From Historical OpenStreetMap data to customized training samples for geospatial machine learning

Zhaoyan Wu[1,2], Hao Li[1,*] and Alexander Zipf[1]

[1] GIScience Research Group, Heidelberg University, Heidelberg, Germany;
zhaoyan.wu@uni-heidelberg.de, hao.li@uni-heidelberg.de, zipf@uni-heidelberg.de
[2] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China;

* Author to whom correspondence should be addressed.
This abstract was accepted to the Academic Track of the State of the Map 2020 Online Conference after peer-review.

After more than a decade of rapid development of volunteered geographic information (VGI), VGI has already become one of the most important research topics in the GIScience community [1]. Almost in the meantime, we have witnessed the ever-fast growth of geospatial machine learning technologies to develop intelligent GIServices [2] and to address remote sensing tasks [3], for instance land use/land cover classification, object detection, and change detection. Nevertheless, the lack of abundant training samples as well as accurate semantic information has been long identified as a modelling bottleneck of such data-hungry machine learning applications. Correspondingly, OpenStreetMap (OSM) shows great potential in tackling this bottleneck challenge by providing massive and freely accessible geospatial training samples [4, 5]. More importantly, OSM has exclusive access to its full historical data [6], which could be further analyzed and employed to provide intrinsic data quality measurements of the training samples. Therefore, a flexible framework for labeling customized geospatial objects using historical OSM data allows more effective and efficient machine learning.

This work approaches the topic of labeling geospatial machine learning samples by providing a flexible framework to automatically generate customized training samples and provide intrinsic data quality measurements. In more detail, we explored the historical OSM data for two purposes: feature extraction and intrinsic assessment. For example, when training building detection convolutional neural networks (CNNs), the OSM features with tags as *building=residential* or *building=house* are certainly of interest while the data quality of such features might play an important role later in the CNNs training phase. Therefore, besides the acquisition of the user-defined OSM features, we provide additional intrinsic quality measurements. Currently, we consider some basic statistics, such as the areas of buildings tagged with different OSM tags, the number of distinct contributors in the last six months, or the equidistance of polygons with *landuse=cropland* etc., since the existing research suggested that the lower equidistance of the current polygon, the better the relative quality of the polygon, which is due to the further refining and editing from users [7]. In the

Wu, Z, Li, H., & Zipf, A. (2020). From Historical OpenStreetMap data to customized training samples for geospatial machine learning
In: Minghini, M., Coetzee, S., Juhász, L., Yeboah, G., Mooney, P., & Grinberger, A. Y. (Eds.). Proceedings of the Academic Track at the State of the Map 2020 Online Conference, July 4-5 2020. Available at https://zenodo.org/communities/sotm-2020
DOI: 10.5281/zenodo.3923040

future, one could also easily extend the current framework and develop other sophisticated quality indicators for specific “fitness-for-use” purposes.

Heterogeneous remote sensing APIs are supported within the framework. User options range from commercial satellite image providers (e.g., Bing or Mapbox) to government satellite missions (e.g., Sentinel-hub) and even user-defined tile map service (TMS) APIs. Corresponding to OSM features, the satellite image would be automatically downloaded via TMS and tiled into proper size. Moreover, this framework supports different machine learning tasks, such as classification, object detection, and semantic segmentation, which requires distinct sample formats. The preliminary test is performed to extract the geographical information of water dams with OSM tag *waterway=dam,* which enables the training of water dams detection CNNs, where users could easily change the geospatial water dams to customized objects as long as the corresponding OSM tags are identified.

The aim of this work is to promote the application of geospatial machine learning by generating and assessing OSM training samples of user-specified objects, which not only allows user to train geospatial detection models, but also introduces the intrinsic quality assessment into the “black box” of the training of machine learning models. Based on a deeper understanding of training samples quality, future efforts are needed towards more understandable and geographical aware machine learning models.

References

[1] Yan, Y., Feng, C. C., Huang, W., Fan, H., Wang, Y. C., & Zipf, A. (2020). Volunteered geographic information research in the first decade: a narrative review of selected journal articles in GIScience. *International Journal of Geographical Information Science*, 1-27.

[2] Yue, P., Baumann, P., Bugbee, K., & Jiang, L. (2015). Towards intelligent giservices. *Earth Science Informatics*, 8(3), 463-481.

[3] Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4), 8-36.

[4] Li, H., Feld, I. N., Herfort, B., & Zipf, A. (2019). Estimating OpenStreetMap Missing Built-up Areas using Pre-trained Deep Neural Networks. *Proceedings of the 22nd AGILE Conference on Geographic Information Science*, Limassol, Cyprus, 17-20.

[5] Chen, J., & Zipf, A. (2017). DeepVGI: Deep learning with volunteered geographic information. *Proceedings of the 26th International Conference on World Wide Web Companion*, Perth, Australia, 771-772.

[6] Raifer, M., Troilo, R., Kowatsch, F., Auer, M., Loos, L., Marx, S., Przybill, K., Fendrich, S., Mocnik, F.-B., & Zipf, A. (2019). OSHDB: a framework for spatio-temporal analysis of OpenStreetMap history data. *Open Geospatial Data, Software and Standards*, 4(1).

[7] Barron, C., Neis, P., & Zipf, A. (2014). A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS*, 18(6), 877-895.