FEW-SHOT LEARNING PIPELINES ON REMOTE SENSING IMAGES

Orr Avrech
Columbia University
oa2429@columbia.edu

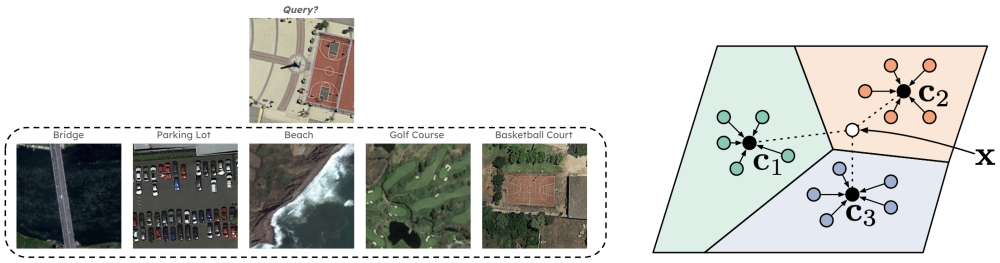
Abstract

The field of remote sensing plays an important role in understanding and mitigating the effects of climate change. However, the specialized nature of this data and its costly annotation processes result in an abundance of unlabeled data and a scarcity of annotated samples. To address this issue, this paper introduces a framework centered around few-shot learning (FSL) techniques. By effectively learning from a limited pool of labeled examples and capitalizing on external unlabeled data sources using self-supervised methods, this approach aims to enable more informed decision-making based on satellite images.

1 Introduction

Supervised deep learning demonstrates exceptional performance when vast annotated datasets are available. Despite the careful curation of several large-scale remote sensing image datasets in recent years—examples include fMoW [2], EuroSat [7], OPTIMAL-31 [12], among others — the annotation process for these datasets demands specialized skills and domain knowledge. Such requirements often lead to higher costs compared to traditional computer vision datasets. In scenarios where there is an abundance of unlabeled data and scarcity in labeled data, notably in remote sensing data, solely relying on supervised learning methods is not sufficient. Furthermore, a multitude of climate change-related tasks heavily rely on satellite imagery, spanning from land use classification to crop type detection. This underscores the importance of being able to generalize well to multiple tasks in this field. Consequently, exploring techniques in few-shot learning, which aims to learn new concepts and tasks only from a limited number of training examples, can have a significant impact on computer vision tasks for climate change. Although FSL deals with limited data scenarios, leveraging existing large-scale external data sources can significantly impact its practical application. Recent works in remote sensing imagery aim to leverage the intrinsic and distinctive attributes of this data. SatMAE [3] used a Masked Autoencoder (MAE) [5] for temporal and multi-spectral remote sensing data. ScaleMAE [9] introduced another pre-training approach based on an MAE, with the objective of learning relationships between data at different scales, aiming to provide robust multi-scale representations.

In this work, our objective is to propose a general framework for few-shot learning in remote sensing tasks. Our approach involves exploring important design considerations and leveraging knowledge from different domains while accommodating the distinct characteristics of our data. In particular, we study the effectiveness of few-shot learning pipelines using self-supervised pre-training and examine their suitability for remote sensing images. We check whether the latest advancements in specialized image foundation models tailored to this domain yield superior results. We also wish to establish an initial benchmark in few-shot learning for remote sensing — a domain that remains relatively under-explored, despite being evidently compelling for climate change-related tasks.



(a) (b)

Figure 1: Example of a meta-training task. (a) A 5-way,1-shot episode taken from OPTIMAL-31. Given a query image (top), find an image (if 1 shot) or set of images (if $k > 1$) with the lowest distance in the learned embedding space. (b) In ProtoNets, prototypes c_k are computed as the average embedding of all support examples for each class. Here, we observe 5 support examples for each of the 3 support class (3w5s). Then, given a query image x, its prediction will be the class with the closest associated prototype in the embedding space (c_2 in this example).

2 Method

2.1 Problem Formulation

The objective of few-shot classification is to produce a model which, given a new learning episode with N classes and a few labeled examples k per class $1, ..., N$, is able to generalize to unseen examples for that episode. In other words, the model learns from a support set $S = \{(x_1, y_1), (x_2, y_2), ..., (x_K, y_K)\}$, where $x \in \mathbb{R}^D$ is an input vector and $y \in \{1, ..., N\}$ is the class label, and is evaluated on an unseen query set. A learning episode is usually referred to as 'N-way, k-shot' to indicate N classes and k samples per class for the support.

2.2 Pre-training

We explore two well-established feature extractors: ResNet [6] and ViT [4]. Initially, we focus on ImageNet-21k pre-training models and subsequently aim to integrate externally acquired data learned through self-supervised techniques. Recent advancements in self-distillation methods, such as DINO [1], fundamentally rely on a straightforward approach: generating self-supervision by exposing two different views to two encoders and aligning one with the other using a predictor. DINOv2 [8] combines the principles outlined in DINO with masked image modeling, showing promising results in both linear and k-NN evaluations. As highlighted in Section 1, masked image modeling techniques currently dominate the landscape of self-supervision in remote sensing data ([3], [9]). This prevalence likely stems from its adaptability to multi-channel (spectral), multi-domain, and multi-scale features. Therefore, evaluating the performance of a dedicated foundational model on satellite images could significantly contribute to our study that proposes a general FSL framework for remote sensing tasks.

2.3 Prototypical Networks

The fundamental idea behind prototypical networks [10] is to learn an embedding space where examples from the same class cluster around a prototype representation. To achieve this, the network learns to map input data into an embedding space, which ideally allows for clear separation and clustering of different classes. Each class' prototype is taken as the mean of its support set in the embedding space. When presented with a new, unlabeled query example, the network embeds this query into the same learned embedding space. Classification is then performed by identifying the closest prototype to this embedded query point, often using distance metrics like Euclidean distance or cosine similarity. Put differently, prototypical networks learn a prototype representation, $c_k \in \mathbb{R}^M$ for each class through a mapping $f_\phi : \mathbb{R}^D \rightarrow \mathbb{R}^M$, where parameters ϕ are learnable. Each prototype is represented as the mean of the support embeddings, $c_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} f_\phi(x_i)$, where S_k denotes the examples labeled with class k in the support set.

Training Configuration				OPTIMAL-31	
ID	Architecture	PreTrain	Meta-Train	5w1s	5w5s
0	ViT-base/16	Sup. (ImageNet-21k)	-	69.4	90.4
1	ViT-base/16	DINOv2	-	**81.3**	**94.9**
2	ResNet-50	Sup. (ImageNet-21k)	-	70.5	92.6
3	ViT-base/16	DINOv2	ProtoNet	78.3	92.3
4	ViT-large/32	ScaleMAE (fMoW)	-	45.7	56.9

Table 1: The effects of architecture, pre-training, and meta-training on downstream few-shot learning performance on OPTIMAL-31 test split using nearest-prototype classification.

2.4 Meta-training

As discussed in Section 2.3, ProtoNets only require a feature extractor to map data points to a learned embedding space. During meta-training, the network uses support sets (few examples per class in each episode) to compute class prototypes within this learned embedding space. The training process involves optimizing the network's parameters to minimize a loss that measures the distance between predicted prototypes and the true class centroids within the embedding space.

3 Experiments

3.1 Setup

Datasets We used two known satellite image benchmarks for our study. OPTIMAL-31 [12] comprises 1860 RGB images across 31 general scene classes. EuroSAT [7] is a land use classification dataset with 10 classes, featuring images from Sentinel-2 available in both RGB and multi-spectral versions. Inspired by previous work on Meta-Dataset [11], we applied similar training and evaluation protocols. Our aim is to encourage future exploration in this domain and to stimulate the development of a larger meta-dataset tailored specifically to remote sensing images and tasks related to climate change and societal impact. In our case, we used a training/validation/test split across the classes of OPTIMAL-31 for both meta-training and evaluation purposes, reserving EuroSAT solely for evaluation.

Evaluation For evaluating the few-shot classification performance, we generate 100 episodes (tasks) from the test split of each dataset. The metric used for evaluation is the mean classification accuracy across these tasks, where we assume nearest-centroid (euclidean) classifier as in ProtoNets. We adopt the convention of evaluating episodes in two configurations: 5-way,1-shot (5w1s) and 5-way,5-shot (5w5s). In each episode, the query set size remains fixed at 30.

3.2 Analysis

Architecture and Pre-training First we evaluate the impact of model architecture selection and the designated pre-training regime. As shown in Table 1, the comparison between ResNet and ViT-base (row 0 vs. row 2) pre-trained on ImageNet-21k shows negligible differences. However, when employing DINOv2 - a self-supervised technique trained on ImageNet-21k as well - a notable improvement becomes evident compared to the supervised pre-training approach (row 3). In fact, this result is quite impressive and surprising: with no training whatsoever for the task at hand, we achieve over 80% (5w1s) accuracy and nearly 95% (5w5s) accuracy on the OPTIMAL-31 test split. This implies that current self-supervised methods trained on natural images yield valuable representations applicable to satellite images as well.

Remote Sensing Representation Learning The initial idea was to use domain-specific self-supervised learning methods, aiming for a more robust common representation adaptable to learning from minimal examples given new tasks. Our choice of ScaleMAE [9], pre-trained on fMoW [2], served as the foundation model under examination. Unfortunately, the results of this pre-training method were very disappointing. Even after carefully trying to reproduce their results on OPTIMAL-31, the adaptation to meta-testing did not work well, demonstrating inferior results to ImageNet

pre-trained representations. Assuming accurate integration of their code, this outcome signals substantial room for improvement in learning robust representations for remote sensing images.

Meta-training In fact, meta-training with ProtoNets has demonstrated instability and considerable sensitivity to hyperparameters. Initially prone to overfitting, the model showed rapid learning on training tasks while struggling to generalize across validation and test tasks. Fine-tuning parameters and incorporating standard data augmentations like flips and rescaling led to similar results on both OPTIMAL-31 and EuroSAT, compared to pre-training alone. In addition, we experimented with non-episodic training, where we use standard supervised training on the dataset's train split and extracted the final feature layer. However, we saw no clear improvement in applying episodic or non-episodic training compared to the simple pre-training with nearest-centroid evaluation.

References

[1] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers, 2021.

[2] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. Functional map of the world, 2018.

[3] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. B. Lobell, and S. Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery, 2023.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[5] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners, 2021.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[7] P. Helber, B. Bischke, A. Dengel, and D. Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2019.

[8] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[9] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning, 2023.

[10] J. Snell, K. Swersky, and R. S. Zemel. Prototypical networks for few-shot learning, 2017.

[11] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol, and H. Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples, 2020.

[12] Q. Wang, S. Liu, J. Chanussot, and X. Li. Scene classification with recurrent attention of vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):1155–1167, 2019.