2021-07-16

MULTI-MODAL SELF-SUPERVISED REPRESENTATION LEARNING FOR EARTH OBSERVATION

Pallavi Jain
Technological University Dublin, pallavi.jain@tudublin.ie

Bianca Schoen Phelan

Robert Ross

Recommended Citation

P. Jain, B. Schoen-Phelan and R. Ross, "Multi-Modal Self-Supervised Representation Learning for Earth Observation," 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2021, pp. 3241-3244, doi: 10.1109/IGARSS47720.2021.9553741.

MULTI-MODAL SELF-SUPERVISED REPRESENTATION LEARNING FOR EARTH OBSERVATION

Pallavi Jain, Bianca Schoen-Phelan, Robert Ross
School of Computer Science
Technological University Dublin
Dublin, Ireland
{pallavi.jain, bianca.schoenphelan, robert.ross}@tudublin.ie

IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9553741

***Abstract*—Self-Supervised learning (SSL) has reduced the performance gap between supervised and unsupervised learning, due to its ability to learn invariant representations. This is a boon to the domains like Earth Observation (EO), where labelled data availability is scarce but unlabelled data is freely available. While Transfer Learning from generic RGB pre-trained models is still common-place in EO, we argue that, it is essential to have good EO domain specific pre-trained model in order to use with downstream tasks with limited labelled data. Hence, we explored the applicability of SSL with multi-modal satellite imagery for downstream tasks. For this we utilised the state-of-art SSL architectures i.e. BYOL and SimSiam to train on EO data. Also to obtain better invariant representations, we considered multi-spectral (MS) images and synthetic aperture radar (SAR) images as separate augmented views of an image to maximise their similarity. Our work shows that by learning single channel representations through non-contrastive learning, our approach can outperform ImageNet pre-trained models significantly on a scene classification task. We further explored the usefulness of a momentum encoder by comparing the two architectures i.e. BYOL and SimSiam but did not identify a significant improvement in performance between the models.**

***Index Terms*—self-supervised learning, unsupervised learning, satellite images**

I. Introduction

Satellite imagery has been popular for land cover mapping, crop assessment, disaster monitoring etc. but is largely lacking in labelled data availability. In cases of a lack of labelled data, transfer learning has been a successful strategy in most computer vision and satellite imagery tasks. But most transfer learning is based on non-EO RGB images, which are different from remote sensing images in terms of spectral and semantic information. Also, unlike non-EO images satellite images are often larger in terms of spectral channels and range of modalities such as including SAR data. This discrepancy raises the goal of having EO (satellite imagery) specific pre-trained models, but yet achieving this when labelled data is scarce in satellite imagery.

To address the issue of labelled data availability, the major focus of machine learning has recently shifted towards learning representations based on self-supervised learning (SSL). With many proposed state-of-art methods, SSL has now reduced the performance gap with supervised learning in computer vision [1], [2]. These self-supervised representation learning methods follow the phenomenon of pre-text tasks and contrastive learning to learn spatially and semantically invariant representations. Pre-text tasks are based on the providing task-specific augmented images such as rotated, patches, or distorted which forces the model to predict the underline task [3], [4], [5]. In contrastive learning approach, an image is transformed into two augmented views i.e. positive pairs and negative pairs; the model tries to bring the two positive pairs (same image views) closer while repulsing the negative pairs (non similar images) [6], [7]. With the success of contrastive learning, recently many non-contrastive learning architectures such as BYOL [1] and SimSiam [2] have been proposed. These model architectures eliminated the need for negative pairs and utilised only the positive pairs of the images to learn the stable invariant representations while also outperforming supervised learning. However, we have seen limited exploration of this approach in satellite imagery, which motivated our work to apply SSL in EO domain over two modalities of satellite images, i.e. Sentinel-1 and Sentinel-2.

In this work we proposed the a cross modal SSL model by exploiting both random multi-spectral (MS) and SAR images as two augmented views of an image. While both sources vary in spatio-spectral information, learning the invariant satellite imagery representations from them can be beneficial for generalising across the EO domain.

For this work, we utilised random single band images from MS and SAR data to learn the satellite image representations by using two state-of-art model architectures i.e. BYOL [1] and SimSiam [2]. Where BYOL utilises the momentum encoder as target network, SimSiam claimed that having shared weight encoder with stop gradient can perform as well as momentum encoder. With this respect we analysed the performance of these two architectures and compared performance of both pre-trained models, along with a more traditional though limited ImageNet trained model.

II. Related Work

Different sensors of EO provides different spatio-spectral properties, but each sensor has its own shortcoming. As MS provides good spatio-spectral data but often suffers from cloud coverage whereas SAR parse the clouds but are difficult to analyse. To overcome such challenges many fusion techniques have been proposed and with the advancement in deep learning many works have utilised the same for various EO domain tasks. Among them some initial popular work has been done

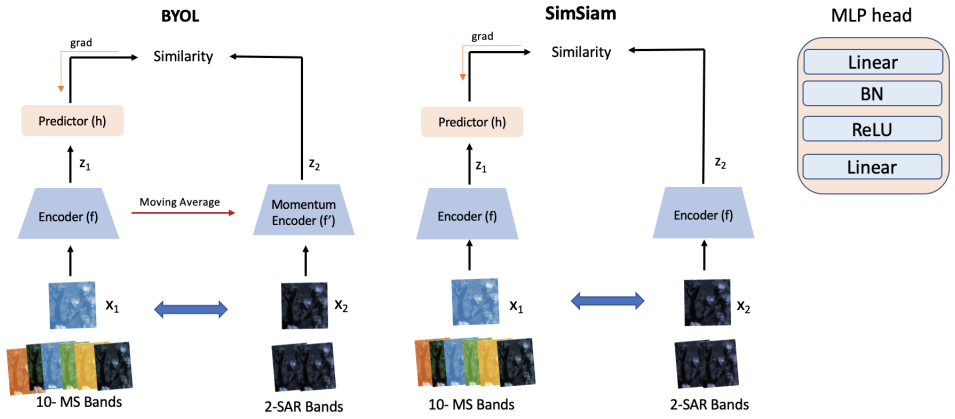

Fig. 1: Model architecture consist of Encoder (f), MLP projection head (z_i), and MLP prediction head (h). **Left side**- BYOL with momentum encoder in target network whereas, **Right side**- SimSiam with same encoder on both network

by Kussul et al. [8] and Ienco et al. [9] for scene classification and land cover mapping respectively with CNN architectures to learn representations for multi-modal data. Also some work utilised SAR images to generate missing part in optical images due to cloud coverage with the use of GANs [10]. Over the time it has been seen that SAR and MS data are interdependent and together can provide better representations for EO domain.

In order to learn better representations, its important to learn invariant representations to generalise it across the domain. With that Self-Supervised learning methods have become increasingly popular as they do not require labelled data. Instead images act as labels themselves, which ultimately helps in learning invariant representations. Most popular self-supervised methods are pre-text based learning, contrastive learning and non-contrastive learning.

Contrastive learning, introduced in SimCLR [6] showed good performance and surpassed the supervised learning in various downstream tasks. The core idea is based on capturing representations for similar (positive pairs) and dissimilar (negative pairs) images by directly comparing the two images [6]. SimCLR is largely sensitive to batch size and required larger batch size of 4096. This made it computationally expensive, later MoCo [7] introduced the momentum encoder with memory bank for past projections, which not only improved the performance but reduced the batch size to 256. More recently, BYOL [1] and SimSiam [2] introduced the architectures similar to contrastive learning but considered only positive pairs with cosine similarity. BYOL utilised the momentum encoder , whereas SimSiam eliminated the momentum encoder and relied on a shared encoder with stop gradient. These self-supervised methods forces model to learn the invariant representations and are then used for downstream tasks.

Though Self-Supervised Learning (SSL) is still relatively unexplored in the satellite imagery domain, it has shown potential in some recent works. As part of pre-text tasks, work by Vincenzi et al. [11] utilised colourisation as a self-supervised task for pre-training and observed that with more spectral information, representations learned by colourisation varies from representation learned through normal RGB images. Also recently, some works have shown that SSL based methods in remote sensing outperformed the supervised learning based models [12], [13] by utilising temporal information as pre-text task with contrastive learning [12] and other multiple pre-text tasks such as image inpainting, relative positions and instance discrimination [13].

These works motivated our approach for utilising MS and SAR images to learn representations by maximising the similarity between the two views in a non-contrastive learning fashion.

III. Methodology

With the success of contrastive learning, many architectures have been proposed recently, to have better and simplified contrastive learning architectures. Initially SimCLR [6] worked around three major concepts, (i) strong data augmentation, (ii) Multi-Layer Perceptron (MLP) projections, and (iii) contrastive loss.

A. Base Architecture

BYOL [1] proposed more efficient training by eradicating the need for negative pairs, with addition of momentum encoder [7], MLP prediction layer, and L2 normalization loss. Since the BYOL architecture serves as the basis of our modeling, we expand here on its key features:

- *Data Augmentation:* Strong data augmentation can come with strong colour distortion, random crop, resize, Gaussian blur, and random flips. These random augmentations creates two views (x_1, x_2) of an image, which are then used to calculate similarity loss.
- *Encoder (f) & Momentum Encoder (f'):* The basic architecture can be divided into online and target networks. The online network consists of a backbone encoder (f), MLP projection head (z), and MLP prediction head (h). Whereas the target network consisted of the same encoder but with moving average, followed by a MLP projection head. The encoder provides the representations vector for two views as f(x_1) and f(x_2).

- *MLP Projection Head (z):* The projection head consists of linear layer, batch normalization (BN), rectified linear unit (ReLU), and final linear layer.
- *MLP Prediction Head (h):* Again this is the same as MLP projection head.
- *Loss:* The projection (z_2) from the target network are then compared with prediction (h) from the online network. The loss function is a mean square error between L2 normalized predictions of h, and z_2 as given in Equation 1.

$$Loss = 2 - 2 \cdot \frac{h}{\|h\|_2} \cdot \frac{z_2}{\|z_2\|_2} \quad (1)$$

where $\|\cdot\|_2$ is l_2 normalisation.

Recently, SimSiam [2] proposed a similar but simplified architecture, which eradicated the momentum encoder, this reduced the requirement of large batch size. Instead they used same encoder on the target network.

B. Multi-Modal Self Supervised Learning

As SSL is not well explored in EO domain, our goal for training is to (i) explore the potential of SSL models such as BYOL and SimSiam for multi-modal satellite imagery, and (ii) learn EO domain representations for downstream tasks with small labelled data.

While the general usefulness of SSL methods is now clear and we see there being a direct benefit in applying vanilla SSL methods to individual bands in EO data. We argue that taking advantage of the multi-modal nature of satellite data may in fact lead to more natural generalisation than can be achieved with traditional 3 channel RGB imagery. Therefore rather than relying simply on simple augmentation techniques to generate alternative image views, in our work we utilised random MS band and SAR bands as the two views , i.e., x_1 is extracted from MS bands ms_n^i while x_2 is extracted from SAR bands sar_m^i, where n and m are random bands sampled from same scene i. Both images are however required to be fully overlapping in geographic views. In addition from selecting from distinct modal sources we do also apply augmentation in the form of random flip, random Gaussian blur, random rotation and random crop-resize as augmentation for both the views on mini batch of data. Other forms of traditional augmentation such as colour distrotion are not applicable here.

IV. Dataset

For this work we utilised the Sen12MS dataset, as it is a diverse repository with 180K triplets of Sentinel-1/2 imagery, as well as MODIS which provides weakly segmented labels. In addition to this we also make use of the recently released IEEE Data Fusion Contest (DFC) subset of Sen12MS data which contains 12,228 pairs of labelled Sentinel-1/2 data. Considering these dataset we utilised two dataset sizes for pre-training our two models: (i) a baseline model with data from DFC with 11,242 pairs of Sentinel-1/2 and (ii) another model with 90K datasets from the original Sen12MS dataset which is 50% of original dataset size. For the work we utilised 10 MS bands, namely Red, Green, Blue, Vegetation Red Edge 5/6/7, NIR, NNIR, SWIR1, SWIR2, and two SAR bands.

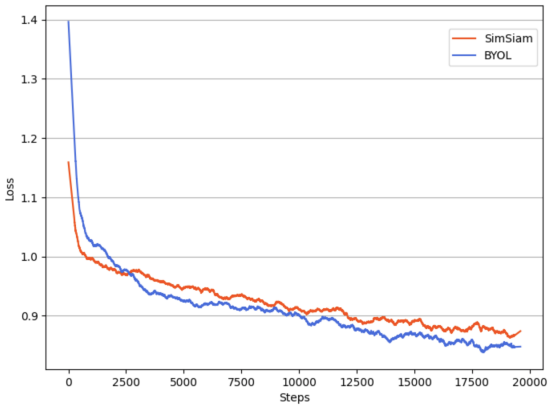

Fig. 2: Loss for BYOL and SimSiam with 90K data. Lines shown are a rolling average over the training steps.

V. Experiment

We used the Sen12MS data for training of both BYOL and SimSiam. After training we evaluated the learned representations on a downstream task, i.e., scene classification on the EuroSAT dataset [14].

The model architecture consisted of ResNet50 [15] with average pooling layer as backbone encoder (f), MLP projection head (z), and MLP prediction head (h). Both MLP heads are similar to BYOL which consisted of linear layer, batch normalization layer, rectified linear units (ReLU) and final linear layer.

We trained two pre-trained models i.e. the baseline model, which is trained on small DFC dataset that is 6% of original dataset, had MLP head hidden dimension as 256 and output dimension as 128. And another model (BYOL-90K, SimSiam-90K) trained on 90K i.e. 50% of Sen12MS dataset, had hidden and output dimensions as per the original BYOL and SimSiam implementation i.e. 4096 and 256 for BYOL and 2048 and 512 for SimSiam.

For the optimiser we utilised Adam optimiser with an initial learning rate of 3e-4. The batch size kept as 32 and baseline model trained for 200 epochs while SimSiam-90K and BYOL-90K model trained for approx 400 epochs. The input image for both the network is a augmented random MS band from 10 MS bands and random SAR image from two bands.

For training we utilised Tesla K80 GPU, with respect to training efficiency, BYOL took approximately twice the time taken by SimSiam.

After the completion of training we discarded the weights of target network and utilised only the encoder weights for the downstream scene classification task.

VI. Evaluation

For the evaluation of our model we utilised the EuroSAT dataset which consisted of 27K, 64x64 images of land coverage from Sentinel-2. We evaluated BYOL and SimSiam

Band	ImageNet Pre-Trained	Baseline SimSiam	Baseline BYOL	SimSiam-90K	BYOL-90K
Blue	0.65	0.74	0.67	0.85	**0.87**
Green	0.62	0.77	0.77	0.86	**0.88**
Red	0.56	0.76	0.77	0.85	**0.86**
RE5	0.61	0.70	0.72	0.80	0.80
RE6	0.51	0.68	0.69	**0.82**	0.74
RE7	0.47	0.70	0.69	**0.81**	0.74
NIR	0.41	0.75	0.74	**0.83**	0.79
NNIR	0.44	0.50	0.52	0.59	0.59
SWIR11	0.58	0.72	0.70	0.77	**0.83**
SWIR12	0.47	0.67	0.68	**0.81**	0.73

TABLE I: Linear evaluation results for all 10 bands on EuroSAT data for 100 epochs. Weighted average F-1 score is used as evaluation metrics.

trained on EO data, along with model trained on ImageNet data (non-EO) to compare performance.

For fair evaluation of our BYOL, SimSiam and ImageNet pre-trained weights, we froze weights for all layers of ResNet50 and trained only the linear classifier with hidden dimension of 128 and softmax layer. For loss function we utilised the cross entropy loss function. In order to compare with ImageNet pre-trained ResNet50 we copied the single channel of an image to the 3-channels prior to application to the input.

As shown in the Table I our baseline BYOL and SimSiam do not show notable difference in results over the 10 bands. Even though the baseline models are trained only on 6% of the datasets, their performance is already outperformed more traditional ImageNet pre-trained models. We see that by increasing the training data for SimSiam and BYOL a jump of 8-10% in performance of most bands is seen, which considerably outperforms the performance of a model derived from non-EO ImageNet pre-trained data.

Considering the relative performance of SimSiam and BYOL, we see that both models perform relatively similar with matching performances on a number of bands. However these performance differences may be due to the small batch sizes that we must make use of due to hardware limitations. Considering the overall training performance, we see from Figure 2 that reductions in loss for both models were broadly continuous over the course of training. Again due to hardware limitation issues we could not allow training to continue until validation performance reduced. We believe that these results offer great promise and that with an increase of data size and computational hardware that further significant improvements can be achieved.

VII. Conclusion

This work applied the self-supervised approach to EO domain data to learn better representations in order to solve the large labelled data availability problem. The SimSiam and BYOL models when trained with only 90K Sen12MS samples outperformed the ImageNet trained model by maximising the similarity between MS band and SAR images. This work also showed that both SimSiam and BYOL model have potential to improve further by increasing the data and training time.

For the work in progress we are continuously evaluating the performance of pre-trained models on Sen12MS data and BigEarthNet dataset. Also our work will explore the applicability for three or more band images as input to have model for RGB images and more.

We emphasise that this work is preliminary and utilising full dataset and other data input strategies can improve the performance of this work.

VIII. Acknowledgement

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106 at the ADAPT SFI Research Centre at Technological University Dublin. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant No. 13/RC/2106.

References

[1] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *arXiv preprint arXiv:2006.07733*, 2020.

[2] X. Chen and K. He, "Exploring simple siamese representation learning," *arXiv preprint arXiv:2011.10566*, 2020.

[3] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1734–1747, 2015.

[4] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[5] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[7] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[8] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 778–782, 2017.

[9] D. Ienco, R. Interdonato, R. Gaetano, and D. H. T. Minh, "Combining sentinel-1 and sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 11–22, 2019.

[10] J. D. Bermudez, P. N. Happ, R. Q. Feitosa, and D. A. Oliveira, "Synthesis of multispectral optical images from sar/optical multitemporal data using conditional generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1220–1224, 2019.

[11] S. Vincenzi, A. Porrello, P. Buzzega, M. Cipriano, P. Fronte, R. Cuccu, C. Ippoliti, A. Conte, and S. Calderara, "The color out of space: learning self-supervised representations for earth observation imagery," *arXiv preprint arXiv:2006.12119*, 2020.

[12] K. Ayush, B. Uzkent, C. Meng, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," *arXiv preprint arXiv:2011.09980*, 2020.

[13] C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, "Remote sensing image scene classification with self-supervised paradigm under limited labeled samples," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[14] P. Helber, B. Bischke, A. Dengel, and D. Borth, "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*, vol. 12, no. 7, pp. 2217–2226, 2019.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.