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. 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 when used on the BigEarthNet dataset. 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.

1. Introduction

The success of computer vision deep learning models relies heavily on the large amount of data available for training on the internet. [**?**] 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. There are multiple ways to address this issue like synthetic data or semi-supervised learning algorithms. After the breakthroughs of LLM and the concept of pretraining foundation models on a wide range of tasks, recent research tried to transfer this concept in the area of remote sensing computer vision.

This report focuses on one foundation model called "DOFA" and evaluates it on a new downstream task for comparison with other foundation models. The used dataset is called BigEarthNet 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. to understand the meaningfulness of the features. Afterward, we use the features as input data for training different classifiers on the classification task.

// TODO: example images

2. Related research

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 and 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 represent for 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 and 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 net-

work (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* training the 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.

// source DOFA paper

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 //https://umap-learn.readthedocs.io/en/latest/index.html

5. Experiments

5.1. Data

We're using the *BigEarthNet dataset* 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 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 **??**). 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 **??**. The classes are mostly uncorrelated with some exceptions like the high likeliness that coniferous forests are usually also mixed forests and marine waters are mostly not in the same images as all the other vegetation forms.

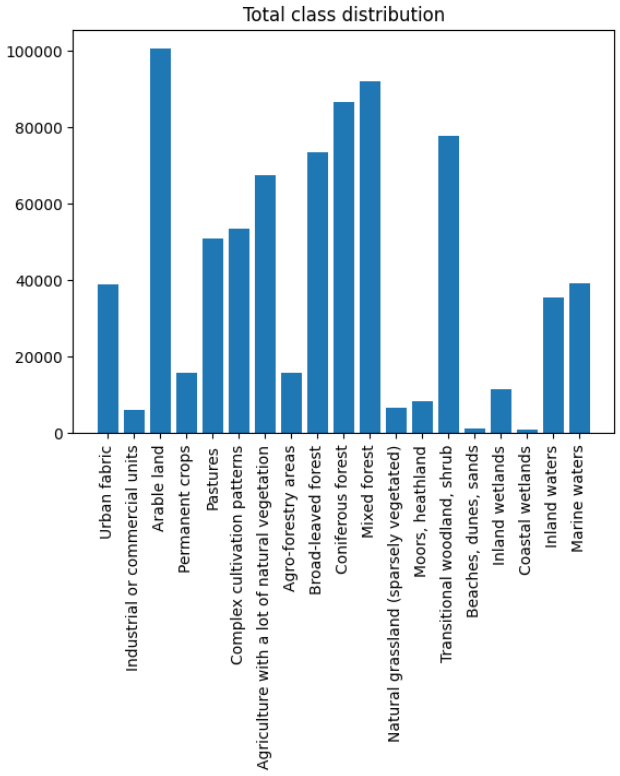

Figure 1. Distribution of the classes in the train dataset. This is a multiclass dataset so one can have multiple classes.

Figure 2. Correlation of classes in the train dataset

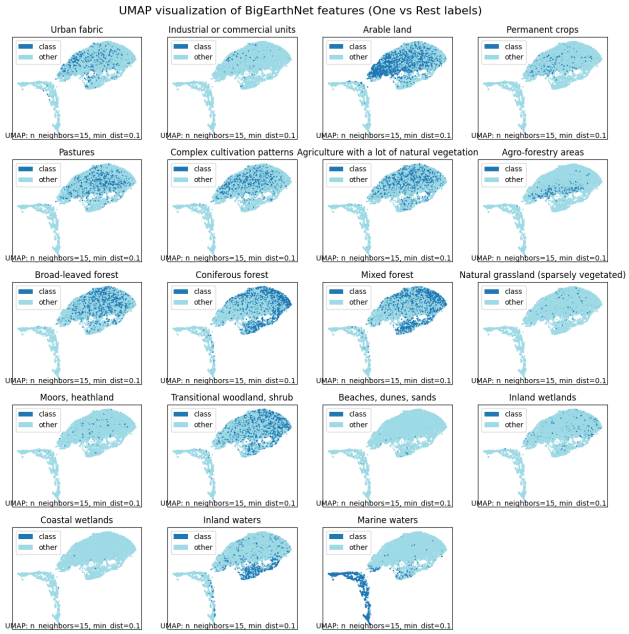

Figure 3. Visualization of the 19 classes in a UMAP transformed space (one vs. rest)

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* (see XXXX for implementation) *, hamming-loss*, and *precision * scores (see XXXX for 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 **??** are the occurrences of different classes in a UMAP transformed 2D space visualized. Because one image has multiple labels we choose the visualization via multiple charts. In this visualization, we expect different classes to be in different data cloud regions. As we can see this 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 @correlation 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 main region while forests, woodlands, etc. are more on the right side.

sdsd	F^2_{macro} %	F^2_{micro} %	hamming loss	P_{macro} %	P_{micro} %
Test	1	6	87837	787	787
Test	2	7	78	5415	787
Test	3	545	778	7507	787
Test	4	545	18744	7560	787
Test	5	88	788	6344	787

Figure 4. Test results of different classification approaches on the 19-class multi-label classification task [1]

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 that data.

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 @test-results are the final results of these experiments.

figure(table(columns: (auto, auto, auto, auto, auto, auto), inset: 3pt, align: horizon, table.header([], [$F^2 macro$" (), "Random Forest", 21.4, 35.8, 0.123, 52, $bold(72)$, "Linear Probing", 22.9, 35.3, 0.124, 49, 70, "MLP", $bold(34.5)$, $bold(47.8)$, $bold(0.113)$, $bold(58)$, 71,), caption: "Test results of different classification approaches on the 19-class multi-label classification task" + footnote("Some classes contain so few examples that they are not in the test dataset and receive a " + F^2 + " score of 0.")) ¡test-results¿

figure(table(columns: (auto, auto, auto, auto, auto, auto), inset: 3pt, align: horizon, table.header([], [$F^2 macro$" (), "Random Forest", 9.46, 33.7, 0.056, 24, 71, "Linear Probing", // todo the rest of the lines 22.9, 35.3, 0.124, 49, 70, "MLP", $bold(34.5)$, $bold(47.8)$, $bold(0.113)$, $bold(58)$, 71,), caption: "Test results of different classification approaches on the 43-class multi-label classification task") ¡test-results-43¿

There are notable differences in the performance of the chosen approaches. MLP has the best performance over most of the metrics. However 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 scores. When we compare the F^2 scores per class we can see the cause for

these values (@scores-by-class). There is a high variance in the performance across classes which reaches from 0

figure(grid(columns: (auto, auto), rows: (auto), gutter: 3pt, image("images/Linear Probing - f2 scores.png"), image("images/MLP - f2 scores.png"),), caption: F^2 + " scores for every class. Left side for the linear classifier, right side for the MLP classifier.") ¡scores-by-class¿

This leads to the conclusion that the features generated by the DOFA model are not well linearly separable across all classes and only some achieve a F^2 score of over 50

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

figure(image("images/MLP - confusion matrix.png"), caption: "Confusion matrices of the MLP classifier on the test set. The matrices are calculated separately for each class.")

// REMARKS ABOUT CONFUSION MATRIX

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 classes, which leads to a mediocre classification performance.

References

[] FirstName LastName. The frobnicatable foo filter, 2014. Face and Gesture submission ID 324. Supplied as supplemental material `fg324.pdf`.