BIGEARTHNET: A LARGE-SCALE BENCHMARK ARCHIVE FOR REMOTE SENSING IMAGE UNDERSTANDING

Gencer Sumbul\textsuperscript{1}, Marcela Charfuelan\textsuperscript{2}, Begüm Demir\textsuperscript{1}, Volker Markl\textsuperscript{1,2}

\textsuperscript{1}Technische Universität Berlin, \textsuperscript{2}DFKI GmbH

\textbf{ABSTRACT}

This paper presents the BigEarthNet that is a new large-scale multi-label Sentinel-2 benchmark archive. The BigEarthNet consists of 590, 326 Sentinel-2 image patches, each of which is a section of i) 120 \times 120 pixels for 10m bands; ii) 60 \times 60 pixels for 20m bands; and iii) 20 \times 20 pixels for 60m bands. Unlike most of the existing archives, each image patch is annotated by multiple land-cover classes (i.e., multi-labels) that are provided from the CORINE Land Cover database of the year 2018 (CLC 2018). The BigEarthNet is significantly larger than the existing archives in remote sensing (RS) and thus is much more convenient to be used as a training source in the context of deep learning. This paper first addresses the limitations of the existing archives and then describes the properties of the BigEarthNet. Experimental results obtained in the framework of RS image scene classification problems show that a shallow Convolutional Neural Network (CNN) architecture trained on the BigEarthNet provides much higher accuracy compared to a state-of-the-art CNN model pre-trained on the ImageNet (which is a very popular large-scale benchmark archive in computer vision). The BigEarthNet opens up promising directions to advance operational RS applications and research in massive Sentinel-2 image archives.

\textbf{Index Terms}— Sentinel-2 image archive, multi-label image classification, deep neural network, remote sensing

\section{1. INTRODUCTION}

Recent advances in deep learning have attracted great attention in remote sensing (RS) due to the high capability of deep networks (e.g., Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Generative Adversarial Networks (GAN)) to model the high-level semantic content of RS images. To train such networks, a very large training set is needed with a high number of annotated images in order to learn effective models with several different parameters. To the best of our knowledge, publicly available RS image archives contain only a small number of annotated images and a large-scale benchmark archive does not yet exist. Thus, the lack of a large training set is an important bottleneck that prevents the use of deep learning in RS. In order to address this problem, fine-tuning deep networks pre-trained on large-scale computer vision archives (e.g., ImageNet) is considered in RS community. However, such an approach has several limitations related to the differences on the characteristics of images between computer vision and RS. Additionally, in the existing archives, RS images are annotated by single high-level category labels that are related to the most significant content of the image. However, RS images typically contain multiple classes and thus each image can be simultaneously associated with different land-cover class labels (i.e., multi-labels). To overcome these problems, we introduce the BigEarthNet that is a new large-scale Sentinel-2 archive\textsuperscript{1} and contains 590, 326 Sentinel-2 image patches. Each patch is annotated with multi-labels provided from the CORINE Land Cover database, which is updated in 2018 (CLC 2018). We propose our archive as a sufficient source for RS image analysis with deep learning. In order to test the BigEarthNet on RS image analysis problems, we focus our attention on image scene classification. To this end, we consider a shallow CNN architecture to be trained on the BigEarthNet. We compare the results obtained by this network with the Inception-v2 [1] pre-trained on the ImageNet. We believe that it will make a significant advancement in terms of developments of algorithms for the analysis of large-scale RS image archives.

\section{2. LIMITATIONS OF EXISTING REMOTE SENSING IMAGE ARCHIVES}

Most of the benchmark archives in RS (UC Merced Land Use Dataset [2], WHU-RS19 [3], RSSCN7 [4], SIRI-WHU [5], AID [6], NWPU-RESISC45 [7], RSI-CB [8], EuroSat [9] and PatternNet [10]) contain a small number of images annotated with single category labels. Table 1 presents the list of the existing archives. These archives become popular for the implementation, evaluation and validation of algorithms in the context of image classification, search and retrieval tasks. However, RS community encounters critical limitations, while using these archives for applying deep learning based approaches. One of the most critical limitations is that the number of annotated images included in the existing archives is very small. Thus, they are insufficient to train modern deep neural networks to reach a high generalization ability as the models may overfit dramatically when using small training sets. In details, training such networks on the existing archive images suffers from the problem of learning a large number

\textsuperscript{1}The BigEarthNet is available at \url{http://bigearth.net}.
Table 1: List of the existing RS archives.

| Archive Name | Image Type | Annotation Type | Number of Images |
|--------------|------------|-----------------|------------------|
| UC Merced    | Aerial RGB | Single Label [2] | 2,100            |
| WHU-RS19 [3] | Aerial RGB | Single Label     | 1,005            |
| RISSC7 [4]   | Aerial RGB | Single Label     | 2,800            |
| SIRI-WHU [5] | Aerial RGB | Single Label     | 2,400            |
| AID [6]      | Aerial RGB | Single Label     | 10,000           |
| NWPU-RESISC45 [7] | Aerial RGB | Single Label | 31,500           |
| RSI-CB [8]   | Aerial RGB | Single Label     | 36,707           |
| EuroSat [9]  | Satellite Multispectral | Single Label | 27,000           |
| PatternNet [10] | Aerial RGB | Single Label | 30,400           |

Table 2: The considered Level-3 CLC classes and the number of images associated with each land-cover class in the BigEarthNet.

| Land-Cover Classes                                      | Number of Images |
|---------------------------------------------------------|------------------|
| Mixed forest                                            | 217,119          |
| Coniferous forest                                       | 211,703          |
| Non-irrigated arable land                               | 196,098          |
| Transitional woodland/shrub                             | 173,506          |
| Broad-leaved forest                                     | 198,944          |
| Land principally occupied by agriculture, with significant areas of natural vegetation | 147,095 |
| Complex cultivation patterns                            | 107,786          |
| Pastures                                                | 103,354          |
| Water bodies                                            | 84,811           |
| Sea and ocean                                           | 81,612           |
| Discontinuous urban fabric                              | 69,872           |
| Agro-forestry areas                                     | 30,674           |
| Peatbogs                                                | 23,207           |
| Permanently irrigated land                              | 135,898          |
| Industrial or commercial units                          | 128,955          |
| Natural grassland                                       | 12,835           |
| Olive groves                                            | 12,938           |
| Scrutrophilous vegetation                               | 11,241           |
| Continuous urban fabric                                 | 11,754           |
| Water courses                                           | 10,572           |
| Vineyards                                               | 9,507            |
| Annual crops associated with permanent crops            | 7,022            |
| Inland marshes                                          | 6,236            |
| Moors and heathland                                     | 5,388            |
| Sport and leisure facilities                            | 5,385            |
| Fruit trees and berry plantations                       | 4,754            |
| Mineral extraction sites                                | 4,618            |
| Rice fields                                             | 3,793            |
| Road and rail networks and associated land              | 3,384            |
| Bare rock                                               | 3,277            |
| Green urban areas                                       | 1,786            |
| Beaches, dunes, sands                                   | 1,378            |
| Sparsely vegetated areas                                | 1,563            |
| Salt marshes                                            | 1,502            |
| Coastal lagoons                                         | 1,405            |
| Construction sites                                      | 1,174            |
| Estuaries                                               | 1,086            |
| Intertidal flats                                        | 1,083            |
| Airports                                                | 919              |
| Dump sites                                              | 950              |
| Port areas                                              | 509              |
| Saltines                                                | 424              |
| Burnt areas                                             | 328              |

of parameters that prevents the accurate characterization of semantic content of RS images. To this end, fine-tuning the models pre-trained on ImageNet is used as a transfer learning approach. However, the profound differences between the image properties of computer vision and RS limit the accurate characterization of RS images when fine-tuning approach is applied. As an example, Sentinel-2 images have 13 spectral bands associated to varying and lower spatial resolutions with respect to computer vision images. There are also differences in the ways that the category labels of computer vision and RS are defined for the semantic content of an image. Thus, fine-tuning pre-trained models for RS images may not be generally applicable to reduce this semantic gap and therefore may lead to weak discrimination ability for land-cover classes. Another limitation of existing archives is that they contain images annotated by single high-level category labels, which are related to the most significant content of the image. However, RS images generally contain multiple classes so that they can be simultaneously associated to different land-cover class labels (i.e., multi-labels). Hence, a benchmark archive consisting of images annotated with multi-labels is required. Although the archive presented in [11] contains images with multi-labels, the sample size of this archive is very small to be efficiently utilized for deep learning. Another limitation of RS image archives is that since researchers generally do not have free access to satellite data together with their annotation, most of the benchmark archives contain aerial images with only RGB image bands. Unavailability of a high number of annotated satellite images prevents to employ deep learning methods in a convenient way for the complete understanding of huge amount of freely accessible satellite data (e.g., Sentinel-1, Sentinel-2). Although the benchmark archive proposed in [9] includes annotated satellite images, the number of images is still small. Aforementioned limitations of existing archives reveal the need for a large-scale RS benchmark archive to be used for training deep neural networks instead of the ImageNet.

3. THE BIGEARTHNET ARCHIVE

To overcome the limitations of existing archives, we introduce the BigEarthNet that is the first large-scale benchmark archive in RS. We have constructed our archive by selecting 125 Sentinel-2 tiles acquired between June 2017 and May 2018. Considered tiles are distributed over the 10 countries (Austria, Belgium, Finland, Ireland, Kosovo, Lithuania, Luxembourg, Portugal, Serbia, Switzerland) of Europe. It is worth noting that considered tiles are associated to cloud cover percentage less than 1%. All tiles were atmospherically corrected by using Sentinel-2 Level 2A product generation and formatting tool (sen2cor) of ESA. Among 13 Sentinel-2 spectral bands, 10\textsuperscript{th} band, for which surface information is not embedded, was excluded. After the tile selection and preliminary processing steps were carried out, selected tiles were divided into 590,326 non-overlapping image patches. Each patch (denoted as image hereafter) is a section of i) 120 × 120 pixels for 10m bands; ii) 60 × 60 pixels for 20m bands; and iii) 20 × 20 pixels for 60m bands. We have associated each image with one or more land-cover class labels (i.e., multi-labels) provided from the CORINE Land Cover (CLC) database of the year 2018 (CLC 2018). The CLC inventory was produced by the Eionet National Reference Centres on Land Cover with the
4. EXPERIMENTAL RESULTS

In the experiments, we have used the BigEarthNet archive in the framework of RS image scene classification problems. To this end, we selected a shallow CNN architecture, which consists of three convolutional layers with 32, 32 and 64 filters having $5 \times 5$, $5 \times 5$ and $3 \times 3$ filter sizes, respectively. We added one fully connected (FC) layer and one classification layer to the output of last convolutional layer. In all convolution operations, zero padding was used. We also applied max-pooling between layers. We considered to utilize: i) only RGB channels (denoted as S-CNN-RGB); and ii) all spectral channels (denoted as S-CNN-All). For the S-CNN-All, cubic interpolation was applied to 20 and 60 meter bands of each image to have the same pixel sizes associated with each band. Weights of the S-CNN-RGB and the S-CNN-All were randomly initialized and we trained both networks from scratch on the BigEarthNet images. In order to show the effectiveness of the BigEarthNet to be used in training, we compared the results with fine-tuning one of the recent pre-trained deep learning architectures. We considered the Inception-v2 network [1] pre-trained on ImageNet as a state-of-the-art architecture. We used the feature vector extracted from the layer just before the softmax layer of the Inception-v2. To employ fine-tuning, we fixed the model weights of the Inception network. We added one FC and one classification layer to the network and just fine-tuned these layers by using the RGB channels of the BigEarthNet images. In the experiments, 70,987 images that are fully covered by seasonal snow, cloud and cloud shadow were eliminated. Then, among the remaining images, we randomly selected: i) 60% of images to derive a training set; ii) 20% of images to derive a validation set; and iii) 20% of images to derive a test set. Both for fine-tuning and training from scratch, we selected the number of epochs as 100 and Stochastic Gradient Descent algorithm is employed in order to decrease the sigmoid cross entropy loss (which aims at maximizing the log-likelihood of each land-cover class throughout all training images). For the performance metrics of experiments, we employed precision ($P$), recall ($R$), $F_1$ and $F_2$ scores, which are widely used metrics for multi-label image classification. As it can be seen from Table 4, the S-CNN-RGB provides better performance than the Inception-v2 in all met-

---

Footnotes:
1. https://land.copernicus.eu/user-corner/technical-library/corine-land-cover-nomenclature-guidelines
2. The lists of images fully covered by seasonal snow, cloud and cloud shadow are available at http://bigearth.net/#downloads.
We would like to note that we plan to regularly enrich the BigEarthNet by increasing the number of annotated Sentinel-2 image patches annotated by crowdsource data. This work was supported by the European Research Council under the ERC Starting Grant BigEarth (759764) and the German Ministry for Education and Research as BBDC (01IS14013A).

6. ACKNOWLEDGEMENTS

This paper presents a large-scale benchmark archive that consists of 500,326 Sentinel-2 image patches annotated by multi-labels for RS image understanding. We believe that the BigEarthNet will make a significant advancement for the use of deep learning in RS by overcoming the current limitations of the existing archives. Experimental results show the effectiveness of training even a simple neural network on the BigEarthNet from scratch compared to fine-tuning a state-of-the-art deep learning model pre-trained on ImageNet. We would like to note that we plan to regularly enrich the BigEarthNet by increasing the number of annotated Sentinel-2 images.

5. CONCLUSION

This work was supported by the European Research Council under the ERC Starting Grant BigEarth (759764) and the German Ministry for Education and Research as BBDC (01IS14013A).

7. REFERENCES

[1] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in IEEE Conf. Comput. Vis. Pattern Recog., 2016.
[2] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in Int. Conf. Adv. Geogr. Inf. Syst., 2010.
[3] W. Shao, W. Yang, and G. S. Xia, “Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification,” Int. J. Remote Sens., vol. 34, no. 23, pp. 8588–8602, 2013.
[4] Q. Zou, L. Ni, T. Zhang, and Q. Wang, “Deep learning based feature selection for remote sensing scene classification,” IEEE Geosci. Remote Sens. Lett., vol. 12, no. 11, pp. 2321–2325, November 2015.
[5] B. Zhao, Y. Zhong, G. Xia, and L. Zhang, “Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 4, pp. 2108–2123, April 2016.
[6] G. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, “Aid: A benchmark data set for performance evaluation of aerial scene classification,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 7, pp. 3965–3981, July 2017.
[7] G. Cheng, J. Han, and X. Lu, “Remote sensing image scene classification: Benchmark and state of the art,” Proc. IEEE, vol. 105, no. 10, pp. 1865–1883, October 2017.
[8] H. Li, C. Tao, Z. Wu, J. Chen, J. Gong, and M. Deng, “Rsi-ch: A large scale remote sensing image classification benchmark via crowdsource data,” arXiv preprint arXiv:1705.10450, 2017.
[9] P. Helber, B. Bischke, A. Dengel, and D. Borth, “Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification,” arXiv preprint arXiv:1709.00029, 2017.
[10] W. Zhou, S. Newsam, C. Li, and Z. Shao, “Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval,” ISPRS J. Photogram. Remote Sens., vol. 145, pp. 197–209, 2018.
[11] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, “Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 2, pp. 1144–1158, February 2018.

| Test Images | True Multi-Label | Inception-v2 | S-CNN-RGB | S-CNN-All |
|-------------|------------------|--------------|-----------|-----------|
| pastures, peatbogs | non-irrigated arable land, coniferous forest, mixed forest, transitional woodland/shrub | non-irrigated arable land, land occupied by agriculture, mixed forest | pastures, peatbogs |
| pastures, land occupied by agriculture, water bodies | coniferous forest, mixed forest, transitional woodland/shrub | non-irrigated arable land, land occupied by agriculture | pastures, land occupied by agriculture, water bodies |
| discontinuous urban fabric, industrial or commercial units | coniferous forest, mixed forest, transitional woodland/shrub | discontinuous urban fabric, land occupied by agriculture, broad-leaved forest, coniferous forest, mixed forest | discontinuous urban fabric, industrial or commercial units |

Table 3: Example of Sentinel-2 images with the true multi-labels and the multi-labels assigned by the Inception-v2, the S-CNN-RGB and the S-CNN-All.

| Method       | P (%) | R (%) | F1   | F2   |
|--------------|-------|-------|------|------|
| Inception-v2 [1] | 48.23 | 56.79 | 0.4988 | 0.5301 |
| S-CNN-RGB      | 65.06 | 75.57 | 0.6759 | 0.7139 |
| S-CNN-All      | 69.93 | 77.10 | 0.7098 | 0.7384 |

Table 4: Experimental results obtained by the Inception-v2, the S-CNN-RGB and the S-CNN-All.