Enhancing Active Learning for Sentinel 2 Imagery through Contrastive Learning and Uncertainty Estimation

David Pogorzelski
Helmholtz-Zentrum Hereon Sediment
Transport and Morphodynamics
Geesthacht, Germany
david.pogorzelski@hereon.de

Peter Arlinghaus
Helmholtz-Zentrum Hereon Sediment
Transport and Morphodynamics
Geesthacht, Germany

Abstract— **In this paper, we introduce a novel method designed to enhance label efficiency in satellite imagery analysis by integrating semi-supervised learning (SSL) with active learning strategies. Our approach utilizes contrastive learning together with uncertainty estimations via Monte Carlo Dropout (MC Dropout), with a particular focus on Sentinel-2 imagery analyzed using the Eurosat dataset. We explore the effectiveness of our method in scenarios featuring both balanced and unbalanced class distributions. Our results show that for unbalanced classes, our method is superior to the random approach, enabling significant savings in labeling effort while maintaining high classification accuracy. These findings highlight the potential of our approach to facilitate scalable and cost-effective satellite image analysis, particularly advantageous for extensive environmental monitoring and land use classification tasks.**

Note on preliminary results: This paper presents a new method for active learning and includes results from an initial experiment comparing random selection with our proposed method. We acknowledge that these results are preliminary. We are currently conducting further experiments and will update this paper with additional findings, including comparisons with other methods, in the coming weeks.

Keywords—Active Learning Sentinel S2, Contrastive Learning, Uncertainty Search, Multispectral Data

I. Introduction

The utilization of satellite imagery analysis has become an essential tool across diverse sectors, encompassing environmental monitoring, urban planning, and disaster response. With the continuous expansion in satellite data availability, there is a corresponding surge in the demand for sophisticated analytical techniques capable of efficiently processing this data and deriving actionable insights. Particularly, tasks such as land use classification and semantic segmentation of satellite images play pivotal roles in tracking and understanding temporal changes in landscapes. Nevertheless, these critical tasks are often hindered by the requirement for extensive labeled datasets, which are both costly and labour-intensive to create.

Active learning is a strategic approach to mitigate the burden of labeling by strategically sampling a subset of the training data pool. This subset ideally represents the broader dataset sufficiently to train effective models with minimal data. Although the concept of active learning has been around for some time, recent strides in deep learning within the realm of remote sensing have demonstrated a promising shift towards self-supervised learning (SSL) methods [2], which have shown considerable success in the active learning domain.

Despite the growing inclination towards using SSL for remote sensing applications, its integration within active learning strategies for remote sensing data remains underexplored. In this paper, we introduce a novel methodology to the existing repertoire of active learning frameworks, specifically tailored for the underdeveloped area of remote sensing active learning algorithms. This contribution aims to bridge the gap in the application of SSL in active learning, enhancing the efficiency and accuracy of remote sensing data analysis.

II. Related Work

A. Active Learning

Active learning is a subset selection methodology utilized in machine learning, particularly when labeled data is scarce or costly to obtain [1]. The fundamental goal of active learning is to strategically select an unlabelled subset S_U from an unlabelled training set D_U in order to train a model with a much lower amount of data, i.e. $|S_U| \ll |D_U|$, while aiming for a similar model performance. The cardinality or the sample size of S is a hyperparameter. The selection is optimized to maximize the performance on a specific task T, e.g. classification or semantic segmentation. Let S_U^* denote the set of all subsets of D_U with size N each. Mathematically, the objective of active learning can then be expressed as

$$\mathrm{S}'_{\mathrm{U}} = \underset{\mathrm{S_U} \in\, \mathrm{S}^*_{\mathrm{U}}}{\operatorname{argmax}}\ \mathrm{J(T)}$$

where J describes the performance. The strategy of maximizing the function is called *query strategy* and the effectiveness of S'_U is typically evaluated through metrics that measure model performance, such as accuracy, precision, or a domain-specific evaluation criterion, reflecting the active learning cycle's contribution towards achieving more informative and representative training samples. In a neural network setting, the performance is usually measured by a loss function. Given a neural network $f(x)$ and a corresponding loss function $L(x)$, the active learning objective can be expressed as

$$\mathrm{S}'_{\mathrm{U}} = \underset{\mathrm{S_U} \in \mathrm{S}^*_{\mathrm{U}}}{\operatorname{argmin}} \mathrm{L}\big(\mathrm{f(x)}\big)$$

The process can be repeated which makes it an *iterative active learning* scheme:

$$\mathbf{S}_{\mathbf{U}}^{(\mathbf{i+1})} = \underset{\mathbf{S^{(i)}}_{\mathbf{U}} \in \mathbf{S}_{\mathbf{U}}^{*(\mathbf{i})},\ \mathbf{S}_{\mathbf{U}}^{(\mathbf{i})} \cap \mathbf{S}_{\mathbf{L}}^{(\mathbf{i})} = \emptyset}{\mathbf{argmin}} \mathbf{L}\left(\mathbf{f}_{\mathbf{S}_{\mathbf{L}}^{(\mathbf{i})}}(\mathbf{x})\right)$$

$$\mathbf{S}_{\mathbf{L}}^{(\mathbf{i+1})} = \mathbf{S}_{\mathbf{L}}^{(\mathbf{i})} \cup \mathbf{a}\big(\mathbf{S}_{\mathbf{U}}^{(\mathbf{i+1})}\big)$$

$$\mathbf{S}_{\mathbf{U}}^{*(\mathbf{i+1})} = \mathbf{S}_{\mathbf{U}}^{*(\mathbf{i})} \setminus \mathbf{S}_{\mathbf{L}}^{(\mathbf{i+1})}$$

where $f_{S^{(i)}_L}$ describes the model learned on the labelled data S_L, a(x) denotes the mapping from unlabelled to labelled data. The process is repeated for all i until a convergence criteria is met, e.g. a predefined accuracy. The advantage of iterative active learning is the continual enhancement of selection decisions with each iteration.

B. Contrastive Learning

A common technique in the training of deep learning models is the combination of pre-training and fine-tuning, together called transfer learning.

Pre-training is the process of training a neural network on a large, carefully pre-processed dataset to develop a foundational model that can be adapted for various tasks. This foundational model captures general features that are useful across different domains. By transferring the learned parameters from the pre-trained model to a new model, one can leverage these pre-existing insights. This method is especially useful as it allows the new model to start from an advanced point of learning. The next step involves fine-tuning, where the pre-trained model is further trained on a specific dataset of interest, allowing it to specialize and adapt to the particular nuances of that data. This approach significantly accelerates the training process and improves the model's performance on specialized tasks. Often, a small amount of data is already enough to fine-tune a pre-trained model.

This approach is well-established in domains like image processing, which benefit from the availability of large datasets. However, in fields where labeled data is scarce, traditional supervised learning approaches are not viable. In such cases, alternative methods such as *self-supervised learning* (SSL) are utilized. SSL leverages unlabeled data to learn feature representations that are useful for downstream tasks. This method enables the extraction of meaningful patterns from data without the need for labels.

A broad overview of SSL in remote sensing is given in [2].

SSL can be used for different objectives like reconstructing data, predicting self-produced labels, or learning a representation that maps semantically similar inputs close together in the representation/ feature space f such that $|f(x_1) - f(x_2)| \to 0$.

In this paper, we focus on the latter part, also referred to as *contrastive learning*.

To map semantically similar inputs together, the model could trivially map all instances to a common point where the similarity task would be fulfilled. However, this would lead to a poor representation since dissimilar points would share similar representations. This is called *model collapsing*. One popular way to prevent model collapsing, is by the means of *negative sampling* which incorporates negative, dissimilar samples as well:

$$sim\big(f(x), f(x^+)\big) \gg sim\big(f(x), f(x^-)\big)$$

where x is referred to as the *anchor point* and x^+as the positive sample and x^- as the negative sample. f is the encoding function, usually a neural network, and sim describes a similarity function where a larger value describes higher similarity between two objects.

There are many ways to generate positive and negative samples for an anchor point and to incorporate these in a training framework.

A common way is to augment the anchor point by transformations like rotation, cropping etc. Depending on the loss function, multiple positive samples are created. A prominent for this is the NT-Xent (Normalized Temperature-scaled Cross Entropy) loss:

$$L(i,j) = -\log\left(\frac{\exp\big(\mathrm{sim}(z_i, z_j)/\tau\big)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(\mathrm{sim}(z_i, z_k)/\tau)}\right)$$

Given a data instance x from the training set D with size $D = |N|$, z_i and z_j are two different, but similar views on x. $z_k$includes all other samples plus z_j, the positive pair. The number of summands is $2N$ since two views are generated for each instance of x. z_i is also referred to as the query, z_j and z_k as key. τ is a hyperparameter and controls the decision boundary between positive and negative samples. A large τ leads to a softer decision boundary and vice versa. The indicator function $\mathbb{1}$ ensures that the query itself is not included in the comparison.

A popular SSL model is MoCo [3]. [2] showed that MoCo generally outperforms other well-known methods in remote sensing classification tasks. Additionally, it was shown that models fine-tuned with 50% labeled data can achieve similar performance to those trained with full supervision. A crucial finding is the significant role of random cropping as a data augmentation strategy in enhancing the effectiveness of the encoder. The Eurosat dataset was specifically highlighted to demonstrate the advantages of SSL in analyzing remote sensing data.

C. Uncertainty Estimation

Another significant query strategy in active learning involves the computation of uncertainty, which is crucial for efficiently selecting informative data points for training. In the Bayesian framework, uncertainty is typically represented by the posterior distribution $p(\theta|D)$, where θ denotes model parameters and D the observed data. However, the computation of this posterior distribution becomes computationally infeasible in deep learning due to the extremely high dimensionality of the parameter space. A practical and widely acknowledged solution to this challenge is Monte Carlo (MC) Dropout [4]. MC Dropout not only simplifies the estimation of the posterior but also provides a computationally efficient approximation by utilizing dropout at both training and inference stages. This technique enables the model to generate different outputs for the same input by randomly dropping units, which effectively samples from an approximate posterior distribution.

It was applied by the same authors in another work for active learning in image classification [5]. The results showed that uncertainty estimation helps reducing the labelling effort.

D. *Combined Methods*

Another study presented in [6] explores a hybrid approach for active learning in remote sensing. Initially, contrastive learning with negative sampling was employed using MoCo to process the data, followed by clustering of the encoded data. Subsequently, Euclidean-distance-based diversity sampling was applied within each cluster to select training samples for the main model. This one-shot procedure demonstrated superior performance over other competitive methods.

In our approach, termed MCFPS, we integrate the beneficial aspects of the methods described above. We utilize negative sample contrastive learning with MoCo to encode data into a lower-dimensional feature space. Subsequently, we explore the neighborhood of each sample to identify and select the most uncertain sample for training our main model.

III. Methodology

Our proposed method consists of a initialization and an iterative learning scheme:

A. *Initialization*

Given a dataset $X \in \mathbb{R}^{N\times d}$, we start by training an encoder f to transform the dataset to $X_{enc} \in \mathbb{R}^{N\times e}$, where $e \ll d$.

B. *Iterative Learning Scheme*

The iterative learning scheme consists of five steps:

1. Model initialization: With the encoded data X_{enc}, we initiate our iterative active learning scheme. This involves initializing a model g to estimate uncertainty using MC dropout.

2. Diversity Sampling: We employ farthest-point-sampling within the transformed space X_{enc} to select S objects from the dataset.

3. Nearest-Neighbour Search: For each sampled object S_i, a nearest-neighbour search is conducted with a neighbourhood size of k.

4. Uncertainty Estimation: Uncertainty for each S_i within its neighbourhood M_i is estimated through t forward passes using model g.

5. Candidate Selection: For classification tasks with C classes the previous step yields the uncertainty $\hat{Y} \in \mathbb{R}^{t\times C}$ for each neighbour in M_i. The overall uncertainty is determined by averaging the outcomes to yield $\hat{Y}_{Mean} \in \mathbb{R}^{C}$ and taking the maximum of $\hat{Y}_{Mean}$

The final step involves selecting the sample with the highest uncertainty in each neighbourhood for inclusion in the set of samples that need to get labelled by a human annotator next.

The labelled data is then used to pretrain the model g that is being used for uncertainty estimation for the next iteration.

The method is also depicted in Figure 1.

```
Algorithm 1 Iterative Active Learning Scheme
Initialization:
    Given dataset $X \in R^{N\times d}$
    Train encoder $f$ to transform $X$ to $X_{enc} \in R^{N\times e}$ where $e \ll d$
Iterative Learning Scheme:
while not converged do
    Step 1: Model Initialization
        Initialize model $g$ with $X_{enc}$ using MC dropout for uncertainty estimation
    Step 2: Diversity Sampling
        Select $S$ objects from $X_{enc}$ using farthest-point-sampling
    Step 3: Nearest-Neighbour Search
    for each sampled object $S_i$ do
        Conduct nearest-neighbour search with neighbourhood size $k$ to find $M_i$
    end for
    Step 4: Uncertainty Estimation
    for each $S_i$ in $S$ do
        for each neighbour in $M_i$ do
            Estimate uncertainty through $t$ forward passes using model $g$
            $\hat{Y} \in R^{t\times C}$ where $C$ is the number of classes
            Compute $\hat{Y}_{Mean} \in R^{C}$ by averaging the outcomes
        end for
        Determine overall uncertainty by taking max($\hat{Y}_{Mean}$)
    end for
    Step 5: Candidate Selection
    Select sample with highest uncertainty in each neighbourhood for labelling
    Use labelled data to pretrain model $g$ for next iteration
end while
```

Fig. 1. The pseudocode of our proposed method

IV. Experementation and Results

A. *Experimental Setup*

The code for our experiments can be found here: https://github.com/autocoast/active-learning-sentinel-s2

For our experiments we used the Eurosat dataset [7], a popular remote sensing dataset with 27000 Sentinel S2 image patches each of size 64x64x13. Every patch is assigned to one out of 10 possible classes. The dataset is almost perfectly class-balanced.

We utilized a pretrained ResNet50 model as our encoder sourced from [8] which was used as a backbone for a contrastive learning with MoCo. The encoder transforms each 64x64x13 patch into a 2048-dimensional vector. Each patch was encoded in this manner. Figure 1 illustrates the encoded data after an additional PCA with 50 components and a further t-SNE with two components were applied. The different classes are distinguished by varying colours, demonstrating the effective separability of the data. To quantify the impact of this method, we applied a k-Nearest-Neighbor (kNN) classifier to the t-SNE-reduced data, achieving a test accuracy of 95%. Conversely, direct encoding of the raw 64x64x13 data into 50 dimensions using PCA, followed by kNN classification with the same train/test split, resulted in a test accuracy of only 68%.

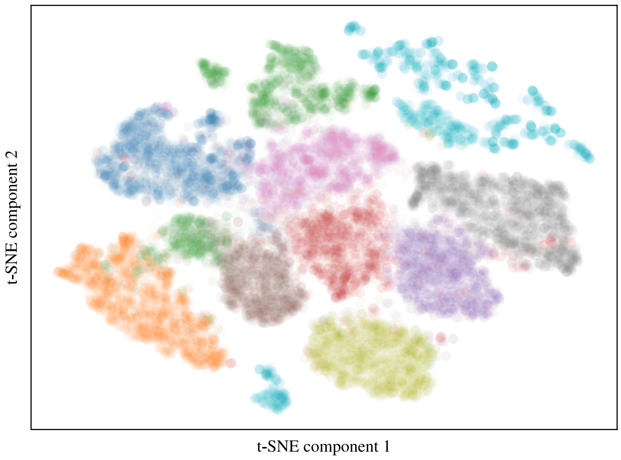

Fig. 2. T-SNE plot of the Eurosat dataset after performing the encoding with SSL and further PCA employment. Each color represents a label in the dataset.

We run our experiments for two scenarios. In the first scenario, we train our models on the original, class-balanced data set. In the second scenario, we modify the training data in a way that the training data becomes unbalanced.

For each of both settings we compare MCFPS against random selection. We run our method iteratively for eight iterations with selecting 64 candidates in each round. We also check the performance of training on the full training data set without a pretrained ResNet50 model.

B. Results

The results show that applying the pre-trained SSL model from [8] already yields to a significant boost in the test accuracy where 90% of test accuracy is already reached with less than 1% of the data in the balanced case and 1.8% in the unbalanced case. This is illustrated in Figure 3.

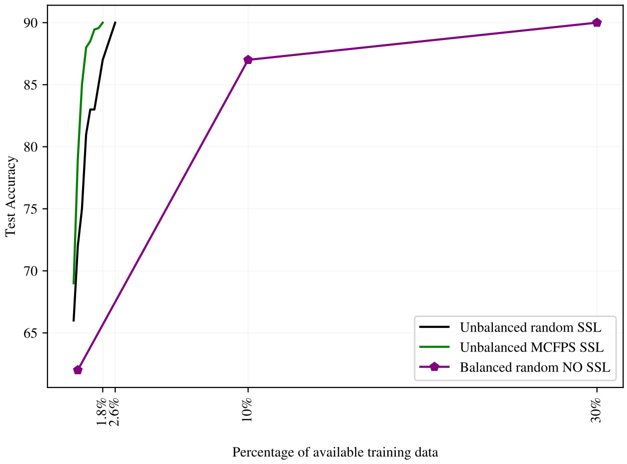

Fig. 3. Utilizing a contrastive learning based pre-trained model significantly boosts the performance of the model. This was already shown in [2].

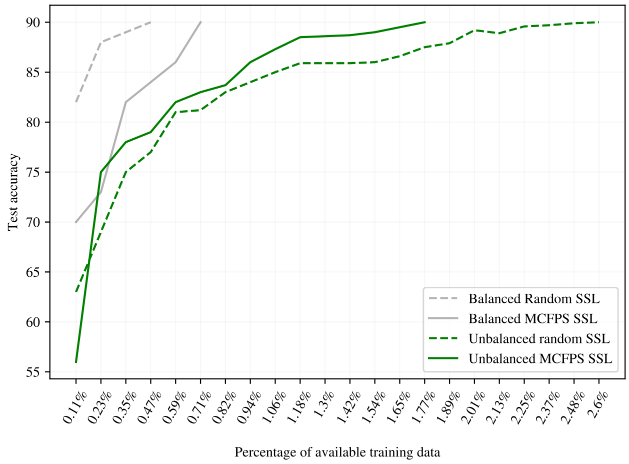

Fig. 4. Test accuracy results for the balanced and unbalanced scenarios. In a balanced setting, the random selection shows better performance than the MCFPS method. The MCFPS however outperforms the random selection method in an unbalanced setting.

For the balanced scenario, we see that performing random sampling yields to better results than MCFPS. When the data is unbalanced, MCFPS reaches the 90% mark quicker than random.

V. Discussion

The results indicate that using certainty and diversity-based active learning can significantly reduce labelling effort in unbalanced scenarios, which are typical in real-world applications. Specifically, employing a model pre-trained with MoCo substantially decreases the labelling effort compared to not using pre-trained models. However, evaluating this effect with only one dataset is insufficient to draw comprehensive conclusions. Additionally, the dataset used in our study was relatively easy to classify after encoding. In future work, we plan to incorporate a real-world dataset that not only focuses on classification but also on semantic segmentation, which is a more prevalent scenario in land usage and land coverage tasks.

Note on Preliminary Results: This paper introduces a novel method for active learning and presents results from an initial experiment comparing random selection with our proposed method. We acknowledge that these results are preliminary. Ongoing experiments are being conducted, and we will update this paper with additional findings, including comparisons with other methods, in the coming weeks.

References

[1] Settles, Burr. "Active learning literature survey." (2009).

[2] Wang, Yi, et al. "Self-supervised learning in remote sensing: A review." IEEE Geoscience and Remote Sensing Magazine 10.4 (2022): 213-247.

[3] He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[4] Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." international conference on machine learning. PMLR, 2016.

[5] Gal, Yarin, Riashat Islam, and Zoubin Ghahramani. "Deep bayesian active learning with image data." International conference on machine learning. PMLR, 2017.

[6] Jin, Qiuye, et al. "One-shot active learning for image segmentation via contrastive learning and diversity-based sampling." Knowledge-Based Systems 241 (2022): 108278.

[7] Helber, Patrick, et al. "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 12.7 (2019): 2217-2226.

[8] Wang, Yi, et al. "SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation [Software and Data Sets]." IEEE Geoscience and Remote Sensing Magazine 11.3 (2023): 98-106.