often with attention mechanisms to attend to salient regions. A landmark example is the "Show, Attend and Tell" model using CNN+LSTM with spatial attention (ICML 2015). Further improvements incorporated object detection features and attention. Anderson et al. (2018) introduced the Bottom-Up and Top-Down Attention model, which uses a pre-trained object detector (Bottom-Up) to propose region features, and an attention LSTM (Top-Down) to weight these features when generating each word. This "bottom-up/top-down" approach became a widely used baseline [1] for captioning and significantly improved content selection in captions. Subsequent approaches replaced RNNs with Transformers for both image and text encoding. For example, transformer-based captioners take grid or region features as a sequence input to a transformer decoder, enabling better long-range dependency modeling in the generated description. Xu et al. (2020) and Cornia et al. (2020) introduced fully-attentive caption models with improved attention refinement (e.g. the Meshed-Memory Transformer).

Vision-Language Transformers: The advent of transformer-based VLMs trained on massive image-text data has pushed captioning to new heights. SimVLM (Wang et al., 2021) proposed a simplified transformer encoder-decoder trained on weakly labeled image-caption pairs, achieving excellent zero-shot captioning. PaLI (Pathways Language–Image model by Chen et al., Google 2022) scales this paradigm to a multilingual 17-billion-parameter model, combining a ViT image encoder with a text decoder. PaLI was trained on 10B image-text pairs and achieved state-of-the-art results on captioning benchmarks [2], while maintaining a modular encoder–decoder design. These works demonstrate the benefit of scaling up and jointly training vision and language components. However, such models are extremely large; fine-tuning them on a new domain is computationally expensive. This has led to research in efficient fine-tuning for VLMs. VL-Adapter (Sung et al., CVPR 2022) introduced lightweight adapter layers for vision-language models (e.g. VL-BART, VL-T5), inserting small bottleneck modules within the transformer blocks instead of updating all weights [3]. With only 4% of parameters trained, VL-Adapter matched full fine-tuning performance on image captioning (COCO) and other tasks [3].

Remote Sensing Image Captioning: Generating captions for overhead imagery is a niche within captioning, with its own dedicated datasets and methods. Early remote sensing image captioning (RSIC) works used CNN encoders (often pre-trained on ImageNet) with RNN decoders, similar to natural image captioners [5]. However, remote images often cover broad areas with multiple objects and varying scales, making it challenging for models trained on natural images to capture all relevant details. Li et al. (2020) identified severe overfitting problems in RSIC models (due to relatively limited data and repetitive scenes) and proposed a Truncation Cross Entropy (TCE) loss that effectively alleviates overfitting [6]

by relaxing the training objective. A recent study by Lin et al. (2024) notes that current approaches typically just fine-tune generic VLMs on RSIC data, rather than developing models tailored to remote sensing imagery's unique characteristics [7]. This presents an opportunity to apply foundation models like PaliGemma to RSIC, while introducing targeted adaptations (e.g. domain-specific modules or training strategies) to bridge the gap. Our work will build on this literature by leveraging a pre-trained VLM and incorporating an efficient fine-tuning method to better suit the remote sensing captioning task.

II. Project Proposal - Methodology

Proposed Approach: I propose to enhance the PaliGemma model for remote sensing image captioning by integrating lightweight adapter modules for fine-tuning. Rather than updating all of PaliGemma's millions (or billions) of parameters, I will insert small neural layers (adapters) at select points in the image encoder and/or language decoder. Each adapter consists of a down-projection to a small latent size, a non-linearity, and an up-projection back to the original size. During training, only these adapter weights (and possibly layer normalization biases) are updated, while the original model weights remain frozen. This technique dramatically reduces trainable parameters (often to ¡5% of the full model [3]) and has proven effective in preserving pre-trained knowledge [3]. By using adapters, we aim to efficiently transfer PaliGemma's general vision-language capability to the remote sensing domain without overfitting or requiring excessive computational resources. Concretely, we will design adapters for PaliGemma's architecture. The PaliGemma encoder (SigLip) is likely a Vision Transformer, which produces a sequence of image features. The decoder (Gemma) is a transformer-based language model generating the caption. I can insert adapters after the self-attention and feed-forward sublayers in these transformers. I will experiment with two settings: (1) Vision-side adapters only, where the image representation are allowed to adjust to remote imagery (capturing domain-specific visual features), and if i have enough time; (2) Dual adapters on both the vision and language side, allowing some adaptation in language decoding.

Novelty and Justification: While adapter-based fine-tuning has been studied in general V&L tasks [3], applying it to a multi-modal remote sensing captioning scenario is relatively novel. Previous RSIC works mostly trained full models or fine-tuned conventional CNN-RNN models; none have used the adapter approach on a large foundation model for this task. This method is novel in that it bridges a state-of-the-art pre-trained VLM (PaliGemma/PaLI family) with the remote sensing domain through efficient adaptation. This avoids training a new model from scratch and leverages rich pre-trained knowledge (common objects, scenes, language patterns) while still allowing domain-specific refinement. The project will proceed as follows. First, I will perform baseline fine-tuning of PaliGemma on the RISC training set to establish a performance reference (using standard training, possibly with some regularization). Next, I will implement the adapter modules and train the model with most weights frozen. I will monitor training behavior (validation loss, metrics) to see if it converges faster or avoids overfitting compared to the baseline.

A. *Evaluation Metrics*

To evaluate the image captioning performance, I will use standard quantitative metrics from the image captioning literature [1] , as well as qualitative inspection: **BLEU (Bilingual Evaluation Understudy):** I will compute BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores which measure n-gram overlap between the generated caption and the reference captions. For instance, BLEU-4 looks at 4-word sequences. These are precision-based metrics that reward overlapping words, weighted by shorter length to avoid verbosity. BLEU is a core metric used in COCO evaluations, but it often correlates poorly with human judgment for captions, so I interpret it with caution (especially since having multiple reference captions mitigates some issues of BLEU). **METEOR:** This metric computes alignment between candidate and reference using synonyms and stemming, and combines precision/recall with a penalty for word order mistakes. It might be more sensitive to small differences and has shown better correlation with human judgment than BLEU in some cases. We will report METEOR to account for synonyms (for example, if the model says "automobile" vs reference "car", BLEU would score zero for that word but METEOR can give partial credit). **CIDEr (Consensus-based Image Description Evaluation):** CIDEr is specifically designed for captioning tasks. It uses TF-IDF weighting for n-grams, rewarding n-grams that are important (common across the reference captions for an image) and penalizing those that might be common in the corpus but not specific to that image. This helps ensure the generated caption is image-specific and not a generic sentence. A high CIDEr score usually indicates the caption captured the unique details of the image well. I expect CIDEr to be a primary metric of success. I will primarily consider CIDEr and SPICE for judging caption quality (as they emphasize relevance and content). BLEU (especially higher-order BLEU-3,4) will be secondary, as it is very strict. In addition to these automatic metrics, we will perform qualitative analysis: manually inspecting a sample of generated captions, particularly for failure cases. For example, I will check if the model hallucinated any objects.

B. *Dataset*

- The dataset is imbalanced, heavily skewed towards training data, which is typical for machine learning tasks.
- Frequent usage of spatial descriptors (e.g., "next", "area", "by") and natural elements (e.g., "green", "trees") highlights the remote sensing context, suggesting captions emphasize geographical features, vegetation, and urban structures.
- The high occurrence of generic words (e.g., "the", "a", "are") could pose challenges in generating precise and discriminative captions, potentially requiring careful fine-tuning or prompt engineering to mitigate generic responses.

APPENDIX

TABLE I
DATASET OVERVIEW

Attribute	Value
Total Images	44,521
Total Captions	222,605
Captions per Image	5
Splits	**Number of Images**
Train	35,614
Validation	4,453
Test	4,454
Data Sources	**Number of Images**
NWPU	31,500
RSICD	10,921
UCM	2,100

TABLE II
CAPTION LENGTH STATISTICS

Statistic	Value
Average Caption Length	12.09 words
Standard Deviation	4.22 words
Minimum Length	5 words
25% Quartile	9 words
Median Length	11 words
75% Quartile	14 words
Maximum Length	51 words
Vocabulary Size	3,720 unique words

TABLE III
TOP 20 MOST FREQUENT WORDS IN CAPTIONS

Rank	Word	Occurrences
1	the	205,467
2	a	117,754
3	are	110,698
4	and	100,628
5	is	77,975
6	there	73,976
7	of	69,650
8	some	62,091
9	many	61,774
10	green	61,480
11	on	55,452
12	in	54,672
13	trees	51,503
14	buildings	47,186
15	with	43,122
16	to	26,766
17	next	24,675
18	area	24,371
19	by	21,306
20	two	18,925

REFERENCES

[1] P. Zhang et al., "Vinvl: Revisiting visual representations in vision-language models," CVF Open Access, https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_VinVL_Revisiting_Visual_Representations_in_Vision-Language_Models_CVPR_2021_paper.html (accessed Apr. 19, 2025).

[2] X. Chen et al., "Pali: A jointly-scaled multilingual language-image model," arXiv.org, https://arxiv.org/abs/2209.06794#: :text=large%20multilingual%20mix%20of%20pretraining,simple%2C%20modular%2C%20and%20scalable%20design (accessed Apr. 19, 2025).

[3] Y.-L. Sung, J. Cho, and M. Bansal, "VL-adapter: Parameter-efficient transfer learning for vision-and-language tasks," CVF Open Access, https://openaccess.thecvf.com/content/CVPR2022/html/Sung_VL-Adapter_Parameter-Efficient_Transfer_Learning_for_Vision-and-Language_Tasks_CVPR_2022_paper.html #: :text=this%20paper%2C%20we%20introduce%20adapter,by%20sharing%20their%20weights %20to (accessed Apr. 19, 2025).

[4] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," arXiv.org, https://arxiv.org/abs/2301.12597#:~:text=Flamingo80B%20by%208.7%25%20on %20zero,can%20follow%20natural%20language%20instructions (accessed Apr. 19, 2025).

[5] Y. Yang et al., "Remote Sensing Image Change captioning using multi-attentive network with diffusion model," MDPI, https://www.mdpi.com/2072-4292/16/21/4083#: :text=,RS%20images%20for%20change%20captioning (accessed Apr. 19, 2025).

[6] Author links open overlay panelQiaoqiao Yang et al., "Meta captioning: A meta learning based Remote Sensing Image captioning framework," ISPRS Journal of Photogrammetry and Remote Sensing, https://www.sciencedirect.com/science/article/abs/pii/S0924271622000351#: :text=Meta%20captioning%3A%20A%20meta%20learning,fitting %20problem.%20Sun (accessed Apr. 19, 2025).

[7] RS-Moe: Mixture of experts for remote sensing image captioning and visual question answering, https://arxiv.org/html/2411.01595v1 (accessed Apr. 19, 2025).