IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9554776

BAG-OF-WORDS FOR TRANSFER LEARNING

Iulia Calota[1], Daniela Faur[1], Mihai Datcu[1,2]

[1]Research Center for Spatial Information (CEOSpaceTech), University POLITEHNICA of Bucharest (UPB), [2]Earth Observation Center (EOC), German Aerospace Center (DLR)

ABSTRACT

Although the number of labeled datasets in Earth Observation (EO) is increasing, there is still a major gap between the Deep Learning (DL) classifiers designed in this field versus the models in Computer Vision. This gap is produced mainly by the number of datasets available, but also by the diversity of data. In EO, there are different sensors acquiring images, from multispectral (MS) or hyperspectral data, to SAR imagery. In this paper, we want to demonstrate how to reduce the divergence created by the diversity of data. We trained several DL architectures on Bag-of-Words from large-scale MS and SAR datasets, and then we used transfer learning on smaller ones and evaluated the results. With this method, we demonstrate that a DL architecture can be trained with any type of large-scale data, transformed into Bag-of-Words, and the trained model can be used further on other types of data, without regard on the number of channels.

Index Terms— Bag-of-Words, Transfer Learning, multispectral data, SAR data, Deep Learning

1. INTRODUCTION

Deep Learning (DL) has become a powerful tool in training classifiers for Earth Observation (EO). Reviews of DL in EO and Remote Sensing (RS) enunciate the advantages and the challenges posed by this relatively new field [1], [2], [3]. Papers on new types of algorithms and architectures for all kinds of data are published every day. In [4], the authors propose a new architecture of the deep Q-network for image classification in polarimetric SAR data. For multi-label classification of multispectral images, the authors in [5] propose a new method that combine a KBranch convolutional neural network and long short-term memory networks. In [6], the authors propose a method based on active deep learning for the classification of hyperspectral images.

Other papers concentrate on applying state-of-the-art architectures used in Computer Vision to imagery of Earth Observation. In [7], the authors apply convolutional neural networks in SAR-optical image matching with Siamese networks. In [8], the authors seek a standardized workflow in Deep Learning.

A known problem in the conjunction of Earth Observation and Deep Learning is the reduced number of labeled datasets. This disadvantage develops from the difficult process of labeling Remote Sensing images. Multispectral images can usually include the RGB channels, as in the case of Sentinel-2 scenes [9], which means that visual inspection is easier. However, this too requires expert inspection, not only in Earth Observation, but also in agriculture, oceanography or urban development. In the case of SAR images, visual labeling is even more challenging. Some automatic labelling algorithms have been proposed, like the ones found in [10] or [11]. Nonetheless, it is difficult to achieve a large-scale dataset, such as ImageNet [12].

One method to overcome this disadvantage is transfer learning, as seen in papers like [13], [14] or [15]. There is one limitation in the current framework of transfer learning for Earth Observation: the compatibility of the input dimension. When using a model trained on Computer Vision datasets such as ImageNet, the images on which the transfer learning or the fine-tuning is done should have the same input dimension as the samples in ImageNet. This means that we cannot use the whole potential of multispectral images or the high resolution of SAR data in classification.

In this paper, we propose a new method for using transfer learning in Earth Observation. With the help of two large-scale datasets, one with multispectral images and the other with SAR images, we create two Bag-of-Words models. These Bags-of-Words are trained on a Deep Neural Network architecture. Afterwards, we chose some smaller-scale datasets with optical, multispectral and SAR data, which are also transformed in Bags-of-Words. These smaller-scale Bags-of-Words are then used for transfer learning on both initial models. With this method, we demonstrate that multispectral images can be used at full potential even on SAR imagery, or SAR images can be used on pre-trained networks with multispectral data.

The rest of the paper is structured as follows: in the second section, we present our methods, the used datasets and architectures. In the third part, we present the results. We end with conclusions and discussions.

2. METHOD

In this section, we present some theoretical aspects of our work, we elaborate more the algorithm used and give details on the datasets that we used.

In order to use our method, a pre-processing of the patches in the dataset is needed. In [16], the authors reduce the training time of convolutional neural networks by using either histograms or Bag-of-Words instead of the original patches of the dataset. This same pre-processing is used in our paper. The algorithm for obtaining the Bag-of-Words can be found in [17]. In this method, which is faster and consumes less memory than the original Bag-of-Words, a random dictionary is generated from the whole dataset. Afterwards, the Bag-of-Words is created by iterating through each image of the dataset and comparing the dictionary to the patches through Nearest Neighbor. If the initial dimension at the input of the network was of the order of 100000, after transforming it to Bag-of-Words, the dimension is of the order of 100. This happens, because the spectral information, given by the channels of the patches, is not anymore present. As stated in [17], we chose our dictionary size as 250.

The datasets used in this paper are BigEarthNet [18], EuroSAT [19], UC-Merced [20] and OpenSARUrban [21]. BigEarthNet contains two separate large-scale datasets, BigEarthNet-S2, containing multispectral patches, and BigEarthNet-S1, containing SAR patches [22]. This dataset contains 590326 patches, both in multispectral and SAR data. The patches have the size 120x120x12, 12 being the number of channels. For the classification, there are 43 labels available, and each patch can have multiple labels, making this dataset multi-class and multi-label.

EuroSAT contains 27000 multispectral patches. The size of each patch is 64x64x13. It is a multi-class dataset, containing 10 classes. Of all the datasets used, EuroSAT resembles BigEarthNet-S2 the most, as the patches originate from the same sensor – Sentinel-2.

UC-Merced contains 2100 optical patches. Each patch has the size 256x256x3. This multi-class dataset contains 21 classes. This is the most balanced dataset used, as each class is represented by 100 samples. A particularity of thus dataset is the data type – uint8, whereas the other datasets use uint16 (multispectral) or float32 (SAR).

OpenSARUrban has 33358 SAR patches. Each patch has a size of 100x100x2, one channel for each polarization. It contains 10 labels of urban scenes, and it is a multi-class dataset. This dataset is similar to BigEarthNet-S1, because the source sensor is the same – Sentinel-1 in Interferometric Wide swath mode used for acquiring Ground Range Detected products.

The Deep Learning architectures we used are convolutional neural networks. We used a three-layered architecture, depicted in Fig. 1. Such convolutional neural networks consist mainly of three types of layers: convolutional layers, pooling layers and fully-connected layers. Convolutional layers carry out the convolution operation between windows of the input and a filter. Pooling layers are used for down-sampling and translation invariance. Fully-connected layers flattens their input and is usually used also for classification [23].

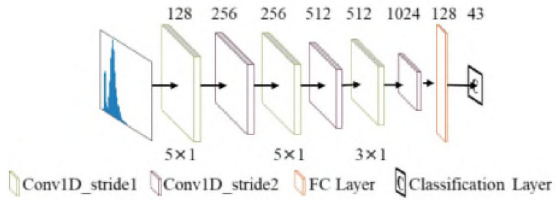

Fig. 1: Three-layered architecture used for training of BigEarthNet.

In order to use these networks for training Bag-of-Words, some adjustments were needed. Instead of 2D convolutional layers, we used 1D convolutional layers (Conv1D_stride1). As histograms-like structures, such as Bag-of-Words, do not need translation invariance provided by the pooling layers [23], we used Conv1D layers with stride of two (Conv1D_stride2).

The Bag-of-Words resulting from BigEarthNet-S1 and BigEarthNet-S2 are used for training the above-described architectures. These pre-trained models are then fine-tuned on the Bag-of-Words acquired from EuroSAT, UC-Merced and OpenSARUrban. In order to see the improvements of using this kind of transfer learning, we trained previously these datasets on the above-mentioned architectures as they are (without using Bag-of-Words) and displayed the results in Table 1. As metrics we used the widely used precision (P), recall (R), F1-score (F1) and F2-score (F2). In the Table, "BoW-BEN-S1" is the model trained with Bag-of-Words generated from BigEarthNet-S1 and "BoW-BEN-S2".

Table 1: Initial results after training all datasets on a three-layered architecture.

Dataset	**P**	**R**	**F1**	**F2**
EuroSAT	0.63	0.55	0.59	0.57
UC-Merced	0.52	0.41	0.46	0.43
OpenSARUrban	0.55	0.51	0.53	0.52
BoW-BEN-S1	0.75	0.65	0.7	0.67
BoW-BEN-S2	0.81	0.73	0.76	0.74

In all our runs, we trained the networks for 100 epochs. All datasets were split in 80% training data and 20% validation data. As optimizer, we used Stochastic Gradient Descent in all trainings. The results in Table 1 for EuroSAT, UC-Merced and OpenSARUrban are trained on a similar architecture to the CNN in Fig. 1. As they have an additional dimension, instead of Conv1D_stride1 layers, Conv2D layers were used and MAX pooling layers replaced the

Conv1D_stride2 layers. Otherwise, the number of filters and other parameters remained the same.

3. RESULTS

In this section, we present our results. The results are displayed in Table 2. In the Table, "BoW-EuroSAT" is the dataset consisting of Bag-of-Words generated from EuroSAT. The same applies for the ones generated from UC-Merced and OpenSARUrban. We fine-tuned the models for 100 epochs. The optimizer used was also Stochastic Gradient Descent.

Table 2: Results after transfer learning.

Dataset	**Pre-trained CNN**	**P**	**R**	**F1**	**F2**
BoW-EuroSAT	BoW-BEN-S1	0.87	0.86	0.86	0.86
	BoW-BEN-S2	0.88	0.87	0.88	0.87
BoW-UC-Merced	BoW-BEN-S1	0.81	0.79	0.8	0.8
	BoW-BEN-S2	0.85	0.81	0.83	0.81
BoW-OpenSAR Urban	BoW-BEN-S1	0.75	0.63	0.69	0.65
	BoW-BEN-S2	0.78	0.7	0.74	0.72

Although the three-layered architecture in Fig. 1 has around 7.78 million parameters to train, overfitting in the case of BigEarthNet is not prominent. This happens because of the reduced input dimension. As the datasets become smaller, even with transfer learning, the overfitting becomes large. However, this is an expected behavior. For more complicate and deeper architectures, we expect to see more overfitting, but the results to be similar or to improve.

Overall, this method produces good results, with a good performance for all datasets, regardless of their sizes. Especially for UC-Merced, the smallest dataset, transfer learning with Bag-of-Words delivers good results. The differences in performance between the pre-trained network with SAR data and the one trained with MS data can be seen also in Table 1. One explanation for this phenomenon comes from the fact that the original patches contain mainly vegetation classes, which are not as distinguishable in SAR scenes as they are in MS imagery.

Transfer learning or fine-tuning with Bag-of-Words as input is also a faster approach. For small datasets, this is not very visible, but for larger datasets, as BigEarthNet, this can be useful when training deeper networks. This happens, because the dataset consisting of Bag-of-Words occupies a smaller amount of memory. Furthermore, the number of parameters decreases, as the input dimension of the CNN is reduced.

Another advantage of using Bag-of-Words is that smaller architectures deliver good results as deeper ones. We re-run the method also on VGG19 and the performance was similar for training and fine-tuning with Bags-of-Words. However, as expected, overfitting was larger on such complex architectures, as there are more parameters to train.

4. CONCLUSION

In this paper, we demonstrated how to reduce the gap between data from different sensors, data types and classification types of different datasets. We used multispectral, optical and SAR datasets to prove that, with pre-processing and transforming the datasets into Bags-of-Words, we can use transfer learning between different types of data, regardless of the initial dimension of the patches.

However, our method does not solve all problems. One of the major difficulties we encountered in our work was at fine-tuning. Each dataset used for fine-tuning needed different settings to achieve a good performance. We needed different settings of learning rates, layers to fine-tune, etc. This is a downside encountered usually in transfer learning and fine-tuning, which requires a lot of time. Moreover, in some runs the training was stuck on a local minima with no changes or significant improvement in performance and metrics. However, because of Bag-of-Words, the training time is reduced significantly.

In the future, we want to test this method also on other architectures and to find a standardized method for fine-tuning.

5. REFERENCES

[1] G. Cheng, X. Xie, J. Han, L. Guo and G. -S. Xia, "Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735-3756, 2020, doi: 10.1109/JSTARS.2020.3005403.

[2] Z. Zheng, L. Lei, H. Sun and G. Kuang, "A Review of Remote Sensing Image Object Detection Algorithms Based on Deep Learning," 2020 *IEEE 5th International Conference on Image, Vision and Computing* (ICIVC), Beijing, China, 2020, pp. 34-43, doi: 10.1109/ICIVC50857.2020.9177453.

[3] A. Alem and S. Kumar, "Deep Learning Methods for Land Cover and Land Use Classification in Remote Sensing: A Review," *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)* (ICRITO), Noida, India, 2020, pp. 903-908, doi: 10.1109/ICRITO48877.2020.9197824.

[4] K. Huang, W. Nie and N. Luo, "Fully Polarized SAR imagery Classification Based on Deep Reinforcement Learning Method Using Multiple Polarimetric Features," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 10, pp. 3719-3730, Oct. 2019, doi: 10.1109/JSTARS.2019.2913445.

[5] G. Sumbul and B. Demİr, "A Deep Multi-Attention Driven Approach for Multi-Label Remote Sensing Image Classification,"

in *IEEE Access*, vol. 8, pp. 95934-95946, 2020, doi: 10.1109/ACCESS.2020.2995805.
[6] P. Liu, H. Zhang and K. B. Eom, "Active Deep Learning for Classification of Hyperspectral Images," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 2, pp. 712-724, Feb. 2017, doi: 10.1109/JSTARS.2016.2598859.
[7] L. H. Hughes, N. Merkle, T. Bürgmann, S. Auer and M. Schmitt, "Deep Learning for SAR-Optical Image Matching," *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Yokohama, Japan, 2019, pp. 4877-4880, doi: 10.1109/IGARSS.2019.8898635.
[8] T. Landry et al., "Applying Machine Learning to Earth Observations In A Standards Based Workflow," *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Yokohama, Japan, 2019, pp. 5567-5570, doi: 10.1109/IGARSS.2019.8898032.
[9] https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook, retrieved: 19.01.2021.
[10] X. Zhang, J. Liu and X. Chi, "Auto-labeling algorithms on CA based interactive segmentation for high resolution remote sensing images," *2015 Sixth International Conference on Intelligent Control and Information Processing* (ICICIP), Wuhan, 2015, pp. 367-370, doi: 10.1109/ICICIP.2015.7388198.
[11] M. Chi, Z. Sun, Y. Qin, J. Shen and J. A. Benediktsson, "A Novel Methodology to Label Urban Remote Sensing Images Based on Location-Based Social Media Photos," in *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1926-1936, Oct. 2017, doi: 10.1109/JPROC.2017.2730585.
[12] http://www.image-net.org/, retrieved: 19.01.2021.
[13] X. Liu, M. Chi, Y. Zhang and Y. Qin, "Classifying High Resolution Remote Sensing Images by Fine-Tuned VGG Deep Networks," *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, Valencia, 2018, pp. 7137-7140, doi: 10.1109/IGARSS.2018.8518078.
[14] Z. Huang, C. O. Dumitru, Z. Pan, B. Lei and M. Datcu, "Classification of Large-Scale High-Resolution SAR Images With Deep Transfer Learning," in *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 1, pp. 107-111, Jan. 2021, doi: 10.1109/LGRS.2020.2965558.
[15] D. Marmanis, M. Datcu, T. Esch and U. Stilla, "Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks," in *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 105-109, Jan. 2016, doi: 10.1109/LGRS.2015.2499239.
[16] I. Calota, D. Faur and M. Datcu. "DNN-Based, Semantic Extraction: Fast Learning from Multispectral Signatures", in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, Waikoloa, 2020.
[17] S. Cui, G. Schwarz and M. Datcu, "Image classification: No features, no clustering," 2015 *IEEE International Conference on Image Processing* (ICIP), Quebec City, QC, 2015, pp. 1960-1964, doi: 10.1109/ICIP.2015.7351143.
[18] G. Sumbul, M. Charfuelan, B. Demir and V. Markl, "Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding," *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Yokohama, Japan, 2019, pp. 5901-5904, doi: 10.1109/IGARSS.2019.8900532.
[19] P. Helber, B. Bischke, A. Dengel and D. Borth, "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217-2226, July 2019, doi: 10.1109/JSTARS.2019.2918242.
[20] Yi Yang and Shawn Newsam, "Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification," *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (ACM GIS), 2010.
[21] J. Zhao, Z. Zhang, W. Yao, M. Datcu, H. Xiong and W. Yu, "OpenSARUrban: A Sentinel-1 SAR Image Dataset for Urban Interpretation," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 187-203, 2020, doi: 10.1109/JSTARS.2019.2954850.
[22] http://bigearth.net/, retrieved: 19.01.2021.
[23] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature*, 2015.