DOFA - Hot topics in Computer Vision

Anonymous HTCV submission

Paper ID *****

Abstract

Large language models have shown impressive performance gains for various tasks in language processing in the last few years [11]. It is reasonable to think that it is possible to overcome some of the problems in remote sensing image processing, like scarcity of labeled data, with the help of foundational vision models.

In this work, we provide an evaluation of the DOFA model [20] when used on the BigEarthNet dataset [4]. This includes an analysis of the feature vectors and different approaches for using these features for multi-label classification.

We've seen reasonable performances across different metrics but no improvement compared to other models trained on BigEarthNet [15].

1. Introduction

The success of computer vision deep learning models relies heavily on the large amount of data available for training on the internet [16]. In the domain of remote sensing though, labeled data is hard to obtain because labeling is a complicated task (requires expert knowledge) and there is a much larger range of sensors [20]. There are multiple ways to address this issue like synthetic data or semi-supervised learning algorithms. After the breakthroughs of LLMs and the concept of pre-training foundation models on a wide range of tasks, recent research tried to transfer this concept in the area of remote sensing computer vision [2].

This report focuses on one foundational model called "DOFA" [20] and evaluates it on a new downstream task for comparison with other foundation models. The used dataset is called BigEarthNet [15] and we will be using the Sentinel-1 data. We will provide results with different metrics on the multi-label classification for the 19-class variant and also the 43-class variant.

We will analyze the meaningfulness of the features computed by the DOFA model by applying a UMAP transformation [10] to understand the meaningfulness of the features. Afterward, we use the features as input data for training different classifiers on the classification task.

2. Related research

Self-supervised learning has been used for some years to accomplish different tasks in remote sensing and can also be used to train larger multi-modal foundational models [5, 6, 20]. In principle, there are two ways to generate meaningful features. The first is to use contrastive learning [3], where two views of the same image should be close in feature space and two views of different images far away from each other. These two views can be simple image transformations or, for example, the same area of the earth from another sensor. There are many examples with different strategies for this contrastive learning approach [1, 5**?**]. Research has shown that this approach tends to produce models that ignore information in the data that is not shared between the augmented views, regardless of whether this information would be useful in later downstream tasks [17]. Because of this, it is key to select an appropriate augmentation strategy as this decision heavily affects the performance of the model [12].

The alternative to contrastive learning is learning to reconstruct images. An example of this approach is SimMIM [19]. It is possible to easily scale these models as they don't rely on image pairs [8], but they might need more fine-tuning to become useful for downstream tasks [9].

Most earlier foundational models were trained on a single sensor type. For example, Scale-MAE [13] is for optical data, SatMAE [5] for Sentinel-2 data, and so on. Even if these models can be used for various downstream tasks in their specific sensor domain, they can't generalize across these domains. According to DOFA, this results in multiple limitations. Models can't use most of the unlabeled data during training because it is from a different sensor type. They also lack universatility when the downstream task differs from the original data as needed channels and bandwidths change based on the used sensor. Overcoming these limitations and

exploiting the usage of data from various sensors is the main goal of DOFA [20].

There is different research that tries to solve this issue of multi-modality like [6, 7, 18].

3. DOFA Model

"DOFA" stands for *"Dynamic-One-For-All"* model, inspired by the concept of neural plasticity from brain science. The main goal is to integrate various data modalities into a single framework, adaptively adjusting to different sensor inputs without needing separate models for each sensor type. This means that the model adjusts different weights based on different wavelengths of the input data. The training is similar to training one large transformer model.

The authors used data from *different sources and sensors* including Landsat, Sentinels, MODIS, EnMAP, Gaofen, and NAIP, which are representative of a large range of spectral bands, resolutions, and imaging types. DOFA is trained using masked images and predicting the missing patches to learn meaningful representations on unlabeled data. It uses *pre-trained models* that were trained on ImageNet to enhance training efficiency and reduce training times, as well as a new distillation loss that should improve the model performance.

The *distillation loss* is derived from the concept of knowledge distillation, where we want to transfer knowledge from a larger model (e.g., trained on ImageNet), which we call the "teacher" model to a smaller model, which we call the "student" model. The basic idea is that the student model (DOFA) is trained so that its output aligns closer to that of the teacher model. This is especially useful when it comes to data that the teacher model has already seen, which in this case should be RGB images (especially for example on landscapes). This setup should accelerate training convergence and enhance the overall performance. This loss is combined with the loss for image reconstruction. Together they guide the model to not only reproduce the input data correctly but also form representations that are informed by the teacher model.

The core idea of adjusting weights is through a second *hypernetwork*. This secondary neural network is trained to generate weights and biases for the main network (transformer-based) to execute, based on the input data. The used characteristics include central wavelengths associated with each input band and their standard deviations. The hypernetwork learns during training to generate effective weights as it receives feedback on the performance of the primary network. This approach makes it possible to use tailored transformations for different modalities *and* train everything networks in one go.

The authors claim that this approach using one foundational model *reduces the computational overhead* and complexity compared to multiple specialized models. Through its multimodal training, the model should be able to handle EO tasks and data types that it hasn't seen before. This concept of multi-modality is supposed to be transferable to other fields like medical image analysis and robotics.

The results in their paper show high performances in different datasets and downstream tasks compared to other approaches. The only dataset in their evaluation that is behind other approaches is the BigEarthNet for which we will provide more results.

[20]

4. Methodology

The images of the BigEarthNet are being fed through the DOFA model and the feature vectors are used for further analysis like UMAP visualization and training data for downstream tasks. The approach relies on PyTorch (DOFA execution, classifier training), scikit-learn (classifier training, metrics), and a Python implementation of the UMAP visualization.

5. Experiments

5.1. Data

We're using the *BigEarthNet dataset* [15] for our evaluations. The goal is to use the model as a feature generator and use these features for solving the multi-label classification problem with 19 classes and also 43 classes from BigEarthNet. This dataset contains around 550,000 image patches from Sentinel-I and Sentinel-II with multi-class labels for every image. We are going to use the *Sentinel-I* images for our experiments.

The data is split into train and test set based on the implementation from the torchgeo repository. The train dataset contains around 270,000 and the test dataset has 125,000 images.

There are high differences in the occurrence of certain classes in the dataset (see Fig. 1). There are, for example, only a few images with wetlands, beaches, and moors. The metrics during the evaluation have to account for this.

We analyzed the correlation between classes of BigEarthNet and the results are displayed in Fig. 2. The classes are mostly uncorrelated with some exceptions like the high likelihood that coniferous forests are usually also mixed forests and marine waters are mostly not in the same images as all the other vegetation forms.

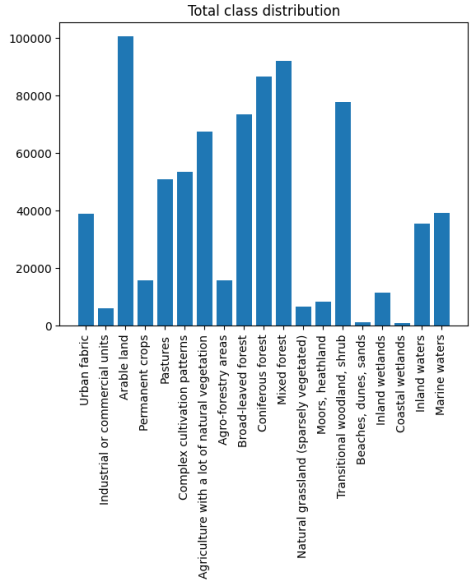

Figure 1. Distribution of the classes in the train dataset. This is a multiclass dataset so one can have multiple classes.

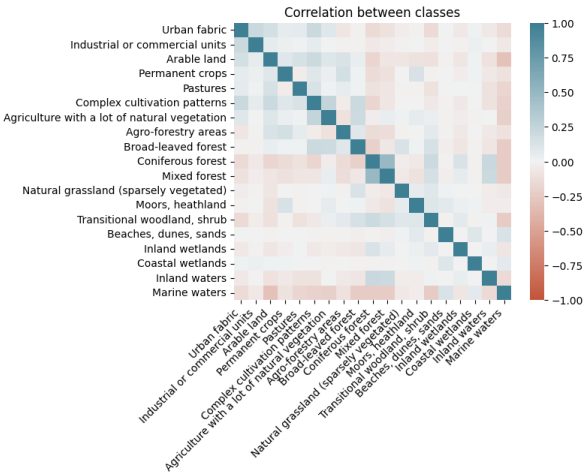

Figure 2. Correlation of classes in the train dataset

5.2. Performance Metrics

We decided to compute multiple metrics for evaluating the performance of the models in the classification tasks. We provide results for *f2-micro, f2-macro* (Implementation), *hamming-loss*, and *precision* scores (Implementation). We also calculate precision and f1 scores for individual classes.

5.3. Results

First, we calculated feature vectors using DOFA for all the images in BigEarthNet (Sentinel-I) and saved them for further use. All the following analyses use these vectors for visualizations or the downstream classification task.

5.3.1 UMAP Feature analysis

In Fig. 3 are the occurrences of different classes in a UMAP transformed 2D space visualized. Because one image has multiple labels, we chose the visualization via multiple charts. In this visualization, we expect different classes to be located in different data cloud regions. As we can see, this assumption holds for a lot of the classes. Most notably, we have a separate region for all marine water images (right bottom chart). As we've seen in Fig. 2, this class is negatively correlated with most of the other classes, so a separation in this visualization is plausible. It appears that arable land is more on the left side of the largest region, while forests, woodlands, etc. are more on the right side.

With the results of this visualization, we expect that the feature vectors of DOFA capture the semantic meaning of the different images, and it should be possible to train a classifier on top of them.

5.3.2 Classification

We used the feature vectors of DOFA to train different ML algorithms for the given classification task. All the approaches use the same data split. In Table 1 are the final results of these experiments.

There are notable differences in the performance of the chosen approaches. MLP has the best performance over most of the metrics. However, the random forest classifier and linear probing also have similar scores in P_{micro} and P_{macro}. All classifiers have significant differences between their micro and macro averaged F^2 scores. When we compare the F^2 scores per class, we can see the cause for these values (Fig. 4): There is a high variance in the performance across classes, which ranges from 0% F^2 score to about 90% F^2 score. This leads to the different averaging variants resulting in different scores. We can also see that MLP manages to achieve higher scores than the linear classifier over more classes. Additionally, the low performing classes are mostly similar and correlate to the classes with low-data availability. It seems that the labels were not enough to learn all of the classes well.

We can conclude that the features generated by the DOFA model are not well linearly separable across all classes and only some achieve an F^2 score of over 50%. The MLP model can separate more complex data and therefore reaches a better performance.

Random forest classifiers can separate high dimensional data, but in this case, as seen in Table 1, their scores are on the same level as the linear classifier.

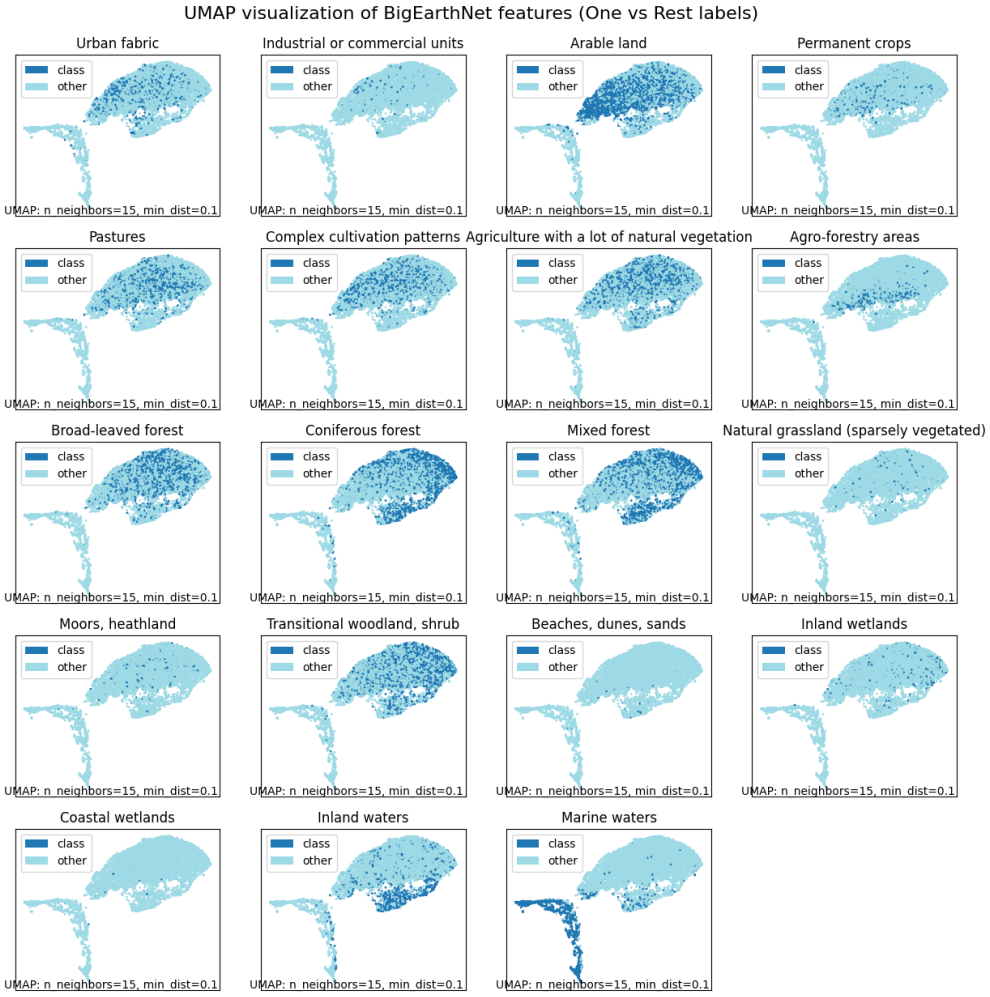

Figure 3. Visualization of the 19 classes in a UMAP transformed space (one vs. rest)

	F^2_{macro} (%)	F^2_{micro} (%)	**hamming loss**	P_{macro} (%)	P_{micro} (%)
Random Forest	21.4	35.8	0.123	52	**72**
Linear Probing	22.9	35.3	0.124	49	70
MLP	**34.5**	**47.8**	**0.113**	**58**	71

Table 1. Test results of different classification approaches on the 19-class multi-label classification task. Some classes contain so few examples that they are not in the test dataset and receive an F^2 score of 0.

In Table 2 we provide test results for the multi-label classification task with 43 classes of BigEarthNet [15]. The results are expected worse than on the 19-class variant, as this is a more complex task. We can see that the DOFA model isn't able to keep up with a specialized model, e.g. [14] even though this comparison has its limitations as this paper used the Sentinel-II images.

In Fig. 5, we see the confusion matrices of the best model over the 19 classes. With 19 classes, it is sensible that most of the classes are quite uncommon, which results in a lot of true negatives. It is interesting to see that most of the errors come from false negatives, which means that the model is unable to detect certain classes in an image.

6. Conclusion

As we have seen, the DOFA model can be used to analyze the Sentinel-I from the BigEarthNet dataset. The features from the model contain the semantic information necessary to train a classifier to predict the related class labels. However, the model struggles to provide good features for the low-availability

	F^2_{macro} (%)	F^2_{micro} (%)	hamming loss	P_{macro} (%)	P_{micro} (%)
Random Forest	9.46	33.73	0.056	24	**71**
Linear Probing	12.63	36.54	0.057	28	66
MLP	15.45	44.46	0.052	**38**	**71**
COMP [14]	**52.8**	**62.3**	**0.04**	/	/

Table 2. Test results of different classification approaches on the 43-class multi-label classification task. For comparison, we provide the results of a specialized model for this task [14]. This comparison is only meant as a reference as this paper uses the Sentinel-II images of BigEarthNet, which includes more spectral bands.

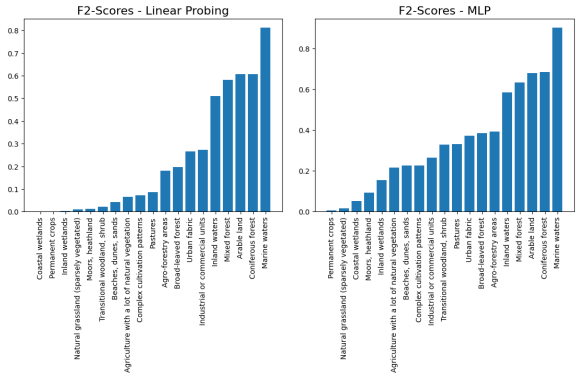

Figure 4. F^2 scores for every class (19 in total). Left side for the linear classifier, right side for the MLP classifier.

Figure 5. Confusion matrices of the MLP classifier on the test set with 19 classes. The matrices are calculated separately for each class.

classes, which leads to an overall mediocre classification performance and doesn't keep up to the state-of-the-art.

References

[1] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning, 2022. 1

[2] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. 1

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 1

[4] Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reben: Refined bigearthnet dataset for remote sensing image

analysis, 2024. 1

[5] 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. 1

[6] Anthony Fuller, Koreen Millard, and James R. Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders, 2023. 1, 2

[7] Jakob Hackstein, Gencer Sumbul, Kai Norman Clasen, and Begüm Demir. Exploring masked autoencoders for sensor-agnostic image retrieval in remote sensing, 2024. 2

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 1

[9] Johannes Lehner, Benedikt Alkin, Andreas Fürst, Elisabeth Rumetshofer, Lukas Miklautz, and Sepp Hochreiter. Contrastive tuning: A little help to make masked autoencoders forget, 2023. 1

[10] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. 1

[11] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2024. 1

[12] Maxim Neumann, Andre Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing, 2019. 1

[13] Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning, 2023. 1

[14] Gencer Sumbul and Begüm DemİR. A deep multi-attention driven approach for multi-label remote sensing image classification. *IEEE Access*, 8:95934–95946, 2020. 4, 5

[15] Gencer Sumbul, Marcela Charfuelan, Begum Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019. 1, 2, 4

[16] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017. 1

[17] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning?, 2020. 1

[18] Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries, 2024. 2

[19] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling, 2022. 1

[20] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J. Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation, 2024. 1, 2