Evaluating Prompt Learning Strategies for Remote Sensing Vision-Language Foundation Models

Stanford CS229

Ethan Hellman* hellman1@stanford.edu
Rodrigo Nieto* rjnieto@stanford.edu
Spencer Paul* spaul2@stanford.edu

1 Introduction

The advancement of language processing through models like BERT[1], GPT [2], and CLIP [3]has been a game-changer. Yet, the potential of multi-modal vision-language models (VLMs) in addressing critical issues like climate change remains largely untapped. Our project explores this frontier, leveraging AI and remote sensing—a combination proving vital in sectors from agriculture to urban planning—to tackle climate change. In this study, we explore the application of prompt learning strategies to enhance vision-language models (VLMs), an approach that could significantly improve AI's role in environmental conservation. Our research is focused on addressing two key questions:

1. Which prompt learning strategies, among CoOp[4], MaPLe[5], and PromptSRC[6], are most effective for remote sensing tasks?
2. How does prompt learning affect domain-specific models compared to general models like CLIP[3] and RemoteCLIP[7]?

Our investigation involves a rigorous evaluation of these strategies on two foundational models, across two remote sensing datasets (EuroSAT[8] and RESISC45[9]), focusing on the accuracy of remote sensing scene classification. This work doesn't just aim to contribute to the academic discourse but seeks to open new avenues for AI applications in environmental conservation.

2 Related Work

2.1 Vision-Language Models

Advancements in vision-language models like CLIP (Contrastive Language-Image Pretraining) by OpenAI have revolutionized multi-modal data processing with their zero-shot learning capabilities and alignment of text and image representations [10]. Furthermore, MAE (Masked Autoencoders) extends to vision-language contexts, addressing the convergence of language and vision model geometries [11][12]. ViLBERT and LXMERT set new benchmarks in multimodal processing with their innovative architectures [13][14][15]. VisualBERT and FLAVA contribute with their simpler yet effective approaches in multimodal interpretation [16][17].

2.2 Remote Sensing Vision-Language Foundation Models

Remote sensing vision-language models have witness recent evolution with RemoteCLIP (2023) adapting the CLIP framework for remote sensing [7], SatMAE (2022) addressing satellite imagery challenges [18], and Dino-MC (2023) introducing self-supervised learning [19]. These models represent significant advancements in remote sensing analysis, improving data interpretation and handling diverse environmental scenarios. That being said, the mantle of state-of-the art is still yet to be claimed and substantiated by a comprehensive study.

*Department of Computer Science, Stanford University

2.3 Prompt Learning

Prompt learning has shifted from manual methods to automated approaches like Context Optimization (CoOp), which introduces learnable vectors for context words in prompts [4]. Conditional Context Optimization (CoCoOp) and Multi-modal Prompt Learning (MaPLe) further this evolution by enhancing generalizability and adapting both vision and language branches [20][5]. Recent methods like RetroPrompt and HealthPrompt expand the application of prompt learning with retrieval-augmented and clinical-text-focused techniques [21][22], demonstrating a trend towards more efficient prompt learning strategies. While further study is merited, MaPLe remains the state-of-the-art in published prompt-learning strategies.

3 Dataset and Features

Our study evaluates various Vision-Language Models (VLMs) using two prominent remote sensing image classification datasets: EuroSAT [8] and RESISC45 [9].

3.1 EuroSAT

EuroSAT is based on Sentinel-2 satellite imagery and comprises 27,000 images categorized into 10 classes, including AnnualCrop, Forest, HerbaceousVegatation, Highway, Industrial, Pasture, PermanentCrop, Residential, River, and SeaLake. Each image is 64x64 pixels with a 10m spatial resolution. The dataset is divided into 13,500 training, 5,400 validation, and 8,100 test images. To align with our focus on few-shot learning scenarios, the models are trained on a limited subset of the training and validation images, while evaluations are conducted on the complete test set.

3.2 Remote Sensing Image Scene Classification (RESISC)

The RESISC45 dataset encompasses 31,500 remote sensing images from nearly 100 countries, distributed across 45 scene classes. Each class comprises 700 images of 256x256 pixels with RGB channels. The spatial resolution ranges between 30 to 0.2 meters per pixel for most scenes. The dataset split includes 15,750 training, 6,300 validation, and 9,450 test images.

4 Methods

4.1 Prompt Learning Strategies

The Context Optimization (CoOp)[4] framework presented in this study advances the adaptation of pre-trained vision-language models for specific image recognition tasks. CoOp innovatively employs learnable vectors to model the context words in prompts, leaving the original pre-trained model parameters unaltered. It features two key variants: a unified context model applicable across all classes, and a class-specific context model with unique context tokens for each class. To illustrate, the prompt $\mathbf{t}$ that is fed to the text encoder $g(\cdot)$ (to generate a classification weight vector) can be represented as the following,

$$\mathbf{t} = [V]_1[V]_2...[V]_M[CLASS], \tag{1}$$

where each $[V]_m (m \in \{1, ..., M\})$ is a vector with a dimension of 512 and M indicates the number of context tokens [4]. This method is optimized using cross-entropy loss to minimize prediction errors. Significantly, CoOp demonstrates superior performance on various datasets compared to manually crafted prompts and enhances data efficiency overall. We chose to include this approach in our study given that CoOp largely kicked off the study of prompt learning and sets an effective baseline against other approaches.

Next, we utilize MaPLe [5], a novel prompt learning strategy that builds off of the work of CoOp. MaPLe is distinct in several ways. Firstly, in addition to the language prompting outlined in CoOp, MaPLe employs vision prompting in the image encoder as well. The prompts in the vision branch are conditioned on the language prompts via a coupling function to enforce synergy between them. This coupling function is learned with the prompts fine-tuning. Secondly, MaPLe uses a deep prompting strategy in both the vision and language branches of the model; in addition to adding context to

the initial text and image inputs, each transformer layer in the vision and text encoders receives a learned prompt derived during fine-tuning. The prompted layer depth is a hyperparameter that can be tuned. For our method, we opt to use a prompt depth of 9 since that was found to be optimal in the MaPLe paper. This approach was selected given that it has remained as the state-of-the-art for prompt learning over the past couple of years.

Lastly, we explore PromptSRC [6], a more recent framework for fine-tuning vision-language models that promises to address overfitting issues observed in other prompt learning strategies. This approach leverages a three-pronged self-regulating approach: 1) Mutual Agreement Maximization, aligning prompted features with frozen model features; 2) Self-Ensembling of prompts, using a weighted aggregation across the training trajectory; and 3) Textual Diversity, enriching text encodings for each class. This framework guides prompt optimization for task-specific and general representations, enhancing both performance on downstream tasks and preserving the generalization capabilities of the pre-trained model. Extensive experiments across various benchmarks demonstrate PromptSRC's efficacy in maintaining robust generalization while optimizing task-specific knowledge. It is for these reasons, and the added potential of beating SOTA, that we include PromptSRC in our study.

4.2 Model Selection

Given the widespread popularity of CLIP-based models, ease of implementation, and inability to compare across all remote sensing foundation models due to resource constraints, we elected to study the behavior of RemoteCLIP [7], the *first* vision-language foundation model tailored for remote sensing. It addresses the scarcity of pretraining data in this domain by leveraging a novel data scaling technique, which converts heterogeneous annotations into a unified image-caption format. This approach uses Box-to-Caption (B2C) and Mask-to-Box (M2B) conversions to enhance data diversity. RemoteCLIP, trained on this enriched dataset, excels in various downstream tasks, including zero-shot image classification and object counting, outperforming existing models and demonstrating robust generalization across diverse datasets. This model marks a significant advancement in the application of vision-language models to remote sensing and promises to serve as a strong backbone for studying various prompt learning strategies.

5 Experiments/Results/Discussion

5.1 Overview

Our study aims to provide novelty in 3 ways:

1. **Consistent Comparison**: While all three of the aforementioned prompt learning strategies have been successful on a variety of datasets including EuroSAT, there has yet to be a consistent comparison of them in the context of remote sensing. Other inconsistencies, such as MaPLe only reporting results in a 16-shot setting, make comparing approaches difficult. Additionally, the authors claim better results than CoOp while training for fewer epochs (5 v 10). Finally, papers may also use differing backbones and versions of CLIP. As such, in our study, we provide consistent comparison across models for more effective comparison.
2. **Deeper Exploration of Remote Sensing**: EuroSAT is the only remote sensing dataset used in previous evaluations of these strategies. It is a relatively easy dataset given its limited number of classes. By introducing RESISC45, we offer deeper insight into how these models perform in the remote sensing domain.
3. **Domain-Specific Foundation Models vs. General Models Learning Prompts on Domain-Specific Tasks**: We address the question of whether it is more efficient to prompt-tune a generally pretrained foundation model or a domain-specific finetuned model for remote sensing tasks. Additionally, we provide insight into how prompt learning strategies relatively improve generally pretrained and domain-specific finetuned foundation model performance.

5.2 Experimental Settings

To mirror real-world scenarios where data is limited, we evaluate all combinations of models + prompting in a few-shot setting (1, 2, 4, 8, and 16 shots). The number of shots is the number of

training examples in each class that the models see during training. Datasets are sampled with a common seed to ensure consistency in training data. Training is done for 10 epochs in all experiments with a batch size of 4. For both CLIP and RemoteCLIP, we use a ViT-B/32 backbone and an ADAM optimizer. All prompts are initialized to "a photo of a," following the handcrafted prompts in the RemoteCLIP study[7]. In addition to shots, we experiment with the number of context tokens in prompts with 2 and 4 context token length settings. Existing prompt learning research will often test for generalization to novel classes. For the scope of our project, we allowed the models to see all classes they are evaluated on during training. This change along with modifying hyperparameters explains the discrepancy in results for similar experiments in prior work.

5.3 Results

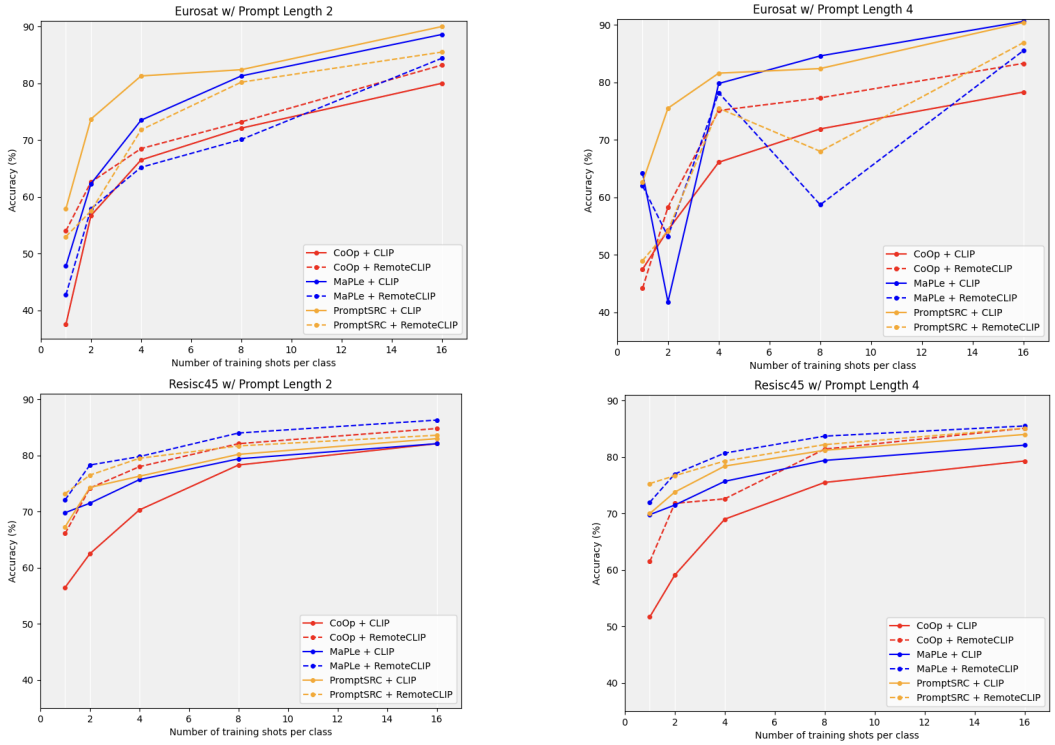

Figure 1: Comparison of accuracy scores amongst different models and prompt strategies across EuroSAT and RESISC45 with varying prompt lengths and shots.

5.4 Discussion

5.4.1 EuroSAT Performance

MaPLe + CLIP with a prompt length of 4 is the strongest performing model + prompt learning strategy pair on EuroSAT achieving an accuracy of 90.6% in the 16-shot setting. This substantially outperforms RemoteCLIP and CLIP baselines of around 81% and 79% in the 16-shot setting respectively[7]. PromptSRC + RemoteCLIP with a prompt length of 4 is the highest-performing domain-specific model achieving an accuracy of 86.9% which also notably outperforms the CLIP and RemoteCLIP baselines. Interestingly, the PromptSRC + CLIP performance is quite strong regardless of the number of shots, outperforming CLIP 16-shot accuracy with only 4 shots. The only prompt learning strategy that underperformed compared to hand-crafted prompt baselines is CoOp with both CLIP and RemoteCLIP.

5.4.2 RESISCS45 Performance

We can see that, unlike the results on EuroSAT, the RemoteCLIP models with prompt learning strategies outperformed the vanilla CLIP model on the RESICS45 dataset. The highest accuracy on

RESISC45 was achieved by MaPLe + RemoteCLIP in the 16-shot setting with a prompt length of 2 yielding an accuracy of 86.3%. Although these results differ from those on EuroSAT, they meet expectations given that the vanilla RemoteCLIP model largely outperforms the vanilla CLIP model on RESISC45 (with a delta of $+9.41$ [7]). However, we see that the CLIP model with prompt learning is actually able to close the gap between the accuracies on RemoteCLIP compared to CLIP without prompting as tested in the RemoteCLIP study [7]. (86.3% to 82.1% constituting a delta of $+4.2$ instead of $+9.41$) [7].

5.4.3 Shot Analysis

For all graphs in Figure 1, we can expect better accuracies across all settings as the number of training shots per class increases. However, it is interesting to note that some model and prompt learning combinations vastly improve with changes in the number of shots, such as MaPLe with CLIP compared with PromptSRC with CLIP. Additionally, we can see that the impact of prompt learning with an increased number of shots looks different for the two datasets. More specifically, it appears that prompt learning with a greater number of shots has a greater impact on performance with the EuroSAT dataset compared to increasing the number of shots on given the RESISC45 dataset.

5.4.4 Model Comparison

CLIP with prompt learning on average beats RemoteCLIP with prompt learning on the EuroSAT dataset across the board. However, RemoteCLIP with prompt learning generally does better than CLIP with prompt learning on RESISC. We believe this discrepancy can be attributed to the difficulty of the datasets. Given EuroSAT's limited class size and variability, a general foundation model may be able to generalize to it easily with a limited number of fine-tuned prompt samples. While RemoteCLIP outperforms CLIP on RESISC45, the boost in performance comes mainly from the model being fine-tuned on thousands of satellite images, not prompt learning. Notably, however, prompt learning closes the overall gap between the two models suggesting that generally pretrained foundation models with domain-adapted prompts are more efficient learners than domain-specific finetuned foundation models with domain-adapted prompts. This potentially has significant implications given the time, data, and computation costs associated with training a domain-specific foundation model. Whereas training a RemoteCLIP or similar domain-specific model can achieve SOTA performance with domain-adapted prompts, similar performance can be seen with generally pretrained vision-language models. This inherently raises questions about the trade-offs between performance and resource constraints.

5.4.5 Prompt Length

Overall, prompt length seems to have minimal impact on performance. The only notable observation is that MaPLe performance is less stable across shots with a prompt length of 4. We see dips at two and eight shots for CLIP and RemoteCLIP. This suggests that MaPLe may be less robust to fewer-shot training and is arguably why their authors only included 16-shot performance results in their paper.

6 Conclusion/Future Work

In conclusion, we find that prompt learning can be an effective strategy to boost performance on remote sensing datasets for vision-language models. Additionally, we see the strongest prompting strategies are MaPLe and PromptSRC. Finally, we conclude that when altering foundation models for the domain-specific task of remote sensing it would appear more data, compute, and time-efficient to prompt tune a general foundation model like CLIP rather than train a domain-specific foundation model and subsequently perform additional prompt tuning. While remote sensing vision-language models can see performance gains over generally pretrained models, the difference is minimal. Thus, we question the merit of extensive pretraining for such marginal gains.

While we conclude that it is more effective to prompt tune a foundation model rather than train a domain-specific foundation model, this hypothesis could be more rigorously explored with more out-of-domain remote sensing datasets as well as comparison across different remote sensing vision-language foundation models. Despite our preliminary analysis, a more holistic study could further support our findings.

7 Contributions

All team members contributed equally to developing the methodology, running experiments, and writing the paper.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[4] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, jul 2022.

[5] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning, 2023.

[6] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting, 2023.

[7] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing, 2023.

[8] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

[9] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, Oct 2017.

[10] Jishnu Jaykumar P, Kamalesh Palanisamy, Yu-Wei Chao, Xinya Du, and Yu Xiang. Proto-clip: Vision-language prototypical network for few-shot learning, 2023.

[11] Jiaang Li, Yova Kementchedjhieva, and Anders Søgaard. Implications of the convergence of language and vision model geometries, 2023.

[12] Minhyeok Lee, Dogyoon Lee, Jungho Lee, Suhwan Cho, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Synchronizing vision and language: Bidirectional token-masking autoencoder for referring image segmentation, 2023.

[13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

[14] Michele Cafagna, Kees van Deemter, and Albert Gatt. What vision-language models 'see' when they see scenes, 2021.

[15] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. X-lxmert: Paint, caption and answer questions with multi-modal transformers, 2020.

[16] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language, 2019.

[17] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model, 2022.

[18] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery, 2023.

[19] Xinye Wanyan, Sachith Seneviratne, Shuchang Shen, and Michael Kirley. Dino-mc: Self-supervised contrastive learning for remote sensing imagery with multi-sized local crops, 2023.

[20] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models, 2022.

[21] Xiang Chen, Lei Li, Ningyu Zhang, Xiaozhuan Liang, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Decoupling knowledge from memorization: Retrieval-augmented prompt learning, 2023.

[22] Sonish Sivarajkumar and Yanshan Wang. Healthprompt: A zero-shot learning paradigm for clinical natural language processing, 2022.