AN ACTIVE LEARNING TOOL FOR THE GENERATION OF EARTH OBSERVATION IMAGE BENCHMARKS

Wei Yao, Octavian Dumitru, Mihai Datcu

EO Data Science, Remote Sensing Technology Institute, German Aerospace Center (DLR)
Münchener Str. 20, 82234 Weßling, Germany
wei.yao@dlr.de, corneliu.dumitru@dlr.de, mihai.datcu@dlr.de.

ABSTRACT

This paper describes an active learning tool for the generation of Earth Observation (EO) benchmark datasets. This tool is able to generate training datasets, based on its active learning strategy with a classification accuracy of around 90%. Afterwards, a data cleaning tool is needed, in order to correct noisy data and provide a clean dataset to be stored in the benchmark database, and for subsequent benchmark verification. The data cleaning procedure is supported by unsupervised learning, using clustering algorithms to group similar patterns, and dimension reduction algorithms to embed them in lower dimension with annotated labels. Moreover, interactive visualizations are implemented in most modules to help better manipulate datasets and get better understandings.

Index Terms— Benchmarks, active learning, interactive visualization, unsupervised learning, data cleaning.

1. INTRODUCTION

1.1. State of the Art

The Earth Observation (EO) community urgently needs to generate large-scale benchmark datasets, to support further machine learning or deep learning tasks. Recently, the booming development of Artificial Intelligence (AI) technologies stressed the importance of good benchmark datasets. Due to the complexity of EO data, this is usually tedious and requires considerable human efforts to prepare a proper dataset, and the available datasets are rigidly supported by pre-defined semantic labels; our studies are thus limited to some existing datasets which cover only a small part of our application needs. Typical examples are the OpenSARUrban dataset [1], the OpenSARShip dataset [2] for SAR applications (using Sentinel-1 data), and the updated BigEarthNet dataset for optical and SAR applications (using both Sentinel-1 and Sentinel-2 data) [3]. The critical questions with the benchmark datasets are: how to prepare training datasets in a fast manner, and how many semantic labels can be assigned to these datasets.

1.1.1. Active learning

Active learning is a member of the machine learning family; it is an interactive learning method which adds human interaction in the loop, and iteratively introduces user supervision to improve the classification accuracy semi-automatically. With the ranked image candidates provided by Support Vector Machine (SVM) classifiers, users are able to generate a large amount of training data with very few samples, and achieve relatively high classification accuracy as well. Previous works have already proved their effectiveness [4]-[5].

1.1.2. Data cleaning

The data cleaning strategy comes naturally after the active learning semantic annotation concept has been defined. Due to classification errors raised by active learning, although with reportedly high accuracy [6], the annotation results may be degraded by noisy data samples. With the generated large amount of annotated data, the data cleaning process then refines the raw data and purifies the resulting datasets, in order to achieve the required benchmark quality. This allows better utilization, and supports further machine learning or deep learning tasks.

1.1.3. Interactive visualization

A chief benefit of interactive visualization is that it encourages users to explore and manipulate the given data to uncover hidden relationships. Moreover, any interactive visualization adds human responses to the verification procedure. Usually, this will be done by domain experts; thus, we obtain more trustworthy and reliable benchmark datasets.

A key element in an interactive visualization approach lies in a suitable dimension reduction. Recently, a novel manifold learning method named Uniform Manifold Approximation and Projection (UMAP) has shown competitive performance with commonly used t-SNE for visualization quality, and is claimed to preserve more of the global structure with superior runtime performance [7]. Hence, we have adopted the UMAP embedding method in our data cleaning tool.

In this paper, we present an active learning-based EO benchmark tool which consists of various modules: a data selection tool, a semantic annotation tool, an interactive data cleaning tool, a benchmark database, and some benchmark verification metrics. Most components are supported by interactive visualization for users to verify existing annotations, clean noisy samples, and obtain more reliable and trustworthy datasets.

2. BUILDING BLOCKS

The benchmarking tool is composed of three groups of blocks, which are shown in Fig. 1.

In the data access step, the Copernicus data hub serves as the data provider, which allows users to query in their databases, and retrieve product metadata. Here, the data selection tool selects valid products from the previously queried information.

Then the active learning-based EO benchmark tool is the main user interface for users to retrieve large amounts of annotated data. These data will be further inspected and cleaned via the data cleaning tool, and after the cleaning check, the processed data will be stored in a database management system.

The last step involves a benchmark verification step which analyses the generated datasets using different metrics.

Moreover, compared with some state-of-the-art benchmark datasets which were mainly generated manually with a fixed range of semantic labels, this tool extends our benchmark robustness as it allows for user-defined labels, rather than just applying already existing labels to maps.

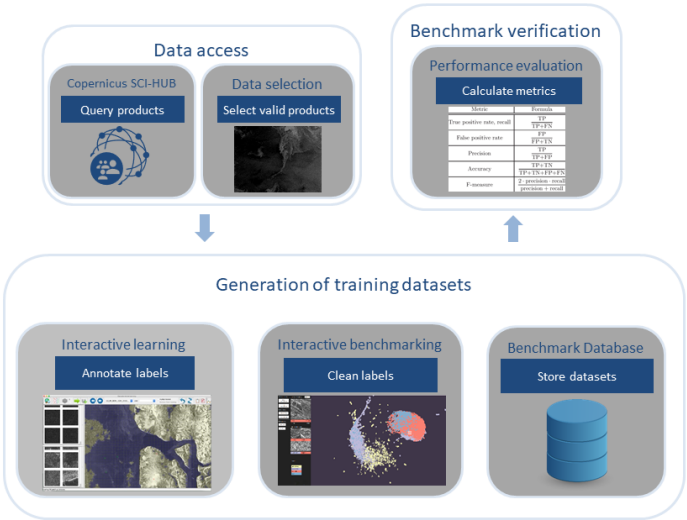

Fig. 1: Framework of the benchmarking tool.

2.1. Interactive visualization-based data selection tool

For the purpose of better selecting Sentinel products from the Copernicus Open Access Hub[1], we developed a complementary light-weight tool which visualizes quick-look images of the products. The Copernicus Open Access Hub provides bash scripts for complex and advanced queries in its database[2]. With the non-interactive script *dhusget.sh*, one can obtain the file *products-list.csv* which contains the product identifiers and the download links.

Fig. 2 shows the tool interface. It starts with loading a Sentinel product list file; then users have the option to select different grid views (with 1, 2, or 4 columns), where less columns allow a more detailed view, and more columns allow an abstract view. A selected quick-look image is bordered in blue and is ready for downloading by clicking the *download* button. During this process, an account for the Copernicus Open Hub access is required.

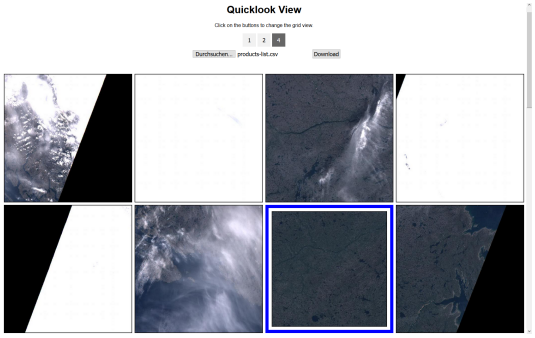

Fig. 2: Data selection tool.

Compared with the single-product quick-look view provided on the Hub, this tool provides a fast overview of a batch of Sentinel products. The visualization allows users to select valid products for subsequent processing, without being impaired by cloud-covered or incomplete content products. Different grid views add the flexibility to visualize data in an abstract-to-fine way, which thus refines the product selection. The codes will be freely available on the first author's GitHub repository[3].

2.2. Active learning-based EO image annotation tool

The logical view in Fig. 3 shows the active learning-based EO benchmark tool and its module interactions. This tool evolved from EOLib [8], where the algorithms have been upgraded for Sentinel-1 and Sentinel-2 products and contain most of the innovative functionality of [6].

The Data Model Generation adds Data Mining-specific information and descriptors (image features) to the EO products having been processed during ingestion. The Database

[1]https://scihub.copernicus.eu/dhus/#/home
[2]https://scihub.copernicus.eu/userguide/BatchScripting
[3]https://github.com/weiyao7/scihub-downloader

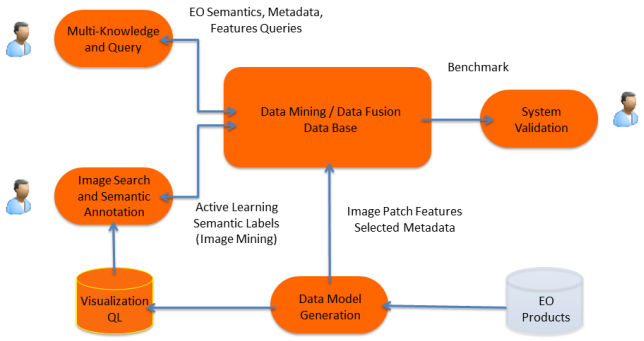

Fig. 3: Logical view of the annotation tool.

Management System provides high-speed storage and data mining functionality whose expected performance (both for processing and retrieval) requires a database-close implementation. This is the actionable information of the data mining tool.

The following components provide more functionality to the user:

- Image Search and Semantic Annotation (ISSA): image mining, query-by-example, retrieval, and adding of semantic annotation to EO image products.
- Multi-Knowledge and Query (M-KQ): multimodal queries based on selected product metadata, image features, and semantic labels.
- System Validation: supports the evaluation of the retrieval and classification performance.

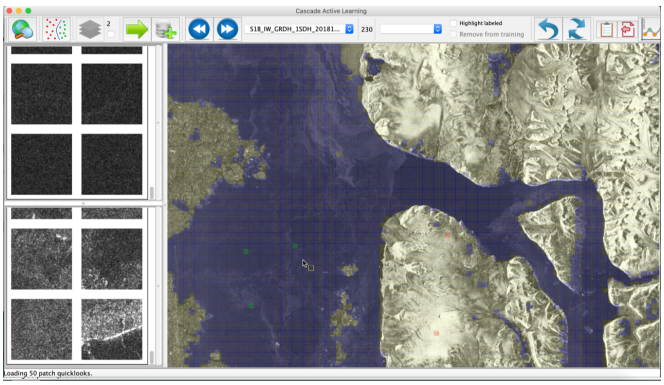

Fig. 4: Image annotation tool: Polar region sea ice example.

Fig. 4 shows an example of polar region sea ice classification, by using the tool.

2.3. Interactive visualization-based data cleaning tool

In contrast to the computer vision community, there are no well-known crowdsourcing activities for EO image annotation yet. Partly due to the complexities of EO imagery, ordinary people generally lack the ability to correctly interpret what is being contained within the imagery.

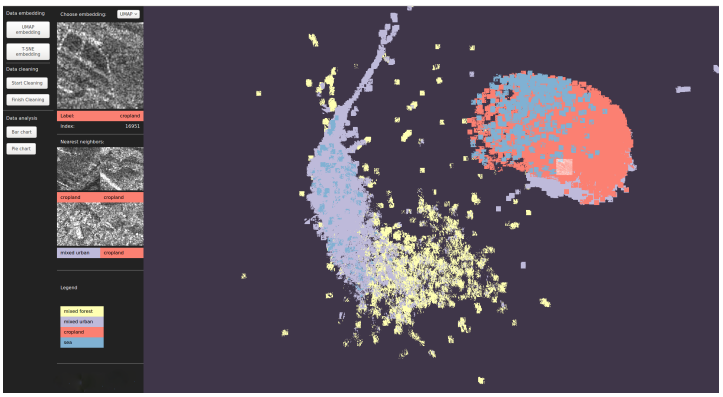

Fig. 5: Data cleaning tool: Sentinel-1 four-class example.

Hence, we developed this interactive visualization-based data cleaning tool to further verify the annotated labels, to help domain experts sweep noisy data, and recognize the same semantic labels with different structures, which should be split in the benchmark datasets.

The design of the data cleaning tool supports the following functions:

- Choose UMAP or the t-SNE embedding method to generate a suitable embedding for the annotated datasets, and generate an annotation map which shows groups of data with different annotations. The tool provides an overview of the annotation distribution, and allows users to zoom into a specific annotation group, which is shown in Fig. 5.
- Generation of a clustering map that displays discovered clusters within an annotation group, thus supporting the data cleaning step.
- Display of the nearest neighbors that allows users to visualize ranked annotated images, and perform data cleaning.

  In the general visualization mode, users can use their mouse to explore a dataset, when hovering over a small thumbnail image, the corresponding larger image will be shown in the upper left corner. When left clicking the thumbnail image, the nearest neighbors will be shown on the left as well. When the users choose the cleaning dataset mode, a right click on an individual image will pop up a context menu, from where the users have the options to assign the image to a new label. When the users finish the data cleaning, there are some options to generate some analytics such as bar charts, pie charts, etc.

Fig. 5 gives an example of Sentinel-1 image exploration. The dataset contains 30,813 non-overlapping image patches, each of which has a size of 120*120 pixels with 20 m resolution. For this, we calculated the SAR-adapted Weber features with a length of 144 dimensions. There are four classes labeled in different colors: yellow represents mixed forest; purple represents mixed urban; red represents cropland; and blue represents sea. The dataset was generated by the CANDELA project[4].

Fig. 6 shows a detailed example which extracts the outliers that lie on the right border of the embedding space in Fig. 5. As shown, all of them contain a prominent black edge border, thus discriminating them from the main thumbnail image clusters. Hence with interactive visualization, the visual inspection efficiency is increased.

The input of this tool are image features, labels, and file paths stored in JSON format; it is independent from the previous tools. Thus, if users have their own versions of images with features and labels, they can easily adapt their data in the tool to visualize data, to clean them, to discover something, etc. The demo codes will also be freely available on the first author's GitHub repository[5].

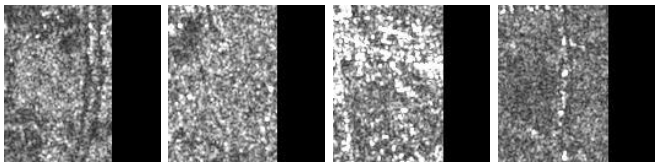

Fig. 6: Outliers shown in the data cleaning tool.

2.4. Our database model and verification metrics for benchmarking

To easily store, access and query data, we use a database to manage the cleaned datasets, and several evaluation metrics are used to measure the dataset quality. For more detailed information, see [6].

3. OUTLOOK

Together with the image annotation tool, it is foreseen that the newly developed data cleaning tool will enhance the reliability and usability of the generated datasets, thus creating trustworthy benchmark datasets, and contribute to the user community.

[4]https://candela-h2020.eu/

[5]https://github.com/weiyao7/benchmark-explorer

4. ACKNOWLEDGEMENT

This work has been supported by the EC-funded project ExtremeEarth (H2020-825258). The authors would like to express their special thanks to Gottfried Schwarz for providing valuable comments on improving the paper's readability.

5. REFERENCES

[1] J. Zhao, Z. Zhang, W. Yao, M. Datcu, H. Xiong, and W. Yu, "OpenSARUrban: A Sentinel-1 SAR image dataset for urban interpretation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 187–203, 2020.

[2] L. Huang, B. Liu, B. Li, and W. Guo, "OpenSARShip: a dataset dedicated to Sentinel-1 ship interpretation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, pp. 195–208, 2018.

[3] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A large-scale benchmark archive for remote sensing image understanding," in *Proceedings of the 2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2019)*, Yokohama, Japan, 2019.

[4] M. Datcu, H. Daschiel, A. Pelizzari, M. Quartulli, A. Galoppo, A. Colapicchioni, M. Pastori, K. Seidel, PG Marchetti, and S. D'Elia, "Information mining in remote sensing image archives: System concepts," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, pp. 2923–2936, 2003.

[5] P. Blanchart, M. Ferecatu, and M. Datcu, "Indexation of large satellite image repositories using small training sets," in *Image Information Mining: Geospatial Intelligence from Earth Observation (ESA-EUSC-JRC 2011)*, Ispra, Italy, Mar. 2011.

[6] W. Yao, C. O. Dumitru, and M. Datcu, "D2.2 data mining v2, Deliverable of the CANDELA project," `https://www.candela-h2020.eu/content/data-mining-v2`, 2020.

[7] Leland McInnes, John Healy, and James Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," `https://arxiv.org/abs/1802.03426`, 2020, arXiv, stat.ML.

[8] M. Datcu, A.-C. Grivei, D. Espinoza-Molina, C.O. Dumitru, C. Reck, V. Manilici, and G. Schwarz, "The digital Earth Observation Librarian: a data mining approach for large satellite images archives," `https://www.tandfonline.com/doi/full/10.1080/20964471.2020.1738196`, 2020, Big Earth Data, vol. 4, iss. 3.