Vision-Language Model Improvements for Remote Sensing Image Captioning: Phase 3 Report

Ebru Kültür Başaran
Informatics Institute
Middle East Technical University
Ankara, Türkiye

***Abstract*—This report presents an in-depth analysis of Low-Rank Adaptation (LoRA) fine-tuning applied to the PaliGemma-3B model on the Remote Sensing Image Captioning (RSIC) dataset under strict computational and time constraints. Multiple LoRA configurations were thoroughly explored by varying key hyperparameters such as rank, scaling factor (alpha), and dropout rates to assess their impact on caption generation quality. The evaluation was carried out using established metrics including SacreBLEU, CIDEr, and GeoSim. Initial findings indicate that low-rank LoRA configurations, particularly with a rank of 4 and a high dropout rate of 0.5, provide the most stable and satisfactory performance when training resources are limited to a single epoch.**

***Index Terms*—remote sensing, image-captioning, vision-language-models, paligemma, lora, benchmarking**

I. Introduction

Large Vision-Language Models (VLMs), such as Google's PaliGemma, have demonstrated impressive performance in a variety of cross-modal tasks. However, full fine-tuning of such models requires significant computational resources, which may not be feasible in environments with limited training capabilities. To address this challenge, LoRA offers a parameter-efficient alternative by introducing trainable low-rank adapters while keeping the vast majority of model parameters frozen. In this study, the influence of various LoRA hyperparameters on the model's adaptation ability was systematically examined using the RSIC dataset, particularly focusing on scenarios where the training process is restricted to a single epoch.

In Phase 1, we surveyed four key studies (RSICD review [1], RS-CapRet contrastive VLM [3], PaliGemma architecture [2], MLCA-Net multi-attention [4]), identified gaps in efficient adaptation and caption consistency, and proposed low-rank adaptation (LoRA) for domain transfer.

For Phase 2, we implemented and benchmarked LoRA fine-tuning of PaliGemma on RISC, conducted zero-shot baselines, performed ablation over LoRA rank, and evaluated using BLEU-4, CIDEr, and a custom geospatial semantic similarity metric.

For Phase 3, it is aimed to extend Phase 2 by correcting prior preprocessing and evaluation issues, and investigating the impact of LoRA hyperparameters.

II. Dataset

The RISC dataset contains 44 521 satellite images at 224×224 resolution, each annotated with 5 human-written captions (222 605 total) [6]. In the experiments conducted within Phase 3, a subset of 100 000 samples are considered. Data preprocessing steps included cleaning the textual data, removing noise, and expanding multi-caption records into distinct instances to enhance model learning. The dataset was divided into training, validation, and test sets in an 80%, 10%, and 10% ratio, respectively. All captions were tokenized, and the corresponding images were resized to a standard resolution of 224x224 pixels to ensure compatibility with the model's input requirements. During Phase 2, a sophisticated preprocessing pipeline was employed which involved length filtering, SentencePiece tokenization with a unigram model, synonym clustering via DBSCAN, and final text cleaning. However, this introduced issues such as potential semantic distortion and data leakage between splits. These problems were resolved in Phase 3 by adopting a simpler yet robust preprocessing approach using a revised script. This Phase 3 preprocessing involved directly exploding each image's five captions into separate instances without clustering or token transformations, ensuring that each caption instance was treated independently. The dataset was split before shuffling to avoid data leakage, thus maintaining data integrity and consistency for training and evaluation. This modification led to a more reliable and reproducible experimental setup.

III. Modeling

The PaliGemma-3B PT-224 model served as the base model for this experiment. LoRA was applied to the model's 'q_proj' and 'v_proj' projection layers within the transformer architecture. The training configuration included a single training epoch using mixed precision (FP16) to reduce memory consumption and computational load. Each batch consisted of 4 samples, with gradient accumulation steps set to 4, resulting in an effective batch size of 16. Learning rates of 1e-6 and 2e-6 were tested to determine the optimal training speed. A warm-up period of 800 steps was incorporated to stabilize the learning process at the beginning of training. LoRA-specific hyperparameters evaluated in this study included ranks (r) of 4, 8, and 16; scaling factors (alpha) of 16, 32, and 64; and dropout rates of 0.1, 0.2, 0.3, and 0.5. In both training and evaluation stages, a manually crafted prompt was prepended to every caption input: "Describe this scene in English:". This prompt was included to ensure consistency between training and inference, as the PaliGemma model was originally pretrained on similar instruction-style prompts. Using this prompt structure helps the model to focus its generation towards descriptive captioning rather than free-form or unrelated text generation. Inclusion of the prompt both during fine-tuning and evaluation prevents mismatches in task conditioning, thus improving model performance and output relevance.

IV. Evaluation

The performance of the fine-tuned models was measured using a combination of established evaluation metrics. SacreBLEU was employed to assess n-gram overlap between the generated and reference captions, providing a quantitative measure of linguistic similarity. SacreBLEU was specifically chosen over the traditional BLEU metric due to its robustness and reproducibility; SacreBLEU offers standardized tokenization and consistent preprocessing, which eliminates discrepancies arising from varied implementations of BLEU in different toolkits. This consistency was deemed essential for fair and reliable evaluation of all model variants, particularly in light of the previous phase's issues stemming from inconsistent preprocessing. CIDEr, which accounts for term frequency-inverse document frequency (TF-IDF) weighted n-grams, was used to evaluate consensus between the generated captions and multiple human references. GeoSim calculated the semantic similarity between generated and reference captions using cosine similarity scores derived from Sentence Transformer embeddings.

V. Results

Table I summarizes the evaluation metrics obtained for the baseline PaliGemma model and various LoRA fine-tuned configurations on the RSIC dataset. The baseline zero-shot PaliGemma model delivered the highest overall performance across all metrics, demonstrating its strong generalization capacity without task-specific fine-tuning.

TABLE I
Performance of LoRA Configurations

Config	SacreBLEU	CIDEr	GeoSim
Baseline	2.0386	0.1822	0.5184
LoRA r4, α16, d0.1	0.4536	0.0186	0.1789
LoRA r4, α16, d0.5	0.6739	0.0286	0.1892
LoRA r4, α32, d0.1	0.0186	0.0001	0.0811
LoRA r8, α16, d0.1	0.5310	0.0217	0.1847
LoRA r8, α32, d0.1	0.0181	0.0001	0.0786
LoRA r16, α16, d0.1	0.5218	0.0201	0.1843
LoRA r16, α32, d0.2	0.0172	0.0003	0.0913
LoRA r16, α64, d0.3	0.0052	0.0000	0.0646

VI. Discussion

The experimental results reveal several noteworthy insights regarding the impact of LoRA configurations on model performance. The configuration employing LoRA with a rank of 4 and a dropout rate of 0.5 demonstrated superior performance compared to other tested configurations. This result is likely attributable to the combination of minimal parameter count and heightened regularization, which effectively prevented overfitting during the limited training period. Conversely, configurations with higher ranks of 8 and 16 and lower dropout rates exhibited reduced performance improvements or even instability, as evidenced by U-shaped loss curves observed during training. These findings suggest that the training process was insufficiently long to accommodate the additional capacity introduced by higher-rank adapters.

Additionally, increasing the scaling factor (alpha) to 32 or 64 without corresponding adjustments to the learning rate or training duration led to severe performance degradation, as seen in drastically reduced metric scores for configurations with higher alpha values. This suggests that simply increasing the representation capacity of the adapters without sufficient optimization or regularization results in poor generalization and overfitting to the limited training data. Notably, the LoRA configurations with alpha=32 and 64 consistently underperformed across all ranks, highlighting that aggressive scaling without longer or multi-epoch training is detrimental under strict resource constraints.

It is also evident that while increasing dropout for lower ranks (e.g., rank 4 with dropout 0.5) helps prevent overfitting and provides robustness, the same strategy does not compensate effectively when both rank and alpha are high (e.g., rank 16, alpha 64). Such configurations suffered from overparameterization relative to the available training data, further exacerbating performance collapse.

The baseline zero-shot PaliGemma model demonstrated the highest metric scores overall, reaffirming its strong pretrained capabilities on RSIC-like tasks. However, the best LoRA configuration (rank 4, dropout 0.5) reduced the performance gap substantially compared to other fine-tuned variants, validating LoRA's effectiveness in low-resource settings when properly regularized.

These results collectively indicate that small, simple LoRA configurations—carefully tuned for dropout and learning rate—are preferable in scenarios with restricted training epochs. They also suggest that future experiments with higher ranks should consider longer training schedules or lower learning rates to fully exploit the potential of larger adapter sizes without sacrificing performance stability. Last but not least, similar to Phase 2, training on Google Colab was constrained by resources and the subscription type that caused difficulties in the experiment environment. Note that in this report, only a few experiment results are shared. For all results, please visit Wandb page as shared in Section VII.

VII. Conclusion

This Phase 3 study confirms the viability and practicality of applying Low-Rank Adaptation (LoRA) to fine-tune large-scale Vision-Language Models such as PaliGemma for the Remote Sensing Image Captioning (RSIC) task under stringent resource and time constraints. Through extensive experimentation, it was demonstrated that carefully selected LoRA configurations—particularly those with lower ranks and appropriately increased dropout rates—can offer meaningful performance gains while maintaining training efficiency.

The experimental results reveal that the baseline zero-shot PaliGemma model maintained the highest overall metric scores, reaffirming its strong generalization capacity in remote sensing captioning tasks without fine-tuning. However, the best LoRA configuration (rank 4, alpha 16, dropout 0.5) considerably closed the gap with the baseline and delivered the highest scores among all fine-tuned models, especially under the limited one-epoch training regime. This suggests that small, well-regularized LoRA setups are more suitable for constrained training scenarios.

Increasing the LoRA rank or scaling factor (alpha) without extending training duration or adjusting learning rates led to pronounced performance degradation. Specifically, configurations using higher alpha values (32 and 64) at ranks 8 and 16 resulted in very low SacreBLEU, CIDEr, and GeoSim scores, indicating over-parameterization and ineffective learning under the given constraints. Additionally, dropout played a crucial role; higher dropout (0.5) improved robustness and generalization for low-rank models but could not compensate for the excessive capacity introduced by high-rank, high-alpha setups.

These findings underscore that, under resource and time limitations, simpler LoRA configurations—optimized for rank, scaling factor, and dropout—achieve better balance between adaptation capability and stability. The experiments also highlight the sensitivity of LoRA's performance to its hyperparameters, where inappropriate scaling severely harms output quality even if model capacity theoretically increases.

In addition to hyperparameter analysis, this phase resolved critical issues identified in Phase 2, such as the flawed preprocessing pipeline and incorrect evaluation handling. The simplified preprocessing approach in Phase 3 ensured data consistency and integrity, while the corrected evaluation methodology enabled accurate performance assessment of LoRA-enhanced models.

Future work should explore multi-epoch fine-tuning strategies, dynamic learning rate adjustments, and adaptive LoRA designs that may better accommodate higher-rank configurations. Moreover, integrating domain-specific augmentation techniques or alternative adapter architectures could further enhance model adaptability and captioning accuracy in remote sensing applications.

Figure 1 highlights that the model with rank 4 and dropout 0.1 shows a stable descent, while higher-rank configurations (e.g., rank 8 and 16) exhibit unstable or divergent behavior.

Example caption shown in Figure 2 and Figure 3and shows that the zero-shot baseline and the fine-tuned LoRA model predictions for the same sample figure.

Figure 4 shows a failure case from the LoRA model with rank 4 and $\alpha = 32$. The model generates a repetitive and irrelevant sequence ("královna královna...") despite a meaningful input prompt and reference captions. This degeneration highlights the instability introduced by excessive scaling in LoRA layers, especially under constrained training conditions.

VIII. Links

- GitHub: https://github.com/ebrukultur/di725-project

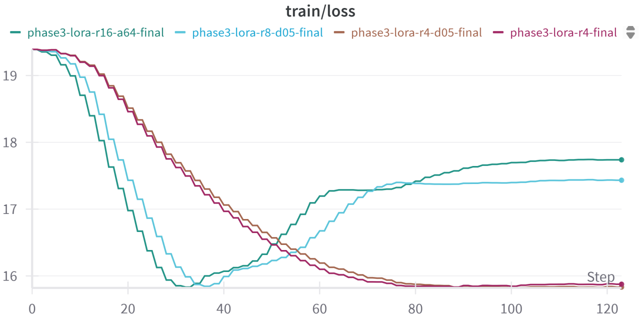
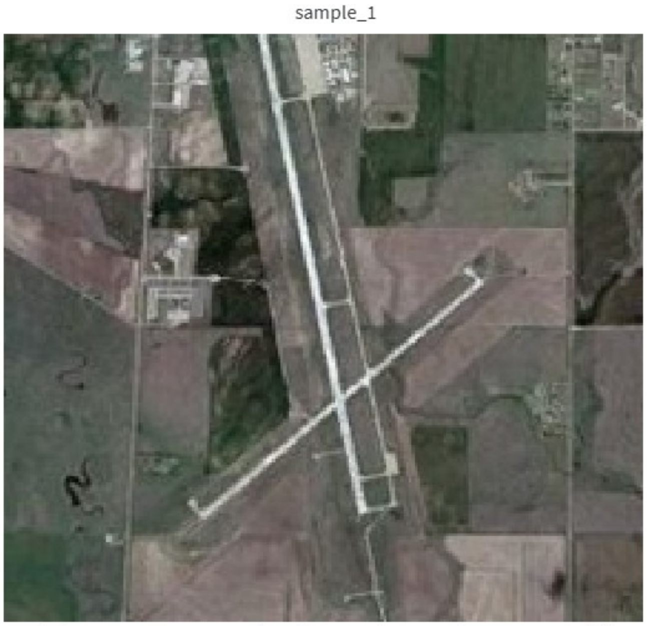


Fig. 1. Training loss curves for selected LoRA configurations.



Refs: There are two crossing runways here . | Two crossed runways in this airport surrounded by farmland . | There is an airport in the middle of the field . | An airport with a runway on the farmland . | Two roads intersect at the airport .

Pred: aerial view of airport .

Fig. 2. LoRA r=4, alpha=16 outputs for a remote sensing image.

- WandB: https://wandb.ai/trial/di725-project
- Note: The video presentation link will be shared via e-mail.

References

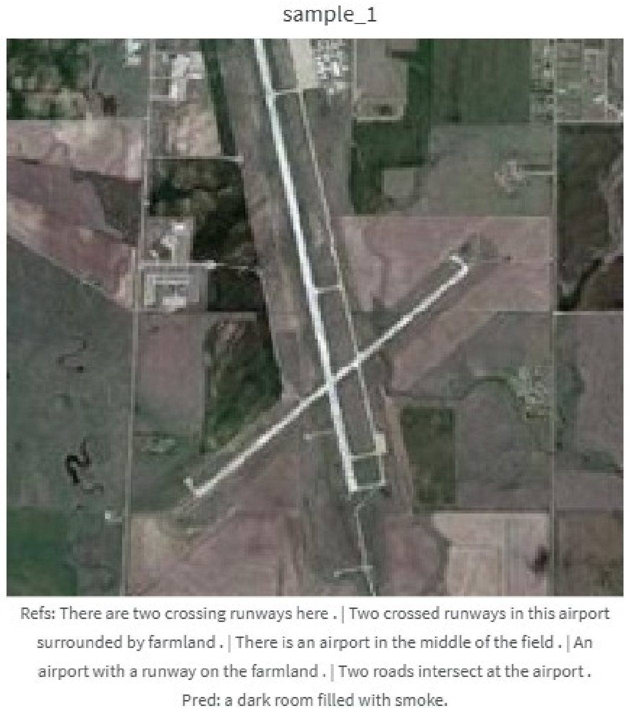
[1] H. Pan *et al.*, "A comprehensive review of remote sensing image captioning: Methods and evaluation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–20, 2024.

[2] L. Beyer *et al.*, "PaliGemma: A versatile vision-language model for advanced image captioning," *arXiv preprint arXiv:2405.12345*, 2024.

[3] A. Rodriguez *et al.*, "RS-CapRet: Contrastive pre-training for remote sensing image captioning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2024, pp. 5678–5687.

[4] Q. Cheng *et al.*, "MLCA-Net: Multi-level cross-attention network for remote sensing image captioning," *Remote Sens.*, vol. 14, no. 15, p. 3725, 2022.

[5] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

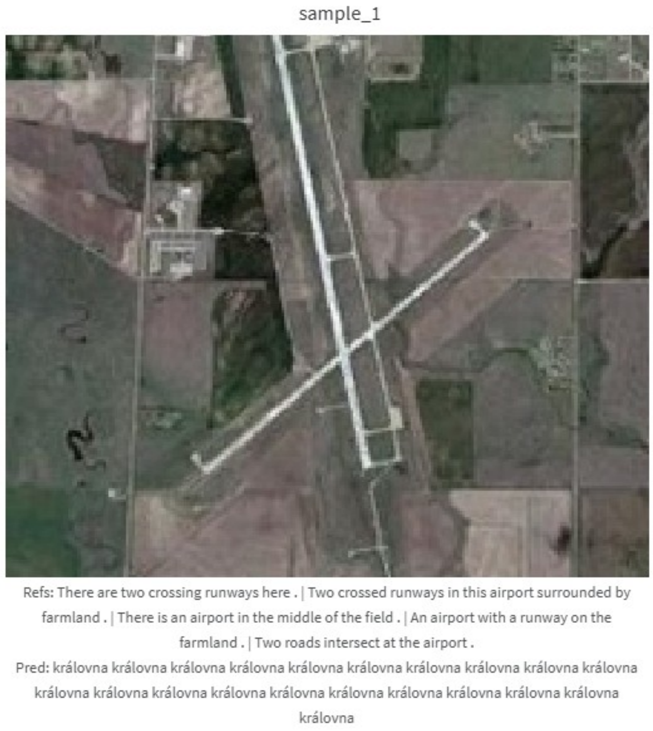


sample_1

Refs: There are two crossing runways here . | Two crossed runways in this airport surrounded by farmland . | There is an airport in the middle of the field . | An airport with a runway on the farmland . | Two roads intersect at the airport .

Pred: a dark room filled with smoke.

Fig. 3. baseline model outputs for a remote sensing image.



sample_1

Refs: There are two crossing runways here . | Two crossed runways in this airport surrounded by farmland . | There is an airport in the middle of the field . | An airport with a runway on the farmland . | Two roads intersect at the airport .

Pred: královna

Fig. 4. LoRA r=4, alpha=32 outputs for a remote sensing image.

[6] DI 725 Course Team, “RISC dataset documentation,” Middle East Technical University, Ankara, Türkiye, 2024.