AN OPENSTREETMAP-BASED DATASET OF BUILDING FOOTPRINTS FOR ANALYSING DIFFERENT TYPES OF LABEL NOISE

Jonas Gütter[1], Anna Kruspe[1,2], Xiao Xiang Zhu[2,3]

[1]German Aerospace Center (DLR), Jena, Germany
[2]Data Science in Earth Observation (SiPEO), Technical University of Munich (TUM), Germany
[3]German Aerospace Center (DLR), Oberpfaffenhofen, Germany

ABSTRACT

We present a dataset consisting of OpenStreetMap imagery and corresponding building footprint labels. Multiple label sets are provided, each containing a different type of label noise. The purpose of the dataset is to enable a systematic analysis of different label noise types in the earth observation domain and to provide a benchmark dataset for noise removal techniques. We also present some preliminary results from experiments on the effect of different label noise types on model performance.

***Index Terms*—** label noise, building footprints, dataset, Deep Learning, OpenStreetMap

1. INTRODUCTION

In the field of Earth observation, data labels for training and testing Machine Learning applications are often subject to noise. The size of the available datasets usually makes it infeasible to annotate features by hand, and so automatic methods or existing maps and databases are utilized to assign labels to the data. In most cases, these approaches are to some degree imperfect and cause the labeling process to introduce noise in the data. An example for such a database is the popular OpenStreetMap project (OSM) which mainly includes information on streets and buildings, but also other kinds of geographic information on a global scale. When relying on these types of annotations, it is important to be aware of the impact that the noise in the labels has on model performance. Currently, the impact of label noise on the performance of deep neural networks is still unclear. Most of the existing research focuses on image classification, whereas for the case of semantic segmentation, the role of label noise is not well understood. Conducting research on semantic segmentation is potentially even more challenging since each image is assigned multiple labels, so that different patterns of labels and of label noise can arise within an image. Therefore, the different kinds of label noise that can be encountered in semantic segmentation are more numerous and complex than the noise types in image classification.

Understanding the role of label noise in training data for semantic segmentation requires datasets that allow the study of different noise types independently from each other. Unfortunately, creating such datasets is difficult for the same reasons that lead to the prevalence of label noise in the first place: The lack of reliable, quality-controlled labels. To circumvent this problem, we create a dataset that is comprised not of real-world imagery, but of OSM imagery and corresponding building footprint labels [1]. By extracting the imagery as well as the labels from the same source, we are able to ensure that there is no label noise in the original labels, and subsequently create further label sets where we can accurately control the type and amount of label noise.

2. RELATED WORK

Label noise is a common problem in classification tasks that has already been extensively studied for general purpose machine learning [1]. When it comes to image classification in Deep Learning, there also exists a number of works that analyzed the role of label noise on model performance, although the findings are inconclusive. Zhang et al. showed that deep neural networks for image classification are able to memorize the training data completely, and demonstrated this by fitting a deep neural network to entirely random labels [2]. An explanation as to why this behaviour does not necessarily hamper model performance was later delivered by Arpit et al., who found that DNNs usually fit to clean labels first and only later begin to memorize noise in the data [3]. This was also confirmed by Arazo et al. [4]. Rolnick et al. showed that deep neural networks can perform very well at image classification tasks even when the training data is corrupted with huge amounts of noise [5]. In contrast to that, Wang et al. and Amid et al. reported clearly deteriorating accuracy values when introducing label noise in their datasets [6, 7].
For the field of semantic segmentation [8] however, there exists less research on the role of label noise. To our knowledge, the most extensive analysis on the impact of label noise on model performance was carried out by Zlateski et al. who

[1]https://zenodo.org/record/4446737#.YAvVEsIxk5k

created a perfectly labeled synthetic dataset of street view scenes, and subsequently produced coarsened labels of lower quality from the initial perfect labels [9]. Their findings are that model performance increases with label quality and with size of the training dataset. Our approach aims at enabling a similar analysis, and we believe that the domain change from street view to aerial images can yield additional useful insights, due to different noise patterns between those domains. Based on the work of Zlateski et al., Acuna et al. created several versions of an image dataset with varying levels of label noise and measured the performance of models for semantic segmentation of boundaries on these datasets. They reported a clearly declining performance with increasing noise levels [10].

Furthermore, despite of a number of at pixel-level annotated datasets being available [11, 12, 13, 14, 15], to our knowledge no dataset has been published yet that provides known types and amounts of label noise alongside clean labels to enable a systematic study of different label noise types.

3. GENERATING THE DATASET

The locations for the training and testing data were selected by taking the coordinates from 24,344 globally distributed cities that are either capitals or have a population greater than 15,000 according to the GeoNames database [16]. Image data from OSM is available in the form of rectangular tiles at different zoom levels and can be downloaded from the OSM tile server [17]. For each of the coordinates, one OSM tile with a width of 0.005 degrees of longitude or roughly 500 meters was downloaded from the OSM tile server. Labels were created by downloading all objects within the tile coordinates that were tagged as buildings from the Overpass API [18]. Since the data that is rendered by the OSM tile server is exactly the same as the data that is queried by the Overpass API (at least apart from changes within a few minutes), we can assume that using the former as labels for the latter will result in a largely error-free groundtruth. From the initial 24,344 images, all images that contained less than 20 buildings were excluded, resulting in 12,316 remaining images.

In addition, aerial imagery from Google Maps and corresponding OpenStreetMap building labels were downloaded for 100 European cities to provide an opportunity for comparison with real-world imagery. The following analyses in chapter 4 only focus on the OpenStreetMap imagery.

Aside from the original labels, three more label sets were created by modifying the original ones, each containing a different type of noise. The modification for the first of these label sets consisted of flipping the value of a certain percentage of pixels in the original image, thereby inserting random noise. This type of noise is not very likely to occur in real-world data, but can serve as a baseline to compare with other noise types. The second modification was the removal of 10% of the existing objects from the original dataset, referring to a common issue when OSM or other sources of Volunteered Geographic Information (VGI) are used for label generation that do not include the complete set of labels. The last label set was modified by adding rectangular shapes in arbitrary locations, simulating the case that buildings are included in the labels that do not appear in the actual image. This type of errors is also rather uncommon, but could for example appear when an outdated map is used for label generation that includes buildings that were torn down in the meantime. To make the amount of noise between those label sets comparable, it was attempted to make the total number of noisy pixels in each image the same over all three label sets. While this goal was achieved for the first and the second label set, the third one with the added objects can have a lesser number of noisy pixels if the locations of the added buildings coincide with the locations of already existing buildings. Examples for clean and noisy labels as well as an OSM tile are shown in figure 1.

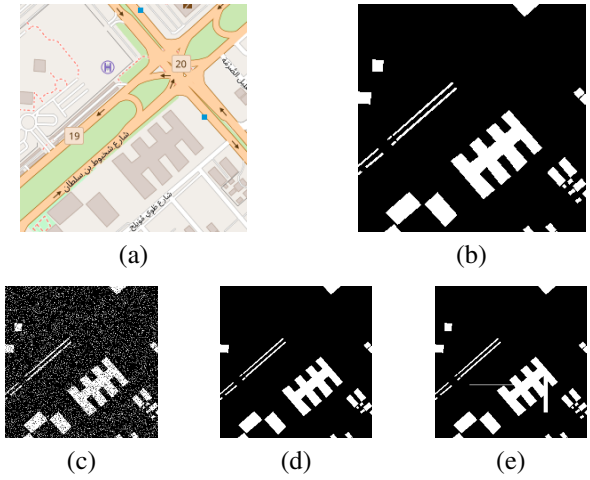

Fig. 1. Example of one sample in the generated dataset; (a): OSM imagery; (b): clean labels; (c): label with random noise; (d): label with deleted buildings; (e): label with additional buildings

4. FIRST ANALYSIS ON THE DATASET

To get an insight if and how different types of label noise affect the performance of a DNN in our setting, we trained a DeepLabV3+ model for semantic segmentation [19] on the dataset. Since previous works suggest that noise increases the required amount of training data [9], we also trained the model on subsets of different sizes of our dataset. Those subsets were further split into 80% training and 20% validation data. Validation was always performed with respect to the clean labels, whereas training was conducted on each of the label sets separately. We chose the binary crossentropy as a loss function and used a batch size of 8 and the Adam opti-

mizer with decaying learning rate for training. Furthermore, dropout was performed before every convolutional layer to avoid overfitting. All experiments were repeated 10 times.

Figure 2 shows the development of the mean accuracies of those 10 repetitions over the training epochs for all the label sets on the highest training size of 9600 images. As expected, the best accuracy is always achieved on the clean dataset. Interestingly, the noisy labels behave very differently: The label set with removed buildings behaves most similar to the clean label set in that it displays a saturating curve and shows very low fluctuations between different runs. However, the curve saturates faster than the one for the clean labels so that the accuracy gap between clean and noisy labels increases with the epochs. In the end, it scores about half a percentage point worse than the clean label set. In contrast to that, the label set with the added buildings reaches its peak accuracy relatively early at epoch 20 and falls down after that again. In the following epochs, it also shows relatively high fluctuations between the different repetitions. This is intuitively surprising, since the labels sets with removed and added buildings are semantically quite similar: Both contain objects in the training images that are not in the labels and vice versa, so the big difference in the training behavior between those label sets is unexpected for us. A reason for this could lie in the imbalanced class distribution of the dataset: Only about 20% of all pixels in the clean training data are labeled as buildings, so adding false positives to a class that already has few samples might have a bigger impact than doing the same to a class with lots of samples. At last, the random noise label set shows a very similar development as the clean label set, except it scores always between a half and one percentage point worse and shows relatively high standard deviations that become smaller in the later epochs.

Figure 3 shows the mean accuracies after 100 epochs of training with different training sizes for all the label sets. For all of the noisy label sets, the increase of the training size from 2000 to 4000 samples seems crucial for getting reasonable results, after that the improvements in accuracy are only marginal. This observation is in line with other works about the impact of sample size on model performance with noisy labels [5]. Furthermore, the fluctuations between different runs are quite high when training with the smallest training size of 2000 samples. In particular, the standard deviation of the label set with removed buildings is extremely high at this sample size, although at all other sample sizes the standard deviation is almost zero. The unusually high standard deviation stems from a few outliers within the 10 runs that produced high training accuracies, yet very low validation accuracies. Our first guess on explaining this behaviour is that in those cases the training data was merely memorized and no generalization capability beyond the training set was reached.

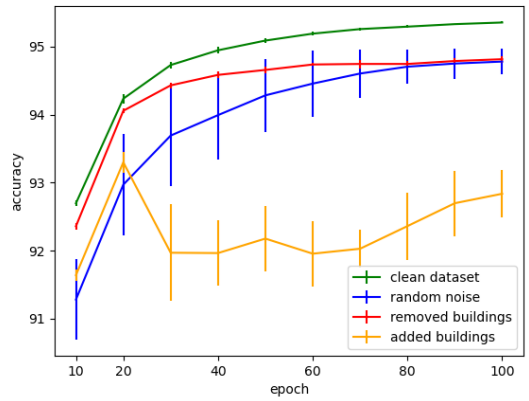

Fig. 2. Development of accuracies over training epochs for the different label sets, trained on 9600 images. Vertical lines show the standard deviation from 10 repetitions.

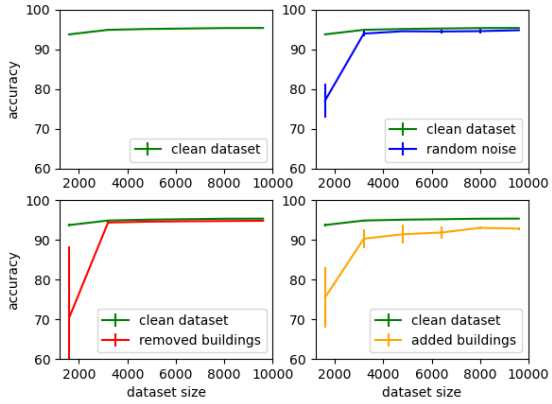

Fig. 3. Accuracy development of the different label sets depending on the size of the training data, after 100 epochs of training. Vertical lines show the standard deviation from 10 repetitions.

5. CONCLUSION AND OUTLOOK

Our dataset can serve as a starting point for understanding the role of label noise in earth observation as well as providing a ground truth for evaluating techniques for detecting and removing label noise. Our short analysis already revealed strong differences in the training behavior on different types of noisy labels, suggesting that more insights can be gained by developing this work further. Future works could include the creation of similar datasets for other geospatial features like roads, and the addition of further noise types that are common in remote sensing, for example when labels are slightly rotated or shifted away from their original position.

The fact that the dataset only consists of artificial map data is

a serious limitation, since observed results cannot automatically assumed to be transferable to real-world imagery. However, the similarity of real-world imagery to OpenStreetMap is probably far bigger than to e.g. medical imagery or any other of the the popular computer vision domains where research about the role of label noise has already been done to a greater extent. This makes our dataset a valuable resource on the middle ground between domain similarity to earth observation and practical usability when it comes to comprehending the role of label noise in remote sensing.

6. REFERENCES

[1] Benoît Frénay and Michel Verleysen, "Classification in the presence of label noise: a survey," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 5, pp. 845–869, 2014.

[2] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[3] Devansh Arpit, Stanislaw Jastrzkebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al., "A closer look at memorization in deep networks," *arXiv preprint arXiv:1706.05394*, 2017.

[4] Eric Arazo Sanchez, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness, "Unsupervised label noise modeling and loss correction," 2019.

[5] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit, "Deep learning is robust to massive label noise," *arXiv preprint arXiv:1705.10694*, 2018.

[6] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy, "The devil of face recognition is in the noise," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 765–780.

[7] Ehsan Amid, Manfred KK Warmuth, Rohan Anil, and Tomer Koren, "Robust bi-tempered logistic loss based on bregman divergences," in *Advances in Neural Information Processing Systems*, 2019, pp. 15013–15022.

[8] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew, "A review of semantic segmentation using deep neural networks," *International journal of multimedia information retrieval*, vol. 7, no. 2, pp. 87–93, 2018.

[9] Aleksandar Zlateski, Ronnachai Jaroensri, Prafull Sharma, and Frédo Durand, "On the importance of label quality for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1479–1487.

[10] David Acuna, Amlan Kar, and Sanja Fidler, "Devil is in the edges: Learning semantic boundaries from noisy annotations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11075–11083.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset," in *CVPR Workshop on the Future of Datasets in Vision*, 2015, vol. 2.

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[14] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan, "Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?," *arXiv preprint arXiv:1610.01983*, 2016.

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[16] Geonames geographical database, "cities15000," `http://www.geonames.org/`, 2019.

[17] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," `https://a.tile.openstreetmap.org/`, 2020.

[18] "Overpass api," `http://overpass-api.de/`, 2020.

[19] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.