UNCERTAINTY-GUIDED REPRESENTATION LEARNING IN LOCAL CLIMATE ZONE CLASSIFICATION

Christoph Koller [1,2], *Muhammad Shahzad* [1], *and Xiao Xiang Zhu* [1,2]

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

ABSTRACT

A significant leap forward in the performance of remote sensing models can be attributed to recent advances in machine and deep learning. Large data sets particularly benefit from deep learning models, which often comprise millions of parameters. On which part of the data a machine learner focuses on during learning, however, remains an open research question. With the aid of a notion of label uncertainty, we try to address this question in local climate zone (LCZ) classification. Using a deep network as a feature extractor, we identify data samples that are seemingly easy or hard to classify for the model and base our experiments on the relatively more uncertain samples. For training of the network, we make use of distributional (probabilistic) labels to incorporate the voter confusion directly into the training process. The effectiveness of the proposed uncertainty-guided representation learning is shown in context of active learning framework where we show that adding more certain data to the training pool increases model performance even with the limited data.

Index Terms— Local Climate Zones (LCZ), Classification, Uncertainty Quantification, Representation Learning, Urban Land Cover

1. INTRODUCTION

Significant performance improvements in remote sensing models result from computational advances associated with models of unprecedented capacity with an ever-growing number of model parameters. In a supervised learning setting, this combination flourishes especially when the data set being modeled is large as well. Yet what the model usually does not tell us is which part of the labelled data is particularly important in boosting the performance. As part of this work, we look at this issue from an uncertainty quantification perspective. We claim that when a notion of label uncertainty for each data point is available (or extracted from the data itself), the model would benefit more when learning from data points with a higher uncertainty value.

To validate the above hypothesis, we analyzed the feature representations learned by a deep neural network within the task of classifying satellite images into local climate zones. A notion of uncertainty is established on the grounds of *distance* (or *similarity*) measures to different class-wise focal points. The smaller the distance to the focal point in the high-dimensional feature representation space, the lower the induced measure of uncertainty. The aspect of human uncertainty in the labeling process is considered by directly embedding the voter confusion between different classes into the class labels used for training. The effectiveness of the proposed uncertainty quantification procedure is validated both visually and by implementing it into an active learning framework.

2. RELATED WORK

Local Climate Zone Classification describes the task of classifying the scene on the Earth's surface into a predefined scheme of climate zones – a popular scheme presented in Stewart et al. [1] that consists of 17 classes of which 10 are urban-related classes and 7 are non-urban-related. Early works of applying this scheme focused on e.g. Urban Heat Islands (UHIs) [2] or urban planning [3]. Recently, Zhu et al. [4] introduced a large-scale benchmark data set to the community, consisting of more than 400,000 Sentinel-1 and Sentinel-2 image patch pairs that were manually labeled in a labor-intensive process. Various works have since then focused on e.g. benchmarking the data set with various neural network architectures [5] or combining the labels of the data set with multi-seasonal Sentinel-2 data [6].

Uncertainty Quantification (UQ) in Deep Learning is a relatively new field in the Earth Observation community which helps to shine some light on the often termed *Black-Box-models* used in machine and deep learning by quantifying possible data and model uncertainties. A general research direction in this domain includes using ensembles of models to receive multiple predictions and derive uncertainties in terms of deviations [7], which has been successfully applied in the remote sensing area [8] [9]. Furthermore, models based on Bayesian reasoning have been established and UQ based on this framework has been implemented recently with the aid of dropout networks [10]. The method is widely used, although the application to remote sensing data is yet limited [11].

Representation Learning focuses on learning intermediate features or representations of data within a modeling pipeline. These features ideally embody rich information about the underlying data and can help downstream models to better predict on it. In [12], the authors explored various remote sensing data sets in order to find good representations for data of this domain. On the other hand, the method was also used by [13] to find representations which are domain-invariant, hence overcome the domain shift typically occurring between the training and test data. The list of possible downstream tasks is extensive and includes for example scene classification [14] or change detection [15].

3. METHODOLOGY

3.1. Data

The data basis of the following analyses is formed by the So2Sat LCZ42 data set [4]. The included image patches come from 42 cities as well as 10 add-on regions all over the world. Each of the patches is of size 32 × 32 pixels, corresponding to an area of 320m × 320m. 18 different channels are available, including 8 Sentinel-1 channels as well as 10 Sentinel-2 channels of which 4 have a ground sampling distance (GSD) of 10m. The other 6 channels were upsampled from a GSD of 20m to match the other channels. For this work, only Sentinel-2 data is considered. For a subset of 10 European cities as well as 9 add-on regions within the LCZ42 cities, the labeling process was further evaluated by asking 10 remote sensing experts to blindly relabel the data independently. This data consists of ca. 250,000 image patches, again including Sentinel-1 and Sentinel-2 bands. For the corresponding label, there exist both a ground truth formed by the majority vote of the expert votes (the original LCZ42 label is additionally considered when there exists a tie between two classes) and a label distribution across all classes formed by the empirical distribution of the expert votes.

3.2. Learning Distributional Labels

Let us denote the climate zones classification data by $\{x^{(i)}, y^{(i)}, y^{(i)}_{\text{true}}\}_{i=1,\ldots,n} \in (\mathcal{X} \times \mathcal{Y} \times \mathcal{K})^n$, where $x^{(1)}, \ldots, x^{(n)} \in \mathcal{X}$ are the LCZ42 image patches comprised of 10 Sentinel-2 bands and $y^{(1)}, \ldots, y^{(n)} \in \mathcal{Y}$ are the label distributions formed by the expert votes. Furthermore, we denote $y^{(i)}_{\text{true}} \in \mathcal{K}$ as the single one-hot encoded ground truth label of image i formed by the majority vote of the expert votes. For the distributions, we follow the idea and notation firstly introduced in [16]. More concretely, for image i define the distributional label via

$$\begin{aligned} y^{(i)} &= \left(y^{(i)}_1, \ldots, y^{(i)}_K\right), y^{(i)}_j \in [0,1] \\ \text{s.t.} \sum_j y^{(i)}_j &= 1 \;\forall\; i = 1, \ldots, n \end{aligned}$$

where $K = 17$ denotes the number of distinct classes (i.e. local climate zones) and $y^{(i)}_j$ expresses the degree of which image i is attributable to class j. For the given expert votes $V^{(i)}_1, \ldots, V^{(i)}_J$, $V^{(i)}_j \in \{1, \ldots, K\} \;\forall i = 1, \ldots, n$ from experts $j = 1, \ldots, J$ regarding image i, we have $\boldsymbol{V}^{(i)}_j = (\mathbb{1}_{\{V^{(i)}_j = 1\}}, \ldots, \mathbb{1}_{\{V^{(i)}_j = K\}})$. This then allows to infer the vote vectors via

$$\boldsymbol{Y}^{(i)} = \left(Y^{(i)}_1, \ldots, Y^{(i)}_K\right), \; Y^{(i)}_k = \sum_j \mathbb{1}_{\{V^{(i)}_j = k\}}$$

Here, $Y^{(i)}_k = m$ means that class k received m votes for image i and it holds that $\sum_{k=1}^K Y^{(i)}_k = M$, where $M = 10$ represents the number of experts. We can now redefine the majority vote via taking the maximum over $\boldsymbol{Y}^{(i)}$ for image i and encoding it in a one-hot manner. Eventually, we combine the $J = 10$ expert votes to form the aforementioned distributional label via the empirical distribution over the different classes. For image i, this leads to

$$y^{(i)} = \frac{1}{M} \sum_j Y^{(i)}_j \tag{1}$$

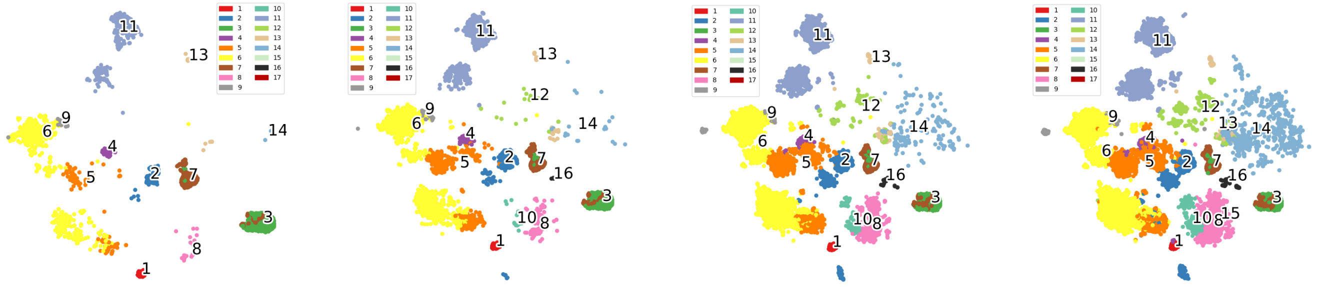
Let further be $f_\theta(x) := g_{\theta_1} \circ h_{\theta_2}(x)$ a neural network classifier which yields the *logits* (unnormalized class estimates) for the input $x \in \mathcal{X}$, dependant on the parameter sets θ_1 and θ_2. Here $h_{\theta_2}(x)$ describes the deep inherent feature representation of the network given input $x \in \mathcal{X}$ (namely the output of the penultimate layer), and $f_\theta(x)$ describes the mapping of this representation into the label space, therefore $g_{\theta_1} : \mathbb{R}^L \to \mathbb{R}^K$, where L is the dimension of the learned deep representation. Training the network is performed in the usual manner by backpropagating the loss through the network. We use the widely-known *cross-entropy loss*, however now considering the distributional nature of the labels. For a batch of data $\{x^{(i)}, y^{(i)}\}_{i=1,\ldots,m}$ the loss is then computed via

$$\begin{aligned} &L_{CE}(f_\theta, x^{(1)}, \ldots, x^{(m)}, y^{(i)}, \ldots, y^{(m)}, \theta_1, \theta_2) \\ &= -\frac{1}{m} \sum_{i=1}^m \sum_{k=1}^K y^{(i)}_k \cdot \log p(y^{(i)} | x^{(i)}, \theta_1, \theta_2)_k \end{aligned}$$

where the predictive distribution $p(y^{(i)} | x^{(i)}, \theta_1, \theta_2)$ is formed via the softmax activation as usual.

3.3. Distance Measures for Deep Representations

During training, not only the prediction $p(y|x, \theta_1, \theta_2)$ but implicitly also the deep representation $h_{\theta_2}(x)$ get optimized with regard to the loss. The latter yet is of a lot higher dimension and allows thus for better separation of the different classes. This motivates the idea of this work to use these representations to filter the data in terms of model certainty. To



(a) 5% most certain samples (b) 10% most certain samples (c) 20% most certain samples (d) 30% most certain samples

Fig. 1: TSNE visualization of the 2048d deep features learned by a ResNet50. The model was pretrained on ImageNet and finetuned on LCZ42 with distributional labels. The distances of the data points to their respective geometric class medians were sorted and only the subset samples are shown in the TSNE plots.

do so, let us first define the set of all points having the same ground truth label, to be

$$\mathcal{S}(X)_k := \Big\{\{x^{(j)}\}_{j=1}^{m} : \{x^{(i)}, y^{(i)}, y_{\text{true}}^{(i)}\}_{i=1}^{n} \in (\mathcal{X}\times\mathcal{Y}\times\mathcal{K})^n \wedge y_{\text{true}}^{(i)} = k\Big\} \text{ for } k\in\mathcal{K},\ X\subset\mathcal{X}$$

Then, using the aggregation function $\Psi(\cdot)$ we can form class-specific focal points, which are derived based on the inputs $x^{(j)}$ belonging to a certain class k. More concretely, we aim to find the class-wise centers of mass or the geometric medians of the deep feature representations. This leads us to

$$\Psi_k(X) = \Psi\big(h_{\theta_2}(\mathcal{S}(X)_k)\big)$$

where $\Psi(\cdot)$ is now taken to be the geometric median, which is defined via

$$\underset{c_k\in\mathbb{R}^K}{\arg\min}\sum_{j=1}^{m_k} ||s_k^{(j)} - c_k||_2 \text{ where } s_k^{(j)} \in \mathcal{S}(X)_k \tag{2}$$

Here, m_k denotes the number of samples with ground truth class k. From this center of mass, we can treat the distance to the individual points as notion of uncertainty. Hence, points further away from the center are equipped with a higher quantity of uncertainty based on their deep feature representations learned by the neural network.

4. EXPERIMENTAL RESULTS & VALIDATION

To validate our approach, we employed ResNet50 architecture as feature extractor, which was pretrained on ImageNet and later fine-tuned on the LCZ42 data until convergence, using the distributional labels defined via (1) to account for the label uncertainty. A geographically separated validation set is used; the test set is a subset of the validation set and is held out. Note that we excluded the LCZ 17 (water), because it accounts for more than 40% of the data, hence leads to a large class imbalance, and contains no label uncertainty, which is unsuitable for the present modeling technique. After training, we extracted the learned deep features $h_{\theta_2}(x_j)$, $x_j \in \mathcal{X}_{\text{train}}$ and computed the geometric class medians according to (2) using the Weiszfeld algorithm.

Once determined, the euclidean distances between the feature representations of all training points and the respective derived geometric class medians are calculated. These distances are said to reflect the uncertainty of the model when classifying the image samples into the predefined LCZ classes. To validate the usefulness of these distances, we refer to Figure 1 which shows exemplary t-distributed stochastic neighbor embedding (TSNE) visualizations with different thresholds for the distances of the individual points to their respective geometric class medians. Though some classes are over-represented when setting a global threshold for the distance to the class median, we can see that the class clusters grow more or less equally. We use this idea and frame the whole problem in the context of active learning where we use the identical ResNet50 architecture, again pre-trained on ImageNet, with different subsets of the training data selected based on their distance to the respective class medians. The subsetting is considered in two different ways: Both a global threshold for the distances, as well as in stratified manner, the latter assures that each class is represented equally. The models are hence relying on the most certain or uncertain data points during training, where the notion of uncertainty is defined via the distances in the high-dimensional feature representation space. All models are evaluated on the identical hold-out test set, and the resulting performance metrics are presented and compared against a random subset of the data matching the respective size of the subset in Table 1.

A straightforward message is delivered by the models trained on a random (R) subset of the initial training data, as they already achieve performance metrics close to the maximum achievable value (using the same architecture, an overall accuracy of ∼71 % was achieved when using the entire training data set). This indicates a large present redundancy in the data set. The other extreme is visible when looking at the uncertainty threshold models, which perform comparably

	OA	AA	κ	CE
R / UT / UTS (5 %)	**66.2** / 60.1 / 58.5	**31.3** / 20.9 / 23.6	**56.7** / 46.1 / 47.9	**1.39** / 1.61 / 1.44
R / UT / UTS (10 %)	62.9 / **63.5** / 61.2	**27.7** / 24.0 / 27.0	**52.9** / 51.6 / 51.8	1.37 / 1.47 / **1.33**
R / UT / UTS (15 %)	**65.8** / 64.9 / 61.4	**31.1** / 24.4 / 26.8	**56.3** / 53.9 / 51.7	**1.25** / 1.35 / 1.31
R / UT / UTS (20 %)	**67.8** / 66.0 / 65.5	**32.4** / 26.1 / 31.7	**58.9** / 55.7 / 56.9	**1.20** / 1.31 / 1.26
R / UT / UTS (30 %)	68.0 / 65.4 / **68.0**	34.3 / 27.3 / **34.8**	59.2 / 55.0 / **59.7**	1.20 / 1.21 / **1.17**

Table 1: Performance metrics on hold-out test set (R = random subsets, UT = uncertainty threshold & UTS = stratified uncertainty threshold, both based on most uncertain samples). The metrics are overall accuracy (OA), average accuracy (AA), Kappa score (κ), and cross-entropy (CE), shown as averages over 3 independent runs. Further performance metrics showed similar trends. The results validate the hypothesis that the most "*uncertain*" samples extracted by the proposed uncertainty-guided representation learning approach induce more diversity during training, and hence makes the trained model more generic.

badly for small subset sizes. This behavior can be partially explained by the large class imbalance present in the training data set, and by the fact that some classes might overall come with smaller label uncertainty, both resulting in an unbalanced training data subset. Note that for larger subsets, the uncertainty-based models are able to overtake the random models in some performance metrics.

5. CONCLUDING REMARKS

In the context of LCZ classification, the So2Sat LCZ42 data set is studied, particularly a subset of European cities for which a label distribution exists. For a sufficiently large data set size, models focusing on data with higher uncertainty values seem to generalize better to the unseen test data. We explain this phenomenon by the richer informativeness of uncertain samples to the model. As an outlook, we would like to focus future work on transferring these findings to feature extraction models that are trained on different data, allowing for an a priori uncertainty assessment of the training data set. This would enable to sort out redundant or certain data samples and empower the model to focus on challenging but informative subsets of the data.

6. REFERENCES

[1] I. D. Stewart and T. R. Oke, "Local climate zones for urban temperature studies," *Bull. Am. Meteorol. Soc.*, 93(12), pp. 1879–1900, 2012.

[2] G. Thomas, A. Sherin, S. Ansar, and E. Zachariah, "Analysis of urban heat island in kochi, india, using a modified local climate zone classification," *Procedia Environ. Sci.*, 21, pp. 3–13, 2014.

[3] F. Leconte, J. Bouyer, R. Claverie, and M. Pétrissans, "Using local climate zone scheme for uhi assessment: Evaluation of the method using mobile measurements," *Build. Environ.*, 83, pp. 39–49, 2015.

[4] X. X. Zhu, J. Hu, C. Qiu, Y. Shi, J. Kang, L. Mou, H. Bagheri, M. Haberle, Y. Hua, R. Huang, et al., "So2sat lcz42: a benchmark data set for the classification of global local climate zones [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, 8(3), pp. 76–89, 2020.

[5] C. Qiu, X. Tong, M. Schmitt, B. Bechtel, and X. X. Zhu, "Multilevel feature fusion-based cnn for local climate zone classification from sentinel-2 images: Benchmark results on the so2sat lcz42 dataset," *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, 13, pp. 2793–2806, 2020.

[6] C. Qiu, L. Mou, M. Schmitt, and X. X. Zhu, "Local climate zone-based urban land cover classification from multi-seasonal sentinel-2 images with a recurrent residual network," *ISPRS J. Photogramm. Remote Sens.*, 154, pp. 151–162, 2019.

[7] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *NeurIPS 2017*.

[8] X. Dai, X. Wu, B. Wang, and L. Zhang, "Semisupervised scene classification for remote sensing images: A method based on convolutional neural networks and ensemble learning," *IEEE Geosci. Remote Sens. Lett.*, 16(6), pp. 869–873, 2019.

[9] F. Lv, M. Han, and T. Qiu, "Remote sensing image classification based on ensemble extreme learning machine with stacked autoencoder," *IEEE Access*, 5, pp. 9021–9031, 2017.

[10] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," *NeurIPS 2017*.

[11] M. Ruβwurm, M. Ali, X. X. Zhu, Y. Gal, and M. Körner, "Model and data uncertainty for satellite time series forecasting with deep recurrent models," *IGARSS 2020*.

[12] M. Neumann, A. S. Pinto, X. Zhai, and N. Houlsby, "Indomain representation learning for remote sensing," *arXiv preprint arXiv:1911.06721*, 2019.

[13] A. Elshamli, G. W. Taylor, A. Berg, and S. Areibi, "Domain adaptation using representation learning for the classification of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, 10(9), pp. 4198–4209, 2017.

[14] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, 55(9), pp. 5148–5157, 2017.

[15] M. Gong, T. Zhan, P. Zhang, and Q. Miao, "Superpixel-based difference representation learning for change detection in multispectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, 55(5), pp. 2658–2673, 2017.

[16] X. Geng, "Label distribution learning," *IEEE Trans. Knowl. Data Eng.*, 28(7), pp. 1734–1748, 2016.