ANALYSING THE INTERACTIONS BETWEEN TRAINING DATASET SIZE, LABEL NOISE AND MODEL PERFORMANCE IN REMOTE SENSING DATA

Jonas Gütter[1], *Julia Niebling*[1], *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

In this work we analyse how training datasize affects the ability of a deep neural network to deal with noisy training labels in a semantic segmentation task with labels from OpenStreetMap. To this end, several versions of the training set were created by introducing varying amounts of label noise, and a model was then trained on subsets of varying size of these versions. The results indicate that the relationship between noise level and model performance is largely independent of the datasize except for very small datasizes where adding label noise has an even more deteriorating effect than usual.

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

1. INTRODUCTION

Deep Neural Networks (DNNs) have become state-of-the-art for the segmentation of remote sensing imagery. Despite the success of DNNs, one of the major challenges that can often hinder their full potential is the need for large amounts of labeled training data. A common way of obtaining such labels in remote sensing is the usage of crowdsourced datasets where the labels are provided mainly by volunteers, as for example in the popular OpenStreetMap project [1]. Labels obtained from crowdsourcing can be subject to noise due to subjectiveness in assessment, incomplete coverage of an area or errors out of carelessness and oversight. Estimating the impact that such noise has on the quality of predictions made by models trained on this data is important to assess whether crowdsourcing is a valid alternative to expert labeling. However doing such estimations is not trivial, since the interaction between noise in training labels and the quality of subsequent predictions is not well researched yet. One of the factors that possibly affects this interaction is the size of the training dataset. Previous studies found that the minimum size of the training dataset required for effective training of DNNs increases with the noise level in the data labels. In this work we aim to validate these findings in the remote sensing domain, thereby answering the question of how training datasize affects the capability of a model to deal with label noise in semantic segmentation. For this, we train a DNN on datasets of varying noise levels and datasizes and compare the predictive performances of the resulting models.

2. RELATED WORK

The role of label noise in Deep Learning has been studied numerous times. One of the first important findings in this regard was stated by Zhang et al. [2], who showed that DNNs are able to memorize noise completely, making the danger of models overfitting to noise seem imminent. It was later shown by Arpit et al. [3] as well as by Arazo et al. [4] that DNNs usually learn clean labels first during training and only later overfit on noisy samples, which alleviates the problem of having noisy labels to some extent. Rolnick et al. [5] found that DNNs can be very robust against label noise. They also analysed the influence of the training dataset size and came to the conclusion that the minimum dataset size required for effective training increases with the noise level in the data. Similar findings were also stated by Wang et al. [6]. To our knowledge, Rolnick et al. and Wang et al. are the only ones who analysed the impact of dataset size on the role of label noise in Deep Learning so far.

Looking at the task of semantic segmentation, there are only a few works that examine the influence of label noise: Zlateski et al. [7] analysed the relationship between time spent on annotating a dataset and predictive performance of a model trained on this dataset, while a study on the influence of label noise in medical image segmentation was performed by Vorontsov et al. [8]. They found that the type of errors makes a big difference for its impact on model performance to the extent that biased errors have far more impact than unbiased ones.

In the field of remote sensing, Mnih et al. [9] were among the first who pointed out that using existing geographic information as training labels can introduce unwanted noise. Other works at the intersection of remote sensing and semantic segmentation include Rahaman et al. [10] who analysed the role of label noise in water body segmentation and found that increasing the noise level can result in massive drops in accuracy, Li et al. [11] who studied the impact of label noise on different classifiers for hyperspectral images, as well as Henry et al. [12] who introduced omission noise and registration noise into road segmentation datasets. Their results suggest that small amounts of noise can even increase predictive performance when a suitable loss function is used.

3. DATA AND MODEL

For our experiments we are using data from OpenStreetMap [1], a freely available geographic database. Building geometries for roughly 10 000 images of 256x256 pixels were downloaded from OpenStreetMap and converted into binary label masks, so that each pixel is either classified as building or as background. Imagery was as well taken from OpenStreetMap. That means that we do not use real-world aerial imagery, but a cartographic view. The reason for using such a dataset is that - since the images are merely the labels rendered among other objects - the label quality is extremely high. It can be safely assumed that the only noise existent in those labels is the one that we introduce on purpose for our experiments. For further details on the generation of the dataset we refer to our previous work [13].

We focus here on the noise type of omission noise, that means objects which are present in the imagery are not represented in the labels. We create several noisy versions of our dataset by removing random objects from the initial clean labels. This way, 10 versions of noisy label sets were created between 10% and 100% label noise. Here, the percentage refers to the area of buildings, meaning that in the dataset with 10% omission noise, approximately 10% of the original building area in each label mask was removed. The labels with 100 % omission noise consequently do not contain any buildings and consist only of the background class. Examples are illustrated in Figure 1.

We train a DeepLabV3+ segmentation network [14] for 30 epochs on subsets of different sizes of all versions of the training dataset and repeat each run 10 times to obtain mean values and standard deviations of our metrics. The number of epochs was chosen after observing that on our clean dataset the accuracy saturates after 30 epochs. The metrics are always calculated on a clean test set.

4. RESULTS

Table 1 shows the pixelwise accuracy as well as the recall, precision and IoU for the building class that were achieved by our experiments. In the following, we focus mainly on the IoU for reasons of brevity and because most of the other metrics display a similar behaviour. Figure 2 shows how the IoU changes with the noise level for different sizes of the training dataset. Not surprisingly, bigger dataset sizes consistently outperform smaller ones. Especially the dataset size of 100 samples shows a drastically worse performance than all the other dataset sizes. The differences in performance between dataset sizes become gradually smaller with increasing noise levels since at a noise level of 100% no learning is possible anymore and so each network necessarily achieves the same performance there, predicting always the background class.

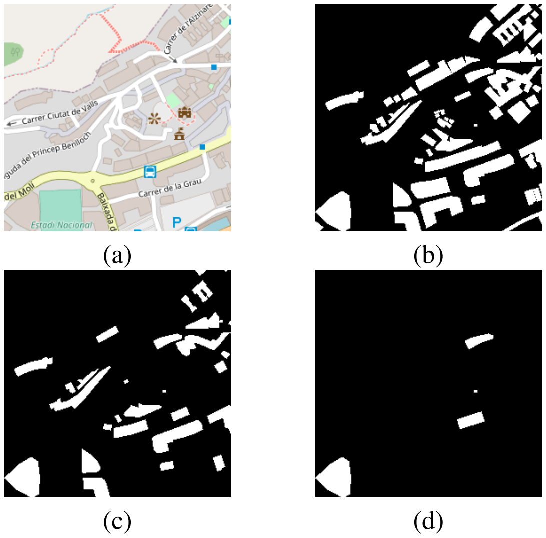

Fig. 1. Training data sample used in the experiments. (a): cartographic imagery from OpenStreetMap. (b): clean labels from OpenStreetMap. (c): labels with 50% omission noise. (d): labels with 90% omission noise.

Comparing the absolute changes in performance between the different dataset sizes can be misleading since lower sizes also have a lower performance right from the start and so just looking at the slope of the performance curve does not necessarily allow to draw conclusions about the impact of the dataset size. For this reason, we also show the development of the IoU metrics relative to their starting values in Figure 3. It can be seen that the development of the relative IoU is similar for each dataset size except for the smallest one with 100 samples. The metric drops a lot faster and reaches its minimum far sooner than for all the other dataset sizes.

Furthermore, to compare our results with the findings from Rolnick et al. [5], Figure 4 shows the pixelwise accuracy plotted against the training datasize for the different noise levels. A stepwise curve as observed by Rolnick et al. cannot be seen here, instead the accuracy increases more gradually with the logarithm of the dataset size.

5. DISCUSSION

The results shown in Figure 3 show that the dataset size does not affect the ability of the model to deal with label noise, except for the smallest datasize of 100 samples, where the predictive performance drops much faster with increasing noise level. This could mean that the vulnerability towards label

		metric	noise level										
			0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
training data size	100	Accuracy	0.837	0.827	0.821	0.815	0.813	0.812	0.811	0.811	0.811	0.811	0.811
		Recall	0.195	0.115	0.069	0.030	0.006	0.004	0.001	0.001	0	0	0
		Precision	0.817	0.820	0.787	0.790	0.776	0.755	0.659	0.213	0	0	0
		IoU	0.185	0.112	0.066	0.028	0.006	0.004	0.001	0	0	0	0
	500	Accuracy	0.905	0.893	0.887	0.873	0.862	0.842	0.832	0.819	0.814	0.812	0.811
		Recall	0.726	0.694	0.608	0.475	0.372	0.251	0.148	0.056	0.019	0.006	0
		Precision	0.822	0.802	0.829	0.850	0.846	0.847	0.847	0.861	0.840	0.760	0.073
		IoU	0.625	0.588	0.534	0.437	0.348	0.238	0.143	0.056	0.018	0.005	0
	1000	Accuracy	0.919	0.911	0.901	0.885	0.881	0.857	0.840	0.828	0.818	0.812	0.811
		Recall	0.844	0.759	0.650	0.528	0.475	0.291	0.196	0.103	0.050	0.006	0
		Precision	0.810	0.835	0.850	0.864	0.850	0.865	0.856	0.848	0.876	0.847	0
		IoU	0.703	0.657	0.581	0.486	0.438	0.278	0.189	0.101	0.050	0.006	0
	5000	Accuracy	0.946	0.939	0.928	0.911	0.889	0.863	0.843	0.828	0.818	0.813	0.811
		Recall	0.885	0.820	0.741	0.626	0.478	0.312	0.168	0.092	0.040	0.010	0
		Precision	0.874	0.897	0.906	0.909	0.916	0.918	0.924	0.919	0.903	0.885	0.071
		IoU	0.785	0.749	0.688	0.589	0.457	0.303	0.166	0.091	0.040	0.010	0
	10000	Accuracy	0.955	0.948	0.937	0.916	0.893	0.869	0.849	0.834	0.821	0.814	0.811
		Recall	0.866	0.840	0.783	0.657	0.501	0.305	0.209	0.127	0.052	0.012	0
		Precision	0.925	0.919	0.922	0.935	0.942	0.947	0.951	0.934	0.926	0.932	0.125
		IoU	0.809	0.782	0.734	0.628	0.486	0.299	0.206	0.126	0.052	0.012	0

Table 1. Selected metrics for the different setups after 30 epochs of training. Metrics are mean values of 10 training runs and were calculated on the same clean test set. Precision, recall and IoU were calculated with respect to the building class.

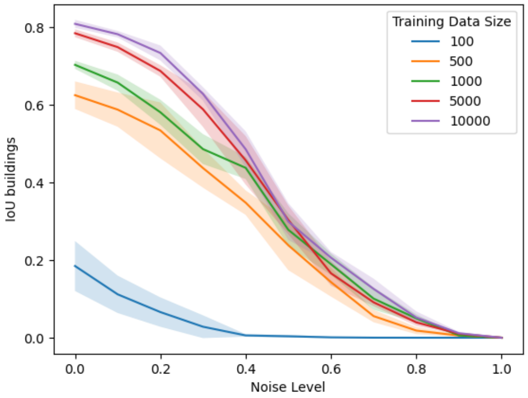

Fig. 2. Mean IoU values of 10 training runs for different sizes and noise levels of the training dataset. The IoU metric was calculated on a clean test set.

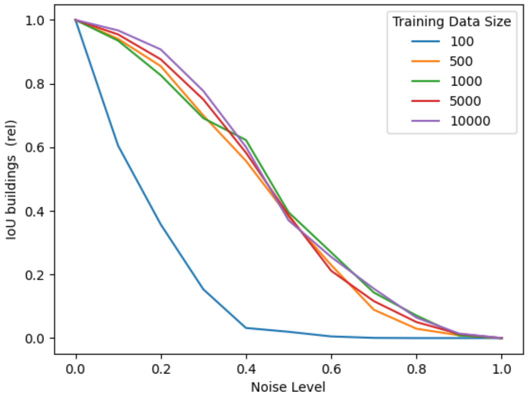

Fig. 3. Mean IoU values of 10 training runs for different sizes and noise levels of the training dataset. The IoU metric was calculated on a clean test set. Values are normalized with respect to the respective IoU on clean labels

noise is increased for very small datasizes. On the other hand, at such a small datasize, bad results could also just occur by chance, when the dataset contains only samples that are hard to generalize from.

Previous work from Rolnick et et al. [5] reported a clear treshold datasize that is necessary for successful learning. Our results don't show such a clear treshold (see Figure 4). However we did see in Figure 3 that there is a big difference in performance between the smallest dataset size and the all other ones, so the existence of a minimum reasonable dataset size still seems plausible with our observations.

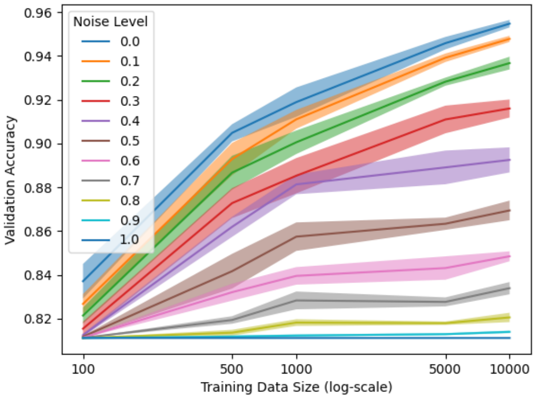

Fig. 4. Pixelwise accuracy plotted against log training datasize. Colored areas denote standard deviations from 10 training runs.

6. CONCLUSION

Our results indicate that dataset size only plays a role for a model's ability to handle label noise when the dataset size is very small. In general, small dataset sizes should be avoided anyway. With our findings in mind, this is even more the case for noisy datasets. Here, we only looked at the narrow case of a specific noise type and non-real-world data. To evaluate the role of dataset size in a more general way, future studies on other noise types and real-world imagery are of importance, as well as a closer look on other use cases in remote sensing, like for example road detection.

7. REFERENCES

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

[2] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, "Understanding deep learning requires rethinking generalization," *International COnference on Learning Representations*, 2017.

[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," *International Conference on Machine Learning*, 2017.

[4] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness, "Unsupervised label noise modeling and loss correction," in *International Conference on Machine Learning*. PMLR, 2019, pp. 312–321.

[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] 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.

[8] Eugene Vorontsov and Samuel Kadoury, "Label noise in segmentation networks: mitigation must deal with bias," in *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections*, pp. 251–258. Springer, 2021.

[9] Volodymyr Mnih and Geoffrey E Hinton, "Learning to label aerial images from noisy data," in *Proceedings of the 29th International conference on machine learning (ICML-12)*, 2012, pp. 567–574.

[10] Mustafizur Rahaman, Md Monsur Hillas, Jannatul Tuba, Jannatul Ferdous Ruma, Nahian Ahmed, and Rashedur M Rahman, "Effects of label noise on performance of remote sensing and deep learning-based water body segmentation models," *Cybernetics and Systems*, pp. 1–26, 2021.

[11] Meizhu Li, Shaoguang Huang, and Aleksandra Pizurica, "A study on the label noise impact on the hyperspectral image classification," *IEICE Proceedings Series*, vol. 64, no. ICTF2020_paper_25, 2021.

[12] Corentin Henry, Friedrich Fraundorfer, and Eleonora Vig, "Aerial road segmentation in the presence of topological label noise," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 2336–2343.

[13] Jonas Gütter, Anna Kruspe, and Xiao Xiang Zhu, "An openstreetmap-based dataset of building footprints for analysing different types of label noise," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 2321–2324.

[14] 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.