MULTI-MODAL SELF-SUPERVISED LEARNING FOR BOOSTING CROP CLASSIFICATION USING SENTINEL2 AND PLANETSCOPE

Ankit Patnala, Scarlet Stadtler & Martin G. Schultz [*]

Juelich Supercomputing Centre
Forschungszentrum Juelich
Juelich

Juergen Gall[†]

Department of Information Systems Artificial Intelligence
University of Bonn
Bonn

ABSTRACT

Remote sensing has enabled large-scale crop classification to understand agricultural ecosystems and estimate production yields. Since few years, machine learning is increasingly used for automated crop classification. However, most approaches apply novel algorithms to custom datasets containing information of few crop fields covering a small region and this often leads to poor models that lack generalization capability. Therefore in this work, inspired from the self-supervised learning approaches, we devised and compared different approaches for contrastive self-supervised learning using Sentinel2 and Planetscope data for crop classification. In addition, based on the dataset DENETHOR, we assembled our own dataset for the experiments.

***Index Terms*—** Optical remote sensing, crop classification, contrastive learning, multi-modal contrastive learning, time-series, self-supervised learning

1. INTRODUCTION

Remote sensing has accumulated vast amount of data with improved capability of new satellite missions such as Sentinel2 [1] and Landsat[1]. These missions cover entire globe at a regular time interval making them valuable resources for crop classification, which heavily relies on temporal patterns. Along with the advent of machine learning to solve complex problems, applications such as crop classification are widely being automated. Though machine learning facilitates large scale crop mapping but labeling is time consuming and requires skilled human efforts. The need for generalize models without the need of additional manual annotations and the development of advanced algorithms in the field of deep learning has motivated the development of techniques such as self-supervised learning. Self-supervised learning relies on pretext tasks, and with recent advancement, contrastive learning [2] has shown promising results. Contrastive learning relies on contrastive losses such as InfoNCE [3] and augmentation. The contrastive learning method relies on augmentation of a data sample and aims to maximize similarity of the data sample and its augmented version but such augmentation for raw satellite data is non-trivial. For tabular data i.e. pixels in our case, the task of obtaining augmented data is not trivial, thus we relied on multi-modal contrastive learning where the augmented version is obtained from different sources; Sentinel2 and Planetscope. The end-user does not necessary require both sources to apply the model to end application and still can avail implicitly the benefits of both sources. In this work, we devised two different types of alignment; point-wise and time-wise for developing a pre-trained model for crop classification. We used DENETHOR [4] dataset to assemble data for our experiments as it provides multiple sources of data for the same geographical region. To our knowledge, there exist no work using contrastive learning on tabular data in the field of remote sensing. We adopted SCARF algorithm [5] for uni-modal self-supervised learning as a baseline.

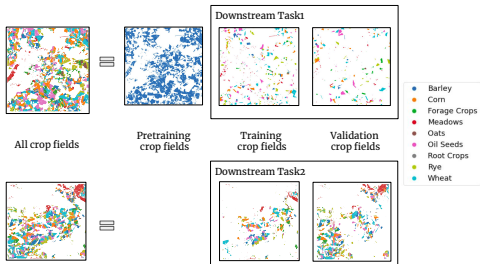

Fig. 1: **Dataset for multi-modal self-supervised learning experiment setup.** None of the fields are overlapping.

2. DATASET

Figure 1 gives a visual description of our strategy in assembling DENETHOR for our experiments. For our experiments, we need a pre-training dataset and different downstream datasets to conduct self-supervised experiments.

[*]Thanks to German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety for funding under grant no 67KI2043 (KISTE).

[†]The author is also associated with Lamarr Institute for Machine Learning and Artificial Intelligence, Germany. The author is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB 1502/1-2022 - Projektnummer : 450058266

[1]https://landsat.gsfc.nasa.gov/appendix/references/

The pre-training dataset is used to pre-train a model using a self-supervised approach and downstream datasets are used to evaulated the pre-trained model. We used both DENETHOR's training and validation set to assemble the dataset for our experiments. We did 70-21-9 random split of cropfields to the DENETHOR's training dataset. The 70% is used to obtain pre-training data whereas 21-9 split is used to obtain training and validation dataset for downstream task1. We did 70-30 random split on DENETHOR's validation dataset to obtain training and validation dataset for downstream task2. DENETHOR's training and validation dataset are from different regions as well as different timespans.

2.1. Point-wise pre-training dataset

For the point-wise pre-training dataset, we randomly selected $100,000$ pixels from each of the 144 timestamps available for Sentinel2. To obtain corresponding pairs from Planetscope, we used the same region and the same timestamp. It is to be noted that Sentinel2 has a pixel resolution of $10m/px$ whereas Planetscope has a resolution of $3m/px$, so we aligned a pixel of Sentinel2 to 3×3 pixels of Planetscope. In total, we used $14,400,000$ Sentinel2 pixels for this experimental setup. The top part of Figure 2 shows the data alignment of point-wise self-supervised learning.

2.2. Time-wise pre-training dataset

Contrast to point-wise which focuses on aligning pixels, in the time-wise setup, we aligned the time series of a Sentinel2 pixel to the corresponding 3×3 Planetscope pixels' time series. Sentinel2 has a temporal revisit time of $5-6$ days whereas Planetscope has a daily visit. We aligned Sentinel2 with relatively coarser temporal resolution to Planetscope to avail benefits of its finer temporal resolution. We randomly selected time series of $150,000$ pixels. The bottom part of Figure 2 shows the data alignment of time-wise self-supervised learning.

2.3. Dataset for downstream tasks

Downstream tasks are used for evaluating the pre-trained model. As shown in Figure 1, we created two downstream tasks i.e. with DENETHOR's training and validation set. We randomly selected 5000 pixels and 1000 pixels for 9 crop types separately for both of them. This yielded two balanced datasets with 45000 training data and 9000 validation data.

3. METHODS

3.1. Model architectures

We used three different categories of networks i.e. Bi-directional LSTM [6](recurrent network), inception time [7] (convolutional network) and position encoded transformer [8]

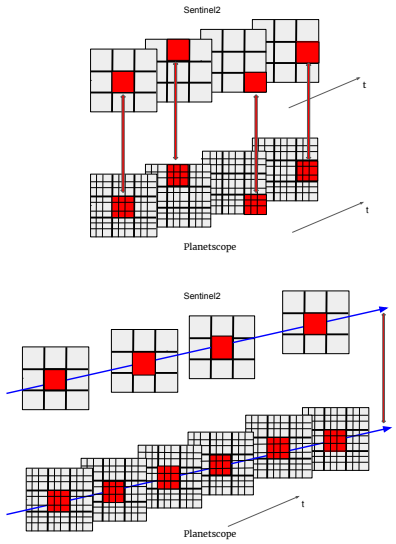

Fig. 2: **Data alignment for point-wise(top) and time-wise(bottom) self-supervised learning methods.**

(transformer network). We obtained 10 variant from each category by varying different hyperparameters such as number of layers, number of hidden dimension etc. using optuna [9].

3.2. Multi-modal self-supervised methods

As shown in Figure 2, in the point-wise method pixels are aligned in both sources whereas in time-wise method time series are aligned. Different architectures were used for self-supervised model based on their feasibility. Inspired from resnets, for point-wise methods we used skipped connection MLP and named it ResMLP. For time-wise self-supervised methods, we used DeiT [10] inspired transformers where we replaced initial 2D convolution layers with 1D convolution to adapt it to time series. For both types, we adapted the original SimCLR loss function to multiple modes.

3.3. Uni-modal self-supervised experiments

For a baseline uni-modal contrastive learning on tabular data, we used random feature corruption from SCARF [5] is used to obtain augmented version of the data. As we used single source here, so we used the original SimCLR [11] loss function. Both the ResMLP and transformer model are kept same as the ones corresponding to Sentinel2 in our multi-modal setup. In the point-wise self-supervised, representation of each pixel is obtained separately whereas in the case of time-wise, the pre-trained model processes a time series input and returns an embedded time series as its representation.

4. EXPERIMENTS AND RESULTS

We randomly obtained 10 different models for each category (LSTM, inception and transformers) and trained them on raw

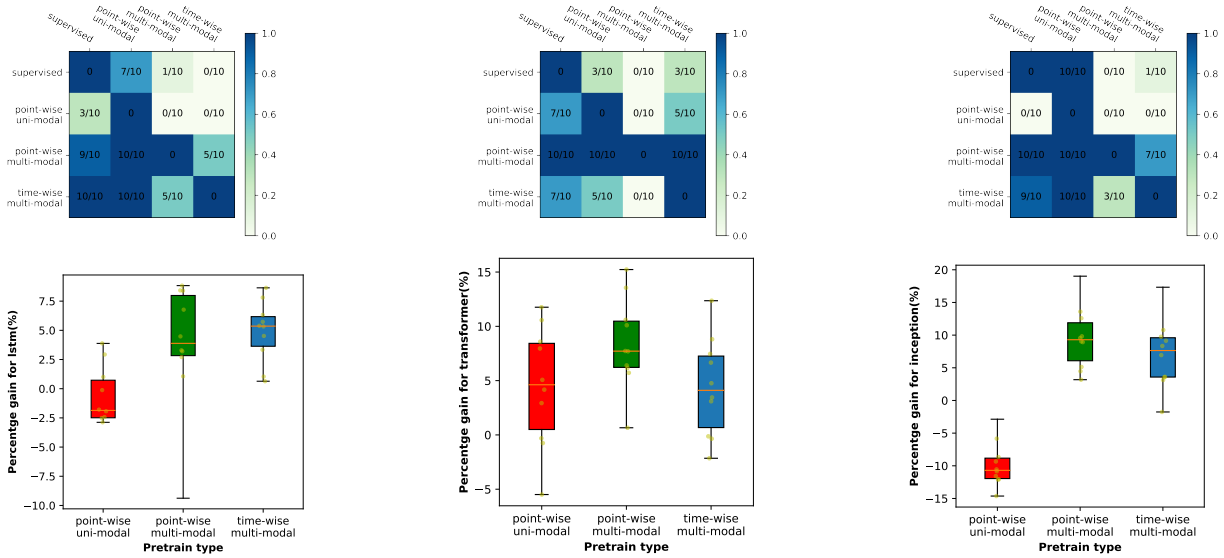

Fig. 3: **Win-matrix and relative gain box-plot of different self-supervised methods on downstream task1.** From left : LSTM, transformer, and inception

refelction values using standard supervised methods as the reference. For the point-wise, we used a 8 layer ResMLP model with both hidden dimension and output dimension as 256. For time-wise, we used encoder part of original transformer with 4 layers, 1D convolution with 256 kernels with an output dimension of 128. After pre-training, the representation are passed through 10 networks with the same hyperparameters as the supervised models. To evaluate the performance, we used win-matrices and box-plots to show the relative gain over the corresponding supervised baseline experiment.

Figures 3 and 4 show our multi-modal self-supervised approach outperforms the uni-modal self-supervised experiments. It is clearly evident that self-supervised learning using multiple complimentary sources learns an expressive representation of crops. The SCARF algorithm showed promising result in OPENML-CC18 [12] benchmark dataset but in the case of our remote sensing dataset, we did not find good results. The time-wise method performed well on downstream task1 (relative box-plots from Figure 3) but failed on downstream task2 (relative box-plots from Figure 4). Transformers being larger model with more hyperparameters are prone to learn noise in the case of noisy data. So, it did not generalize well to data from other region and time.

Despite using fine temporal information in the time-wise self-supervised learning, we did not find improvement in the scores. The reason could be the batch size. SimCLR loss function is highly dependent on the batch size due to the use of contrastive type loss. Transformers when compared to ResMLP are larger in size, hence consumes more memory. Thus to fit in the limited memory, the batch size of our time-wise experiment setup is restricted to 256 whereas for ResMLP, we could able to fit 1024 data in a batch.

The fine temporal resolution definitely gives more dense information to the model. The future prospects wiil be on how to effectively use the fine temporal resolution to improve the time-wise self-supervised model. We proposed an idea to adapt multiple source to the original BERT [13]. With BERT kind of setup, it will be easier to facilitate large batch size in a time-wise self-supervised setting. In addition, we want to add an auxiliary task exploiting data's seasonality.

5. REFERENCES

[1] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini, "Sentinel-2: Esa's optical high-resolution mission for gmes operational services," *Remote Sensing of Environment*, vol. 120, pp. 25–36, 2012.

[2] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.

[3] Aäron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[4] Lukas Kondmann, Aysim Toker, Marc Rußwurm, Andrés Camero, Devis Peressuti, Grega Milcinski, Pierre-Philippe Mathieu, Nicolas Longépé, Timothy

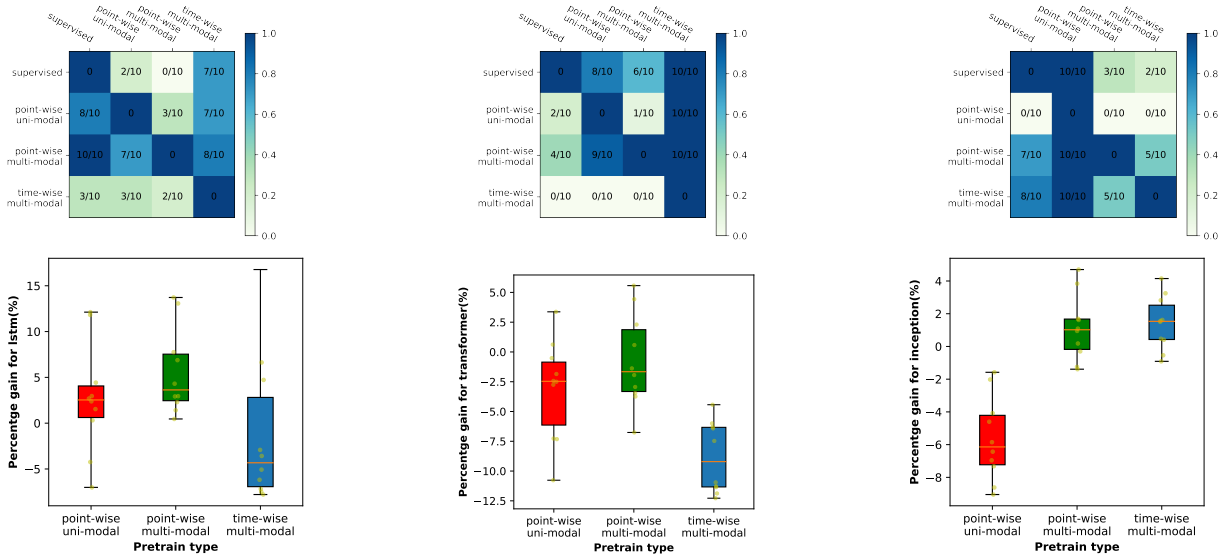

Fig. 4: **Win-matrix and relative gain box-plot of different self-supervised methods on downstream task2.** From left : LSTM, transformer, and inception

Davis, Giovanni Marchisio, Laura Leal-Taixé, and Xiao Xiang Zhu, "DENETHOR: The dynamicearthNET dataset for harmonized, inter-operable, analysis-ready, daily crop monitoring from space," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[5] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler, "SCARF: self-supervised contrastive learning using random feature corruption," *CoRR*, vol. abs/2106.15147, 2021.

[6] Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana, "Modelling radiological language with bidirectional long short-term memory networks," in *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, Auxtin, TX, Nov. 2016, pp. 17–27.

[7] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean, "InceptionTime: Finding AlexNet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, sep 2020.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, vol. 30.

[9] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, New York, NY, USA, 2019, KDD '19, p. 2623–2631.

[10] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou, "Training data-efficient image transformers & distillation through attention," *CoRR*, vol. abs/2012.12877, 2020.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton, "A simple framework for contrastive learning of visual representations," *CoRR*, vol. abs/2002.05709, 2020.

[12] Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael G. Mantovani, Jan N. van Rijn, and Joaquin Vanschoren, "Openml benchmarking suites," 2021.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.